Public Administration and Public Policy/71

Handbook of Research Methods in Public Administration

edited by
Gerald J. Miller
Marcia L. Whicker

Handbook of Research Methods in Public Administration

edited by

Gerald J. Miller
Marcia L. Whicker

Rutgers University
Newark, New Jersey

MARCEL DEKKER, INC. NEW YORK · BASEL · HONG KONG

ISBN: 0-8247-0213-1

This book is printed on acid-free paper

Headquarters
Marcel Dekker, Inc
270 Madison Avenue, New York, NY 10016
tel 212-696-9000, fax 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel 44-61-261-8482, fax 44-61-261-8896

World Wide Web
http //www dekker com

The publisher offers discounts on this book when ordered in bulk quantities For more information, write to Special Sales/Professional Marketing at the headquarters address above

Current printing (last digit)
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

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

1. *Public Administration as a Developing Discipline (in two parts),* 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
4. *Personnel Management in Government: Politics and Process,* Jay M. Shafritz, Walter L. Balk, Albert C. Hyde, and David H. Rosenbloom
5. *Organization Development in Public Administration (in two parts),* edited by Robert T. Golembiewski and William B. Eddy
6. *Public Administration: A Comparative Perspective, Second Edition, Revised and Expanded,* Ferrel Heady
7. *Approaches to Planned Change (in two parts),* Robert T. Golembiewski
8. *Program Evaluation at HEW (in three parts),* edited by James G. Abert
9. *The States and the Metropolis,* Patricia S. Florestano and Vincent L. Marando
10. *Personnel Management in Government: Politics and Process, Second Edition, Revised and Expanded,* Jay M. Shafritz, Albert C. Hyde, and David H. Rosenbloom
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
13. *Encyclopedia of Policy Studies,* edited by Stuart S. Nagel
14. *Public Administration and Law: Bench v. Bureau in the United States,* David H. Rosenbloom
15. *Handbook on Public Personnel Administration and Labor Relations,* edited by Jack Rabin, Thomas Vocino, W. Bartley Hildreth, and Gerald J. Miller
16. *Public Budgeting and Finance: Behavioral, Theoretical, and Technical Perspectives, Third Edition,* edited by Robert T. Golembiewski and Jack Rabin
17. *Organizational Behavior and Public Management,* Debra W. Stewart and G. David Garson
18. *The Politics of Terrorism: Second Edition, Revised and Expanded,* edited by Michael Stohl
19. *Handbook of Organization Management,* edited by William B. Eddy
20. *Organization Theory and Management,* edited by Thomas D. Lynch
21. *Labor Relations in the Public Sector,* Richard C. Kearney
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

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

24. *Public Administration: A Comparative Perspective, Third Edition, Revised,* Ferrel Heady
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
30. *Personnel Management in Government: Politics and Process, Third Edition, Revised and Expanded,* Jay M. Shafritz, Albert C. Hyde, and David H. Rosenbloom
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, Revised and Expanded,* edited by Michael Stohl
34. *Handbook on Human Services Administration,* edited by Jack Rabin and Marcia B. Steinhauer
35. *Handbook of Public Administration,* edited by Jack Rabin, W. Bartley Hildreth, and Gerald J. Miller
36. *Ethics for Bureaucrats: An Essay on Law and Values, Second Edition, Revised and Expanded,* John A. Rohr
37. *The Guide to the Foundations of Public Administration,* Daniel W. Martin
38. *Handbook of Strategic Management,* edited by Jack Rabin, Gerald J. Miller, and W. Bartley Hildreth
39. *Terrorism and Emergency Management: Policy and Administration,* William L. Waugh, Jr.
40. *Organizational Behavior and Public Management: Second Edition, Revised and Expanded,* Michael L. Vasu, Debra W. Stewart, and G. David Garson
41. *Handbook of Comparative and Development Public Administration,* edited by Ali Farazmand
42. *Public Administration: A Comparative Perspective, Fourth Edition,* Ferrel Heady
43. *Government Financial Management Theory,* Gerald J. Miller
44. *Personnel Management in Government: Politics and Process, Fourth Edition, Revised and Expanded*, Jay M. Shafritz, Norma M. Riccucci, David H. Rosenbloom, and Albert C. Hyde
45. *Public Productivity Handbook*, edited by Marc Holzer
46. *Handbook of Public Budgeting*, edited by Jack Rabin
47. *Labor Relations in the Public Sector: Second Edition, Revised and Expanded*, Richard C. Kearney
48. *Handbook of Organizational Consultation*, edited by Robert T. Golembiewski
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
51. *Handbook of Organizational Behavior*, edited by Robert T. Golembiewski
52. *Handbook of Administrative Ethics,* edited by Terry L. Cooper
53. *Encyclopedia of Policy Studies: Second Edition, Revised and Expanded,* 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
59. *Public Administration: A Comparative Perspective, Fifth Edition*, Ferrel Heady

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

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, Revised and Expanded*, edited by Robert T. Golembiewski and Jack Rabin
65. *Handbook of Public Administration: Second Edition*, edited by Jack Rabin, W. Bartley Hildreth, and Gerald J. Miller
66. *Handbook of Organization Theory and Management: The Philosophical Approach*, edited by Thomas D. Lynch and Todd J. Dicker
67. *Handbook of Public Finance*, edited by Fred Thompson and Mark T. Green
68. *Organizational Behavior and Public Management: Third Edition, Revised and Expanded*, 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
77. *Handbook of Public Information Systems,* edited by G. David Garson
78. *Handbook of Global Economic Policy,* edited by Stuart S. Nagel
79. *Handbook of Strategic Management: Second Edition, Revised and Expanded,* 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, Revised and Expanded,* 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, Revised and Expanded*, edited by Terry L. Cooper
87. Handbook of Organizational Behavior: Second Edition, Revised and Expanded, 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, Revised and Expanded*, 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

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

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, Revised and Expanded*, edited by Aman Khan and W. Bartley Hildreth

Additional Volumes in Preparation

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

Handbook of Developmental Policy Studies, edited by Stuart S. Nagel

Handbook of Conflict Management, edited by William J. Pammer, Jr., and Jerri Killian

ANNALS OF PUBLIC ADMINISTRATION

1. *Public Administration: History and Theory in Contemporary Perspective*, edited by Joseph A. Uveges, Jr.
2. *Public Administration Education in Transition*, edited by Thomas Vocino and Richard Heimovics
3. *Centenary Issues of the Pendleton Act of 1883*, edited by David H. Rosenbloom with the assistance of Mark A. Emmert
4. *Intergovernmental Relations in the 1980s*, edited by Richard H. Leach
5. *Criminal Justice Administration: Linking Practice and Research*, edited by William A. Jones, Jr.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Preface

The need for more rigorous and systematic research in public administration has grown as the complexity of problems in government and nonprofit organizations has increased. This book describes and explains the use of research methods that will strengthen the research efforts of those solving government and nonprofit problems.

This book is aimed primarily at those studying research methods in masters and doctoral level courses in curricula that concern the public and nonprofit sector. Thus, students in programs in public administration, nonprofit management, criminal justice, nursing, and education, to mention a few, will be provided detailed information on conceptualizing, planning, and implementing research projects of many different types.

The book is also aimed at consumers of research reports. For example, government executives who fund research must be able to determine whether the research objectives set out in the project are properly conceptualized and whether the research methods chosen are appropriate to the objectives and concepts. This volume will inform such research consumers.

Gerald J. Miller
Marcia L. Whicker

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Contents

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Contributors

LESLIE R. ALM, PH.D. Associate Professor, Department of Public Policy and Administration and Political Science, Boise State University, Boise, Idaho

EVAN M. BERMAN, PH.D. Associate Professor, Department of Public Administration, University of Central Florida, Orlando, Florida

CHIEH-CHEN BOWEN, PH.D. Assistant Professor, Department of Psychology, Cleveland State University, Cleveland, Ohio

WILLIAM M. BOWEN, PH.D. Associate Professor, Cleveland State University, Cleveland, Ohio

STEVEN R. BROWN, PH.D. Professor, Department of Political Science, Kent State University, Kent, Ohio

LYNN BURBRIDGE, PH.D. Assistant Professor, Department of Public Administration, Rutgers University, Newark, New Jersey

CARMEN CIRINCIONE, PH.D. Assistant Professor, Department of Political Science, University of Connecticut, Storrs, Connecticut

PHYLLIS D. COONTZ, PH.D. Associate Professor, Graduate School of Public and International Affairs, University of Pittsburgh, Pittsburgh, Pennsylvania

PATRIA D. DE LANCER, PH.D. Assistant Professor, School of Public Affairs and Administration, University of Illinois at Springfield, Springfield, Illinois

DAN W. DURNING, PH.D. Research Associate, Carl Vinson Institute of Government, University of Georgia, Athens, Georgia

VATCHE GABRIELIAN, PH.D. Associate Director, National Center for Public Productivity, Graduate Department of Public Administration, Rutgers University, Newark, New Jersey

NICHOLAS A. GIANNATASIO, PH.D. Assistant Professor, Department of Political Science and Public Administration, University of North Dakota, Grand Forks, North Dakota

ELIZABETH A. GRADDY, PH.D. Associate Professor, Program in Public Policy, School of Public Administration, University of Southern California, Los Angeles, California

RONALD JOHN HY, PH.D. Professor and Chair, Department of Geography, Political Science, and Sociology, University of Central Arkansas, Conway, Arkansas

GEORGE JULNES, PH.D. Visiting Research Specialist, Institute for Public Affairs, University of Illinois at Springfield, Springfield, Illinois

DAVID KANE, M.D., PH.D. Statistician, Numeric Investors, Cambridge, Massachusetts

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

DEIRDRE M. MAGEEAN, PH.D. Associate Professor, Department of Resource Economics and Policy, Smith Center for Public Policy, University of Maine, Orono, Maine

SARMISTHA R. MAJUMDAR, M.A., M.C.R.P. Doctoral Candidate, Department of Public Administration, Rutgers University, Newark, New Jersey

MICHAEL MARGOLIS, PH.D. Professor, Department of Political Science, University of Cincinnati, Cincinnati, Ohio

GERALD J. MILLER, PH.D. Associate Professor, Department of Public Administration, Rutgers University, Newark, New Jersey

CHANGHWAN MO, M.P.A. Doctoral Candidate, Department of Public Administration, Rutgers University, Newark, New Jersey

ALANA NORTHROP, PH.D. Professor, Department of Political Science, California State University at Fullerton, Fullerton, California

DONIJO ROBBINS, PH.D.* Instructor, Department of Public Administration, Rutgers University, Newark, New Jersey

CARMINE P. F. SCAVO, PH.D. Associate Professor, Department of Political Science, and Director, M.P.A. Program, East Carolina University, Greenville, North Carolina

SALLY SELDEN, D.P.A. Assistant Professor, Department of Public Administration, Syracuse University, Syracuse, New York

MACK C. SHELLEY II, PH.D. Professor, Department of Statistics and Political Science, Iowa State University, Ames, Iowa

MARCIA L. WHICKER, PH.D. Professor and Chair, Department of Public Administration, Rutgers University, Newark, New Jersey

DANIEL W. WILLIAMS, PH.D. Professor, School of Public Affairs, Baruch College, New York, New York

---

* *Current affiliation*: Assistant Professor, Department of Public Administration, University of Maine, Orono, Maine.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1
Introduction

Gerald J. Miller and Marcia L. Whicker
Rutgers University, Newark, New Jersey

The purposes of this handbook are varied and they build on each other. First, it provides a comprehensive survey of quantitative methods used in public administration research whether in government administration of public programs or in academic research involving theory-building and theory-testing. Second, the authors document illustrative past uses of quantitative methods in public administration. They link scientific quantitative techniques to their uses in public administration literature and practice in the past and present. Third, the chapters explore potential emerging uses of quantitative methods in public administration. These chapters illustrate to students, faculty and practitioners how various quantitative methods may be used to help answer emerging theoretical and public policy questions.

The audience for this handbook is multifaceted. First, a primary audience for the handbook is faculty and academic researchers as well as practitioners who use quantitative methods in their work, especially to expand the knowledge base of public administration and public policy. Second, doctoral students will find the book especially suitable for use as a text in methods seminars and as a reference in other graduate seminars revolving around past and emerging research problems. Third, Masters of Public Administration program students will have these chapters for their use in the courses covering research methods and program evaluation in their programs.

The book has four significant strengths. First, the exposition here contributes to the improvement and sophistication of research and research methods used in public administration research wherever done, in the university, in the public agency, or among consultants and researchers funded by foundations and other such organizations. Second, it stands as a reference manual for researchers as they deal with various quandaries in carrying out their various projects. Third, the chapters expose doctoral students to the wide variety of methodologies available to them. Finally, we hope that the authors give Masters students an awareness of the variety of methods available to them as well, but we hope that the chapters provide a high level of comfort to students in using quantitative methods, whether in understanding work they read or in their own research. Thus, the revolution of desktop computing has made powerful research methods readily available to current and future students. This handbook will increase their awareness and ease in dealing with those methods, both for consuming studies that they use in their jobs as well in carrying out research projects.

The chapters are grouped in nine main areas:

1. The Big Picture
2. Describing and Measuring Phenomena

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

3. Data Collection and Manipulation
4. Research Issues and Design
5. Association and Testing Hypotheses
6. Data Across Time
7. Techniques with Multiple Independent Variables
8. Modeling
9. Clustering Techniques

Following this introduction, Phyllis D. Coontz discusses ''Ethics in Systematic Research.'' Her discussion highlights the real and difficult problems researchers face continually. In the next chapter, the editors describe ''Levels of Data, Variables, Hypotheses, and Theory.''

Beginning Part 2 on describing and measuring phenomena, Changhwan Mo explains ''Univariate Measures for Directly Measurable Phenomena.'' William M. Bowen and Chieh-Chen Bowen then outline ''Typologies, Indexing, Content Analysis, Meta-Analysis, and Scaling as Measurement Techniques.''

Part 3 is devoted primarily to the procedures underlying survey research—data collection and manipulation. It begins with Donijo Robbins' treatment of ''Questionnaire Construction.'' Alana Northrup then describes ''Sampling and Data Collection.'' Finally, Carmine P. F. Scavo gives useful insight into ''Constructing Data Sets and Manipulating Data.''

In Part 4, research issues are discussed, especially those involving research design. In the first of these chapters, Nicholas Giannatasio outlines the ''Threats to Validity of Research Designs.'' More generally, Vatche Gabrielian thoroughly discusses the alternatives to quantitative research in ''Qualitative Research Methods: An Overview.''

Returning to quantitative research, Part 5 covers association and testing hypotheses. Leading off, Michael Margolis considers ''Statistics for Nominal and Ordinal Data.'' Beyond these methods, Carmen Cirincione explicates ''Analysis of Variance.'' Finally, Leslie R. Alm discusses the appropriate uses of ''Linear Correlation and Regression.''

Going beyond static pictures of phenomena, the next part looks at data sets collected from multiple points. Lynn Burbridge first explains the uses and misuses of these data sets in ''Cross-Sectional, Longitudinal, and Times-Series Data: Uses and Limitations.'' Dan Williams then describes a major use for these data sets in ''Forecasting Methods for Serial Data.'' Finally, Deidre Mageean outlines ''Demographic Techniques for Cohort Analysis and Population Trends.''

In situations with multiple independent variables, the next part deals with their manipulation and interpretation. First, Elizabeth A. Graddy explains ''Multivariate Regression Analysis in Public Policy and Administration.'' Then, Mack C. Shelley, II provides insight into a specific case in ''Multivariate Techniques for Dichotomous Dependent Variables.''

Of increasing importance, modeling moves center stage in Part 8. In the initial chapter Evan M. Berman looks at ''Causal Modeling and Path Analysis.'' Then, Ronald John Hy cast special light on ''Economic Modeling.'' David Kane then moves into one of the most important uses of models in ''Computer Simulation.'' Finally, introducing a new and increasingly important technique, Patria D. de Lancer explains ''Data Envelopment Analysis.''

In the final part, authors describe data clustering techniques. First, George Julnes surveys ''Principal Component Analysis, Factor Analysis, and Cluster Analysis.'' Then, Steven R. Brown, Dan Durning, and Sally Coleman Selden take a look at ''Q Methodology.''

The Appendix chapter on ''Algebra'' is provided by Rina Majumdar.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

2
Ethics in Systematic Research

Phyllis D. Coontz
University of Pittsburgh, Pittsburgh, Pennsylvania

I. OVERVIEW

This chapter focuses on ethical issues that arise in the conduct of social research. Ethical issues necessarily emerge during the research process because the methods researchers use are intrusive—researchers invade peoples' lives through the questions they ask and by the behavior they observe. Moreover, in order to do research, social scientists need the cooperation of others—this is so regardless of the type of research one does (e.g. field work or telephone surveys) or the setting in which the research is carried out (e.g. in a hospital or business organization). The relationship between the researcher and the participant of research is fiduciary in nature and is based on trust. Thus, the researcher has a responsibility to protect the rights of those who agree to participate in research and participants expect to be treated humanely and ethically. Ethical research practices require taking the appropriate steps to insure that the rights of participants are respected and protected.

Although ethics and research go hand in hand, not all researchers act ethically nor are ethics automatically integrated into the practice of research. This is not to suggest that people are naturally unethical or deliberately act in unethical ways, but rather to stress the complexity of the research process and its potential to impact the lives of others—either socially, psychologically, or physically. Since the effects from research may not always be apparent, the good researcher anticipates the potential consequences from the study. Thus, learning to do good research not only involves using the appropriate methods to study an issue, but also employing ethical standards throughout the research process.

What is meant by the term *ethics*? According to Kimmel (1988), ethical issues are moral issues and both are related to values. When we speak of ethics, we are speaking about the values we hold (what we deem important or an inalienable condition). Such values are reflected in our norms and prescribe our behavior, i.e. what is expected and what we consider to be ''right.'' Questions about what the ''right thing'' to do is arise whenever there is uncertainty, ambiguity, or conflict around our values. Smith (1985) refers to such uncertainty as ethical dilemmas. In a research context, ethical dilemmas can apply to the conduct of research, the subject matter of research, the balance between personal goals and professional goals, the decision of whether or not to investigate a topic, and the uses of research findings (Kimmel, 1988: 33–35).

Ethical dilemmas are related to the goals, processes, and outcomes of social science. Within this context, three general areas are of concern: the ethical treatment of human subjects, the ethics of data collection and analysis, and the ethical uses of scientific knowledge (Reese

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and Fremouw, 1984). I discuss each of these areas in this chapter. I also review the relevant federal regulations pertaining to the use of human subjects in research, the role of the IRB (Institutional Review Board[1]) at universities and colleges with respect to the use of human subjects, and highlight various codes of ethics developed in the social science community (see for example, the American Anthropological Association, 1971; the American Psychological Association, 1981; the American Sociological Association, 1981; and the National Association of Social Workers, 1979). The ethical regulations developed by the government provide more explicit rules for ethical conduct than do the codes of professional associations (Gillespie, 1987). While government regulations are designed to protect society and its members and offer specific steps to be followed by researchers, professional codes emphasize individual responsibilities for ethical research and tend to be more abstract. To underscore the range of ethical dilemmas that can arise in research, I draw upon actual cases that have raised questions, sparked controversy, or led to reform. These cases are not isolated, aberrant, or even exhaustive instances of ethical dilemmas, but rather are intended as heuristics for examining the sorts of ethical problems that can arise in the course of doing research and alerting the researcher to the range of potential ethical dilemmas.

II. TREATMENT OF HUMAN SUBJECTS

Much of the current debate on ethical research pertains to the treatment of human subjects. The impetus for this interest can be traced back to the atrocities by the Nazis during World War II. These came to light during the Nuremberg Trials when countless abuses committed by doctors and scientists on humans were revealed. The Nazi's human experiments were conducted against the will of those affected and included such practices as injecting healthy prisoners with various diseases (e.g., malaria, epidemic jaundice, and spotted fever) and poisons; simulated high altitudes in order to examine the effects; and experimentally inducing wounds (Katz, 1972). The Nuremberg Trials focused world wide attention on the abuse of human subjects and resulted in The Nuremberg Code of 1949 which set forth 10 moral, ethical, and legal principles about medical experimentation on humans (see Box 1). It was The Nuremberg Code that first established the concept of ''voluntary consent'' in human experimentation and has since served as a model for developing and assessing ethical practices in the social and behavioral sciences.

Box 1: The Nuremberg Code

1. The voluntary consent of the human subject is absolutely essential.
2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random or unnecessary in nature.
3. The experiment should be so designed and based on [previous research] that the anticipated results will justify performance of the experiment.
4. The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.
5. No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur, except perhaps, in those experiments where the experimental physicians also serve as subjects.
6. The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

7. Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.
8. The experiment should be conducted only by scientifically qualified persons.
9. During the course of the experiment the human subject should be at liberty to bring the experiment to an end.
10. During the course of the experiment the scientist in charge must be prepared to terminate the experiment if . . . continuation of the experiment is likely to result in injury, disability, or death of the experimental subject (1949: 181–182).

The importance we attach to the treatment of human subjects is related to the value that our culture attaches to the rights of individuals. We expect those who participate in research will be treated with respect and protected from harm. Despite the high value we attach to individual rights and federal regulations and various professional codes of conduct intended to guide researcher conduct, we regularly learn of new instances of unethical research practices. An obvious safeguard against this is to be attuned to the ethical implications of one's research.

The first exposure the novice researcher is likely to have with ethical issues is in the course of doing research for a thesis or dissertation. At the most general level, dissertation and thesis research requires some form of IRB oversight (regardless of how perfunctory) at universities and colleges who receive federal support for research. Since most dissertation/thesis research involves some contact with human subjects, it is a good idea to obtain a copy of your institution's IRB guidelines, discuss them with other students and faculty, and have others review your research protocol before submitting it for IRB review.

According to Dienner and Crandall (1978) the ethical treatment of human subjects applies to potential harm, informed consent, privacy and confidentiality, and deception. To reiterate an earlier point, ethical dilemmas arise when the goals, objectives, and outcomes of research are unclear or conflicting. Thus to cause harm or injury to others, to coerce someone to engage in activities against their will, to invade others' privacy without their permission, or to mislead or deceive participants are all actions that violate the spirit of trust between the researcher and the participant. IRB guidelines and the professional codes of ethics are there to delineate researcher's obligations and it is the researcher's responsibility to be familiar with his/her ethical obligations to participants of research, to colleagues, professional audiences, sponsoring agencies, and to the public and society at large (Gillespie, 1987: 503). Let us now examine each of these four areas of the ethical treatment of human subjects in greater detail.

III. POTENTIAL PHYSICAL AND PSYCHOLOGICAL HARM

Although physical harm to participants in social research is highly unlikely, people can be harmed personally (by being embarrassed or humiliated), psychologically (by losing their self-esteem), and socially (by losing their trust in others) (Diener and Crandall, 1978). Basic to the research process is whether the researcher's desire to advance knowledge or gain insight can be achieved without compromising fundamental rights of participants. Although it may be difficult to predict whether one's investigative procedures will harm participants, the researcher nevertheless should take measures to assess potential risks and benefits associated with his/her research. In its code of professional ethics, the American Psychological Association (APA) states: ''[R]esearch procedures likely to cause serious or lasting harm to a participant are not used unless the failure to use these procedures might expose the participant to risk of greater harm, or unless the research has great potential benefit'' (1990: 395). In other words, the re-

searcher should weigh the scientific value from the research against the potential risk to participants. If there is little scientific value to a study, then exposing participants to potential risk cannot be justified.

Determining potential risk is not always apparent at the outset of a study, but may surface sometime after a study has begun. Similarly, some participants may be at higher risk than others simply because of a pre-existing physical or psychological condition. The classic study of obedience to authority by psychologist Stanley Milgram (described more fully in Milgram's book, *Obedience to Authority* published in 1974) illustrates subtle risk from research and why it should be assessed before research is begun.

Milgram's study was designed as an experiment to examine how ordinary people could be induced to obey authority. The larger question intriguing Milgram was how the Holocaust happened. Participants were told the study was about the effects of punishment on learning. Participants were assigned to the role of teacher and given the task of administering increasingly stronger electric ''shocks'' (up to 450 volts) to a group of experimental confederates who posed as learners. The experiment was rigged so that confederates would not actually receive ''shocks'' (although the confederates were hooked up to an electrical shock box controlled by participants, no actual shocks were ever administered). Instead confederates acted out the pain when the real participants administered the ''shocks.'' Participants were unaware that confederates feigned the pain. Milgram planned the experiment so that when participants met confederates at the outset, confederates revealed they had a ''heart condition'' (in reality they did not). The researcher reasoned that such information could mitigate against administering shocks.

Upon reaching a certain level of electrical shock and hearing the staged pain reactions of confederates, some participants refused to continue administering the shocks and withdrew from the study. Others, however, continued to administer increasing levels of ''shocks'' (in spite of the knowledge of a pre-existing heart condition). Milgram's research was troubling because it showed that some participants were willing to obey the instructions of the researcher regardless of the harm, albeit staged, to confederates.

When the experiment was over, participants were naturally relieved to learn that they had not actually physically harmed confederates. However, some participants reported experiencing stress as a result of their actions even though the stress turned out to be short-lived. The criticism against Milgram focused mainly on his failure to take adequate measures to protect participants from undue stress associated with administering pain to others (Baumrind, 1964; Kelman, 1967). Critics also noted that Milgram had made no effort to determine prior to the experiment whether participants should be excluded for physical or psychological reasons. Other concerns were raised in regard to the effects that the experiment might have on participants' longer term self-concept—how would participants' perception of themselves be affected by the knowledge that they were capable of inflicting pain on another when asked to do so (Baumrind, 1964).

The Miligram experiment reminds us that psychological and social risk may result from one's research and while it may not always be easy to gauge the level of risk prior to the research, if the research deals with sensitive issues, the researcher should consider the long term impact that such issues might have on participants. Assessing such potential harm requires putting yourself in the participant's shoes and exploring the possible effects from all aspects of the research. Although most social science research does not use an experimental design, the real issue in assessing potential risk has less to do with design than with the issues being examined in the research. When these issues are sensitive or have the potential to trigger psychological reactions or erode trust, then the researcher is obliged to consider the various ways participants could be affected by the research. For example, researchers may ask questions that can threaten, embarrass, or humiliate participants. Participant observers can unintentionally harm others through their own active involvement as participants as Whyte did in his study of Street Corner

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Society. Whyte reports having voted four different times in a single election (Whyte 1981: 313–314). While it is reasonable to assume that little damage was done to the opposition candidate by Whyte's illegal votes, his actions are not irrelevant and cannot be dismissed.

What measures can researchers take to minimize risks to participants? Researchers have the obligation to inform participants of foreseeable risks or possible discomforts before a study begins and should give participants ample time to think about the implications of their participation. Researchers can also screen out participants who may suffer from psychological or physical problems that could be exacerbated by participating in the research. If stress or potential harm is a possible or anticipated outcome, measures should be taken to assess the degree of stress or harm anticipated from the study. One common way stressful effects can be neutralized is by debriefing participants after the study and providing them with procedures for contacting the principal investigator should problems develop. Debriefing sessions provide participants with an opportunity to discuss their feelings about their involvement and are useful for neutralizing negative reactions. Federal regulations mandate informing participants of the risks involved in any study that is federally funded. Such notification falls under the rubric of ''informed consent'' which I will now discuss.

IV. INFORMED CONSENT

There are two underlying principles involved in informed consent. One is the belief that participants have the right to chose whether to participate in research without fear of coercion or pressure. The key here is that participation is voluntary. The other principle is based on the belief that participants have the right to be given information that is relevant and necessary for making the decision to participate. Necessary information usually refers to information that bears upon the consequences to the participant as a result of participation. The researcher is obliged to disclose potential risks (whether physical, psychological, or social) involved by participation. Disclosure of potential risks does not mean full disclosure of the research purpose or the methods to be used, but rather how participants will be affected. A key feature of informed consent is that the information identifies the known effects from participating in the study. To provide such information requires the researcher to assess potential risk beforehand. Remember, it is not the amount of information provided, but rather the quality of the information provided, and its relevance for making an ''informed'' decision about participating.

Key elements in disclosure include a description of the general purpose of the study, a statement that participation is voluntary and that participants are free to withdraw at any time, a clear description of the potential risks and benefits involved (research may benefit a group or add to our knowledge about an issue valued by the participant), the name, address, and phone number of the person(s) responsible for the research, and a brief description of what will be done with the information once it is collected. Regulations for federally funded research require that participants sign a written consent form when more than ''minimal risk'' is anticipated. According to federal regulations, ''minimal risk'' refers to risk that is no greater than what can be expected in daily life. Signed consent protects both participants and researchers. Keep in mind that federal regulations *do not exempt research* that deals with sensitive issues such as drug use, sexual behavior, or criminality. IRBs require signed consent when doing research on sensitive topics or when dealing with special categories of participants such as juveniles.

It is assumed that informed consent can only be obtained from those who have the ability to give it, i.e. adults rather than children and those who are mentally competent to understand the meaning of the information they are asked to provide. Minors constitute a special protected category of participants. The protections already accorded minors may be extended by proposed

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

legislation in The Family Privacy Protection Act of 1995. This legislation seeks to protect minors from intrusive research and to safeguard parental rights in restricting the activities in which their children participate. There are two types of parental consent relevant here, ''passive'' consent and ''active'' consent. ''Passive'' parental consent requires parents to respond only if they do *not* want their child to participate in a research project. The process assumes that a nonresponse to consent is an affirmative response. ''Active'' parental consent assumes that a nonresponse is a refusal to participate. ''Active'' consent is required unless a researcher has obtained exemption from the IRB. In order to be exempted, the researcher must document that the research could not be completed using ''active'' consent procedures, that no more than ''minimal'' risk is involved for participants, and that every effort will be made to protect human subjects and inform them of the research procedures involved.

Signed consent forms protect researchers from potential liability (and they protect IRB institutions from liability). However, participant consent does not remove the researcher's responsibility to minimize risk and it should never be used to justify unethical practices. Most IRB guidelines contain sample consent forms. The consent form I am currently using in a study assessing drug treatment needs among newly arrested individuals with Jim Nesbitt at the University of Pittsburgh is shown in Box 2 below. This consent form is more explicit than is usually required because the study deals with the sensitive issue of drug use and is being done with a specially protected group of participants, prisoners.

Box 2

Approved __/__/__
Psychosocial IRB
University of Pittsburgh

CONSENT TO PARTICIPATE IN A RESEARCH STUDY

Title: Substance Abuse and Need for Treatment Among Arrestees Study

Investigators:	Phyllis D. Coontz, Ph.D.	James Nesbitt, M.P.A.
	University of Pittsburgh	University of Pittsburgh
	3G01 Forbes Quad	A223 Crabtree Hall
	Pittsburgh, PA 15260	Pittsburgh, PA 15213
	(412) 648–2654	(412) 624–3109

Description: The purpose of this study is to learn more about the drug use patterns and treatment needs of persons recently arrested for some type of criminal conduct. The Pennsylvania Department of Health has asked the University of Pittsburgh to conduct this study. In order to do this, we are asking about 650 individuals from around the state to participate in the study. If you agree to participate, you will be asked a number of questions that are of a personal nature that focus on your drug use. The interview will take approximately an hour to complete. We will not be asking you for your name, the names of anyone else, or the specific dates or specific places of any of your activities.

You will also be asked to provide a urine sample—in private with no one watching—which will be analyzed for the presence of drugs. No police, court or correctional personnel will have access to these samples or their results. Your urine sample will be tested and disposed of in a private licensed laboratory. The urine

container will be identified by code number only. No names will be used for the urine samples.

We will not ask you any questions about child abuse or neglect. Your questionnaire and urine test results will not be available to authorities or to any members of your family. We will not ask for or record your name or any other identifying information during the interview. We will not ask for or record your name or any other information that could identify who you are. The interview is not being tape recorded. If there are some questions that you don't want to answer, that's OK, you can skip them. Your participation in this study is voluntary and your participation in the urine testing is also voluntary. If you are willing to answer the questions in the interview, but do not want to participate in urine tests, you can still be part of the study. If you do not want to be a part of this research project or if you change your mind, you can quit anytime without any effect on you or your record. Your arrest status will not be affected if you do not participate in the study.

Page 2.

Risks and Benefits: The risks of this study relate to some of the questions that you will be asked during the interview. As indicated above, some of these questions are of a personal nature involving your use of illegal drugs. The interview will be conducted in private so that no one can overhear your responses or know what you are answering. You will not be asked your name or that of anyone else in either the interview or for the urine test. The benefits from this study are that you will help us learn more about how much drug use goes on among arrestees and how much need there is for treatment. There has never been a study examining these issues in Pennsylvania.

Costs and Payments: There will be absolutely no cost to you for your participation. If you agree to participate in the study, you will be compensated $10 when the interview is completed and you've given a urine sample. You may chose to receive the $10 in either a voucher at the commissary or in a cash payment.

Confidentiality: All information you give the researchers will be kept confidential. No personal information about you or anyone else will be asked of you. The interview and urine sample will be coded by number so that you can never be identified. Your identity will not be revealed in any description or publication of this research. You will be given copies of this consent form and the Federal Confidentiality Certificate. As indicated above, a Confidentiality Certificate protects the study staff from being forced, even under subpoena, to research any research data in which anyone is identified.

Right to Refuse to Participate: You are free to refuse to participate in this study and may end the interview at any time. Your participation or refusal to participate will not affect your arrest status. If you are willing to answer the interview questions, but are unwilling to give a urine sample, you may still participate in the study.

****

Voluntary Consent: I certify that I have read the preceding or it has been read to me and that I understand its contents. Any questions I have pertaining to the research will be answered by Phyllis Coontz, Ph.D. (412) 648–2654. Any question I have about my rights as a research subject will be answered by the office of the Senior

Vice Chancellor of Health Sciences, University of Pittsburgh (412) 647–8475. A copy of this form will be given to me. My signature below means that I agree to participate freely in this study.

Date Subject's Initials

Page 3.

Investigator's Certification: I certify that I have explained to the above individual the nature, purpose, potential benefits, and possible risks associated with participating in this study, have answered any questions that were raised, and have witnessed the above signature.

Date Signature of Research Staff/Interviewer

Questions about informed consent and ''protected'' participants were recently raised in relation to a study of cyberporn conducted by researchers at prestigious Carnegie Mellon University (CMU) in Pittsburgh (*The New York Times*, July 16, 1995). The study, titled *Marketing Pornography on the Information Superhighway*, examined uses of computer networks (i.e. Usenet), especially adult oriented computer bulletin board systems. The researchers identified consumers (whose usernames were supplied by the bbs operators) in over 2000 cities in 50 states and in 40 countries and analyzed the sorts of information they consumed on line. Researchers tracked the number of times that pornographic images were retrieved by computer users (a total of 6.4 million downloads). The findings sparked numerous debates about the appropriate uses of the Internet, censorship by universities and colleges,[2] as well as ethics around informed consent.

The study's principal investigator, Marty Rimm (a student), did not obtain consent from those whose computer files were accessed nor had the bbs operators. The researchers tracked Internet users' behavior without their knowledge—and clearly without their consent. In Pennsylvania, it is illegal to knowingly distribute sexually explicit material to anyone under the age of 18. Does downloading pornographic images constitute the ''distribution'' of those images? Since some of the students on college campuses today are under 18 years of age, the issues of parental consent and censorship are also relevant. Should the university obtain ''passive'' or ''active'' consent from parents to use campus computers? Or should the university prohibit underage students from using campus computers? Relevant to this discussion is whether users (or parents) would have given Rimm permission to track their Internet behavior had they known what the legal ramifications were in Pennsylvania or that their Internet behavior would be exposed.

Clearly CMU's actions indicate that the university administration perceived the risk of possible litigation and moved quickly to avoid it by banning Usenet groups from campus computers. The result has been a hue and cry over censorship and controlling the use of the Internet. Aside from the Constitutional issues involved, this case is also troubling for what it suggests about the breach in the fiduciary responsibility of faculty to monitor students' work. The fact that the principal investigator was a student enrolled at CMU and operated under the guidance of faculty advisors is not insignificant. That the study was completed without the consent of Internet users indicates that those with oversight responsibility either believed that consent was unnecessary or simply failed to consider the array of ethical implications involved in the research. IRB review would be helpful in sorting through potential risk and thus prevent such controversies from happening.

Another concern related to informed consent is the impact that obtaining a signed consent form might have on the recruitment of participants. Obtaining written consent could discourage participation and reduce the response rate. ''Passive'' consent procedures generally produce response rates between 80 to 96 percent, but obtaining a comparable response rate using ''active'' consent procedures increases the cost by as much as four times (follow-up telephone calls, multiple mailings, additional meetings with parents, and the additional time involved). Another consideration relates to the effect that informed consent can have on responses themselves. It has already been noted that participants may be more inclined to give socially desirable responses when they have a sense of what the researcher is looking for. And there is some evidence showing that participants who are told the purpose of a study do not behave as those who have not been told the purpose (Singer, 1978). The concern for the researcher is that obtaining consent could undermine a study's validity.

V. PRIVACY AND CONFIDENTIALITY

According to Westin, privacy refers to ''the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others'' (1968: 7). Sieber expands the notion of privacy to include confidentiality arguing that confidentiality ''refers to agreements between persons that limit others' access to private information'' (1982: 146). Thus, privacy refers to persons and confidentiality refers to information. The right to privacy is the individual's right to determine when, where, to what extent, and to whom his or her attitudes, beliefs, and behavior will be shared. One way in which a participant's privacy can be invaded is through the use of concealed devices such as microphones, cameras, or tapping into computer lines (as in the CMU study discussed above). When such devices are used with the participant's knowledge and consent, their use poses no problem. However, ethical dilemmas can arise whenever the desire to observe behavior under ''natural conditions'' interferes or invades a person's right to privacy.

Clearly, information that is shared anonymously protects the privacy of participants, but this safeguard is not always feasible when certain sampling procedures are used. For example, some researchers sample from organizational lists that contain the names, addresses, and phone numbers of employees. The researcher is obliged to take appropriate measures to protect the identities of those who agree to participate in research. A common way this is done is by removing any personally identifying information from the data collection instrument itself. Sometimes, the researcher may use a follow-up strategy to increase the return rate or may find it is necessary to verify or correct information already gathered. In such cases, the researcher is again obliged to protect the identities of participants. A method for this is to use a coded ''master file'' that links a participant's name to an i.d. number and keep such a file locked in a file cabinet to which only those with responsibility for the research project have access.

When face to face interviews are conducted, a common way to protect the identities of those who have been interviewed or of the organizations or the communities being studied is to use pseudonyms, fictitious histories, or global descriptions. This approach is not always foolproof as is seen in the study by Arthur Vidich and Joseph Bensman of a small town in upstate New York which they fictitiously named ''Springdale'' (1958). This case illustrates how easily people can be identified by researchers' descriptions when they contain too much identifying information. In this case, although the researchers promised participants that pseudonyms would be used, they did not attempt to alter the backgrounds, occupations, or other personally identifying information of participants. Consequently, people's identities were easily recognizable in the published results. Those who participated were outraged and felt betrayed by what they

considered a breach of confidentiality that had been assured by the researchers. Despite the pseudonyms, townspeople were able to identify each other. The reason participants were upset by the study's findings is that they were not particularly flattering, thus participants were embarrassed by the study's depiction of their town and themselves and angry for believing the assurances of researchers. In addition to making the identities of the townspeople transparent, Vidich and Bensman were further criticized by their colleagues for failing to obtain the consent of their participants.

It should also be noted that research data are not considered *privileged*, and can be subject to subpoena. Returned surveys and questionnaires, notes, field records, and files can all be accessed by the federal government under certain administrative provisions, such as the Freedom of Information Act or the Federal Property and Administrative Services Act. Gelles reminds us that ''[R]esearchers who engage in research that deals with illegal, sensitive, or taboo topics run the risks of being forced to turn over material they pledged would be kept confidential, of engaging in legal battles, or of spending time in jail for contempt of court'' (1978: 422).

One may inadvertently uncover, during the course of research, information about illegal behavior, drug use, or child abuse, that may place the participant or others at risk. In the case of discovering child abuse, although researchers are not classified as mandated reporters, one must decide what to do if maltreatment is uncovered during the course of the research. This decision cannot be made independent of the assurances that have been given to participants about privacy and confidentiality.

An example of this sort of ethical dilemma is vividly illustrated in Inciardi, et al.'s (1992) ethnography of crack cocaine in Miami. The researchers write:

> Upon entering a room in the rear of the crack house (what I later learned was called the freak room), I observed what appeared to be the gang-rape of an unconscious child. Emaciated, seemingly comatose, and likely no older than 14 or 15 years of age, she was lying spread eagled on a filthy mattress while four in succession had vaginal intercourse with her. After they had finished and left the room, however, it became clear that, because of her age, it was indeed rape, but it had not been ''forcible'' rape in the legal sense of the term. She opened her eyes and looked about to see if anyone was waiting. When she realized that our purpose there was not for sex, she wiped her groin with a ragged beach towel, covered herself with half a tattered sheet (affecting a somewhat peculiar sense of modesty), and rolled over in an attempt to sleep. Almost immediately, however, she was disturbed by the door man, who brought a customer to her for oral sex. He just walked up her with an erect penis in his hand, said nothing to her, and she proceeded to oblige him.
>
> When leaving the crack house a few minutes later, the dealer/informant explained that she was a ''house girl''—a person in the employ of the crack-house owner. He gave her food, a place to sleep, and all the crack she wanted; in return, she provided sex—any type and amount of sex—to his crack-house customers.
>
> When I first walked into that room—and I can still vividly picture the scene—my reaction was one of highly repressed outrage. My thought was to somehow get between the men and the child, provide a distraction, play it by ear. But as I made a move toward the group, my protector took me by the arm, quite firmly I might add, and said in a very matter-of-fact way: ''You can't do anything. Just let it be. If you do anything, I'll have to kill you. It's as simple as that. I brought you here, I vouched for you. You interfere, and if they (pointing to the men with the child) don't do you in, I will'' (1993: 154–55).

The researchers go on to tell us that it would have served little purpose to contact the police or child protection agencies. It became clear that the child involved had been addicted to crack for a year and had no intention of leaving the crack house since it was the only place she had to live. Field workers know that developing rapport is the only way to gain access to

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

certain settings, especially illegal ones. By developing relationships with those involved in the crack industry the researchers were able to observe things usually inaccessible to researchers. Eventually, the researchers were able to persuade the child to enter a drug treatment program. They emphasize the fact that the crack business is filled with degradation, brutality, despair, and exploitation, there is little or nothing overtly that an outsider can do since it would likely lead to serious violence. However, subtle intervention is an option for the researcher who is accepted and trusted. Consider the consequences of disrupting the flow of everyday practices on the street. The researchers suggest that everyone loses when the doors to the crack industry are closed.

VI. DECEPTION

Deception is perhaps the most controversial aspect of the treatment of human subjects because it is widely used and there is a lack of consensus about whether it is appropriate. The most common way that participants are deceived involves intentionally misleading them about the purpose of the research, e.g., the Milgram study. Deception has been justified on the grounds that it is necessary in order to preserve the natural mental state of participants. As we have seen, informing participants of the purpose of the study or obtaining their consent can effect both the response rate and responses—participants might respond in ways different from how they would ordinarily respond if they did not know the purpose of the study, thus rendering the findings meaningless. As I mentioned with regard to the CMU study, participants sometimes try to present a favorable image of themselves or may try to assist the researcher by responding the way they think the research expects them to respond. Deception provides the researcher with a way to divert participants' attention away from the topic of the research.

The frequent use of deception in research was documented by Adair et al. (1985) who found that 58 percent of the empirical studies published in three leading social psychological journals used deception. In a study that compared deceived participants with those who were not deceived, Smith (1981) concluded that participants are willing to accept some deception when the research seems justified by its scientific importance. Baumrind (1981) argues that deception is never justified because it is unethical since it involves lying to participants, however, the code of ethics of the APA (1990: 394) does not rule out deception, but specifies the conditions under which deception is allowable—when methodological requirements necessitate it. The APA adds the proviso that researchers using deception have a ''special responsibility'' to determine whether there are alternative procedures available and to ensure that participants are provided with an explanation as soon as possible'' (1990: 394–95).

One of the more controversial studies involving deception was Laud Humphrey's study of anonymous sex in public restrooms (1975). While a doctoral candidate in sociology, Humphreys became a participant-observer in a number of homosexual acts occurring in ''tearooms''—public restrooms. He assumed the role of a ''watchqueen'' (this refers to someone who is a lookout to warn those having sex of approaching strangers) and observed sex in public restrooms. Besides observing this behavior, Humphreys wanted to learn more about the lifestyles, backgrounds, and motivations of those who engaged in anonymous sex in public restrooms. To do this, Humphreys developed rapport with some of the men he observed. To expand his sample, he traced the registration numbers of the cars of some of the men he had observed in order to learn their home addresses. Once he located these men, Humphreys posed (these participants did not recognize him from the public restrooms) as a health service interviewer and asked these men to provide (voluntarily) demographic and attitudinal information. At no time did Humphreys reveal that he was aware of the respondent's participation in the tearoom subculture.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Humphreys' research was applauded by some, but criticized by others. One of the ethical questions raised by Humphreys' research involved the extent to which deception can be justified. Humphreys argued that informing the participants about the nature of the study would have compromised his ability to do it. However, Baumrind argues that ''intentional deception in the research setting is unethical, imprudent, and unwarranted scientifically'' (1985: 165). According to Baumrind (1985), deception is unacceptable because it violates a participant's right to informed consent and violates the trust implicit in the researcher-participant relationship. Baumrind further notes that the almost routine use of deception undermines the researcher enterprise because it leads some potential participants to suspect (and thus reject) of all research. Suspicions about the motives of research by part of participants challenges the claim that deception will produce valid information (Baumrind, 1985).

VII. RESOLVING ETHICAL DILEMMAS

It should be clear from the preceding discussion that there are no easy or patent answers to the ethical dilemmas that arise in research. In fact, by definition an ethical dilemma is a conflict situation in which the researcher must reconcile between two or more courses of action—whether the conflict is related to basic human rights such as privacy, autonomy, or protection from harm, obtaining a good response rate, or to more lofty goals such as advancing knowledge. How then does the researcher resolve ethical dilemmas?

Kimmel (1988) provides us with some guidance here. According to Kimmel (1988), research decisions are based on two sorts of ethical theories—teleological theory and deontological theory. A *teleological* theory of ethics holds that an action is right or obligatory if it or the rule under which it falls produces the greatest possible balance of good over evil. In short, the consequences of an act determine its value. An act is considered morally right if it leads to desirable outcomes. On the other hand, *deontological* theorists argue that considerations other than consequences are what is relevant in moral decision-making. Deontologists argue that certain acts are to be viewed morally right because they are intrinsically good. Thus, certain conduct is either right or wrong, irrespective of the outcome. Most social scientists embrace a teleological approach to ethics. The morality of acts should be determined on the basis of the ends they serve. If we embrace a teleological perspective, then we are obliged to weigh the significance of the scientific knowledge to be gained from the research we engage in against the potential costs or harm to participants of the research.

VIII. INSTITUTIONAL REVIEW BOARDS (IRBS)

In the final analysis, the individual researcher is responsible for deciding which course of action to take when faced with ethical dilemmas. Since potential risks and benefits are not always apparent, I hope that it is clear that the advice and opinions of others can be of enormous help. Increasingly, ethical decisions about supported research is the responsibility of IRBs. According to federal regulations, each IRB should have at least five members with varying backgrounds that ensure the adequate review of research proposals (including dissertation proposals). To provide a cross-section of expertise, the members must include at least one nonscientist (such as a lawyer, ethicist, or member of the clergy), at least one member not affiliated with the research institution, along with persons competent to review specific research activities (e.g., sociologists or anthropologists). Researchers are required to submit a written protocol to the IRB that describes the proposed research and outlines the measures to be used to protect the

rights of participants. Of particular interest to IRBs are informed consent and confidentiality. Once reviewed, IRBs can then approve, modify, or disapprove the research. The basis for their action comes from federal regulations outlined by DHHS (these appear in the January 26, 1981 issue of the Federal Register).

In reviewing a research protocol, IRBs are concerned that researchers meet the following conditions: (1) risks to participants are minimized by sound research procedures that do not unnecessarily expose subjects to risks; (2) risks to participants are outweighed sufficiently by anticipated benefits to participants and the importance of the knowledge to be gained; (3) the rights and welfare of subjects are adequately protected, (4) the activity will be periodically reviewed; and (5) informed consent has been obtained and appropriately documented.[3]

As mentioned earlier, in addition to federal regulations, social scientists are guided by ethical codes for the treatment of research participants developed by professional societies. Box 3 below contains excerpts from the ethical codes of the American Anthropological Association (AAA), the American Sociological Association (ASA), and the American Psychological Association (APA) regarding the treatment of research participants. Complete copies of these codes can be obtained directly from these associations.

Box 3: Treatment of Research Participants

From the American Anthropological Association

In research, anthropologists' paramount responsibility is to those they study. When there is a conflict of interest, these individuals must come first. Anthropologists must do everything in their power to protect the physical, social, and psychological welfare and to honor the dignity and privacy of those studied . . .

The aims of the investigation should be communicated as well as possible to the informant. Informants have the right to remain anonymous . . .

There is an obligation to reflect on the foreseeable repercussions of research and publication on the general population being studied.

The anticipated consequences of research should be communicated as fully as possible to the individuals and groups likely to be affected.

From the American Sociological Association

Individuals, families, household, kin and friendship groups that are subjects of research are entitled to rights of biographical anonymity . . .

The process of conducting sociological research must not expose subjects to substantial risk of personal harm. Where modest risk or harm is anticipated, informed consent must be obtained.

To the extent possible in a given study, researchers should anticipate potential threats to confidentiality. Such means as the removal of identifiers, the use of randomized responses, and other statistical solutions to problems of privacy should be used where appropriate.

Confidential information provided by research participants must be treated as such by sociologists even when this information enjoys no legal protection or privilege and legal force is applied.

From the American Psychological Association

In planning a study, the investigator has the responsibility to make a careful evaluation of its ethical acceptability. To the extent that the weighing of scientific and human

values suggests a compromise of any principle, the investigator incurs a correspondingly serious obligation to seek ethical advice and to observe stringent safeguards to protect the rights of human participants.

Considering whether a participant in a planned study will be a ''subject at risk'' or a ''subject at minimal risk,'' according to recognized standards is of primary ethical concern to the investigator.

While these professional organizations use slightly different wording for ethical principles, notice that each code states that the responsibility for ethical research practices rests with the individual researcher. IRBs have the responsibility of approving the protocols for research conducted through their institutions, and thus are concerned with the legal implications of noncompliance. As mentioned earlier, IRBs want to avoid possible litigation from ethical violations. Of course, avoiding liability should also be of concern to the researcher since she/he can be personally sued for failing to meet ethical standards.

IX. THE ETHICS OF DATA COLLECTION AND ANALYSIS

Ethical concerns are not limited to the treatment of human subjects, but also arise during the process of the collection, analysis, and reporting of social research data. Learning to design good research assumes that the methods used to collect data will have intellectual integrity and be trustworthy. Once collected, data can be manipulated in various ways that undermine the aims of social science. The expectation in social research is that the data be collected and interpreted ''objectively.'' But interpretive objectivity can be compromised by unethical research practices—Babbage (1969) identifies three ways that this can occur during the ''interpretive'' process. One violation occurs when researchers select only those data that fit the research hypothesis—this is referred to as ''cooking'' the data (1969). Another way that objectivity is manipulated is by ''trimming'' the data. This refers to the practice of massaging the data to make them look better (see for example Huff's 1954 classic *How to Lie with Statistics*). The third way that ''objectivity'' can be compromised is by ''forging'' the data—which refers to the fabrication of data. Attempts at replication serve as ''checks'' for faulty research processes, but such attempts may be especially rare in cases involving large-scale research investigations that are prohibitively expensive (Fisher, 1982; Kimmel, 1988). The ethics of scientific investigation are to observe and report all data accurately and completely, even if it means that one of the researcher's treasured theories is threatened by such data.

Data analysis not only makes sense of the data that are collected, but also contributes to the level of understanding on a particular topic. Other researchers use our findings to frame their research; if the analyses are not correct, we have misled others and wasted their time, money, and effort. Equally relevant is that others who may not be researchers, but are in a position to formulate policy, may rely on erroneous results. In his critique of two widely cited studies of rape, Neil Gilbert (1992) distinguishes between what he calls ''advocacy research'' and social science. Focusing on a widely cited figure that one out of every two women will be a victim of rape (from the Koss and Russell research that appeared in the *Ms*. Magazine Campus Project on Sexual Assault), Gilbert combines critical thinking and data from other studies to show that this the figure inflates the prevalence of the problem.

Gilbert argues that Koss and Russell have intentionally distorted the extent of rape to advance an ideological agenda. The kind of research Koss and Russell have done, Gilbert argues is really ''advocacy research'' which is research that operates under the guise of social science in order to persuade the public and policymakers that a problem is vastly larger than commonly

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

thought (1992: 9). Advocacy research uses four techniques to manipulate public perception: 1) by measuring a problem so broadly (i.e. operational definitions) that almost anything would fit the definition; 2) by measuring a group that is at higher risk for the problem and then projecting the results to the general population; and 3) by claiming that smaller studies that define the problem differently, use diverse methodologies, and come up with varying results, form a cumulative block of evidence that supports the current findings; and 4) by any combination of the preceding three points (Gilbert, 1992: 8). Proponents of this approach believe that ''playing fast and loose with the facts is justifiable in the service of a noble cause'' (Gilbert, 1992: 9).

While advocacy studies may serve some useful purpose by bringing serious problems to the attention of policymakers, they do little to elevate our understanding of an issue since data upon which their claims stand are distorted. In the long run, overstating the magnitude of a problem and manipulating the conceptualization and operationalization of a problem to include almost anything ultimately trivializes it. Advocacy research is nothing more than a foil for an ideology.

Social scientists have the obligation to promote knowledge regardless of the source of that knowledge (i.e. whether it is their own or others). It is also helpful to remember that it is not possible for any researcher to ensure that their research will not be misused or that the methods of social science will not manipulated for purposes other than advancing our understanding of an issue. One way to avoid the possibility of misuse is by writing as clearly and precisely as possible. Clear writing makes it less likely that others can misinterpret results and conclusions and clear writing makes it more likely that the limitations of research are understood—making it more likely that the misuse of information can be detected by others. Advocacy research justifies distortion tactics by arguing they are necessary to get an issue on the policy agenda. While advocacy research may draw attention to an issue, distortion obfuscates understanding and may actually undermine public support.

X. ETHICAL DILEMMAS IN APPLIED SETTINGS

When researchers conduct studies in organizational or other real-life settings, they usually encounter ethical problems that are almost solely political in nature. According to Carol Weiss, ''[S]omething else besides research is going on; there is a program serving people'' and the research is only an appendage of the situation (1972: 92). In short, the applied researcher works as a ''hired'' gun for an organization, and in this capacity, the applied researcher is expected to promote the interests of that organization—applied researchers are ''advocates'' for the policies of the organization. One thing that should be factored into evaluation is that the program being evaluated cannot be held constant—the actions being observed are ''in progress'' which means that a combination of internal and external forces come to bear on activities as they occur. Inevitably the applied researcher must try to balance the dynamics of the setting while at the same time collect reliable data. The applied researcher should never forget the fact that any given organization is part of a larger organizational system, the nature of which will impact outcome.

Evaluation research is likely to present a number of ethical dilemmas for the researcher. Evaluation results are used to justify decisions about the expenditure of resources. Thus, evaluation results can impact decisions about a program's future—whether it should be continued or stopped or whether its budget and personnel should be increased or cut back. In applied settings, there are a number of vested interests at work. At the most basic level, evaluation research poses problems related to whose interests are being served and whose point of view should be represented during the evaluation. It should be kept in mind that collateral interests and points of view are likely to be independent of the aims of an evaluation—such interests tend to reflect

the social and political institutions to which programs (and thus program evaluations) are attached.

XI. THE USES OF SCIENTIFIC RESEARCH

Ethical dilemmas can also arise after a study has been completed, for example, when knowledge (research findings) is misused or when widely accepted procedures and principles with proven utility are improperly implemented (e.g. advocacy research). The inappropriate utilization of research findings outside clearly stated boundaries can have serious and far-reaching methodological consequences. Ethical questions arise, for example, when the findings from research that has been supported by private industry are kept from the general public or manipulated to intentionally mislead the general public. Consider the current legal debate over the tobacco industry's deliberate withholding of the addictive effects of nicotine.

There is little question that the ''products'' (i.e. findings), from social science research will be used by others. The results from social research have long been used to support policy decisions. In the Supreme Court decision of *Brown v. Board of Education* of Topeka in 1954, the unanimous opinion of the court cited several studies showing that segregation had a detrimental psychological effect on black children. When research is used to bolster social policy it is reasonable to expect that the data supporting the policy have not been ''cooked, trimmed, or forged.'' Social scientists since Brown have continued to champion the benefits of integration and civil rights, with many testifying in cases involving school desegregation, busing, and affirmative action.

The ethical concern in such social policy debates involves questions about how much responsibility researchers should bear for applications that are destructive or contrary to prevailing scientific and public sentiment. While policy is never formulated in a vacuum, when research findings are used to demonstrate the need for prescriptive measures, the results are expected to be based on objective data. Similarly, the role of the social scientist as researcher is expected to be kept separate from the role of the social scientist as citizen. While one may argue whether a value-free science is possible, objectivity continues to be the *sine qua non* of science, and according to such a view, scientific findings should be nonmoral in their application. The methods of social science are designed to be free of personal biases, preferences, and values. Thus there should be nothing in the findings of scientific work that hints to what purposes the products of that work should be put (Lundberg, 1961). Scientific work should stand on its objectivity. This is not to say that social scientists are detached from their environment. In their role as citizens, social researchers may take ideological and moral positions—opposing nuclear weapons, acid rain, or racial oppression, but such views should not determine how researchers structure their research.

Sociologist Howard Becker (1967) argues that research is not value-free, but rather is always contaminated by personal and political views. That it is though does not mean that researchers should forsake the standards of good scientific work and advocate for one side of a political debate in the name of science. Becker urges us to keep in mind the objective of our work—what it is we are trying to do in our research—which is to understand and explain social/behavioral phenomena (1967).

Also pertinent to uses of scientific knowledge is the issue of the timing of reporting research findings. Is the common good better served when research findings are withheld until they have gone through a peer review process (so that we can have confidence in their validity) or when findings are reported immediately. The early reporting of findings can influence public understanding of an issue while delayed reporting could impact those who might otherwise have

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

benefited in some way from the early reporting of results (Bermel, 1985). Consider the controversy involving the screening of blood during the early years of the AIDS epidemic. Although there was mounting evidence that HIV could be transmitted by transfusion, the blood bank industry refused to acknowledge this risk. Blood bank officials refused to implement screening procedures during the early years of the AIDS epidemic. Only when facing litigation from patients who had contracted the AIDS virus from contaminated blood and pressure from the CDC, did the nation's blood bank industry began to screen donors' blood. A decision was made to deliberately withhold research findings that documented the risk of HIV from blood transfusions by the blood bank industry.

XII. CONCLUSIONS

The primary objective of this chapter has been to raise the ethical sensitivity of those who will be conducting social research and to show the myriad ways in which ethical dilemmas can emerge in the conduct of research. Research ethics present a set of principles against which the actions of researchers (and science) are judged. As is evident from the ethical dilemmas presented here, research ethics do not constitute a hard and fast list of dos and don'ts; rather ethics provide a set of standards that are to be used in the practice of research. Each stage in the research process discussed above presents its own dilemmas for the researcher. In considering the *ethical treatment of human subjects*, researchers are expected to design their studies so as to protect the rights of participants and treat them with respect and dignity. During the process of research, the researcher is expected to be objective and unbiased in conducting research. The methods of research provide the blueprint for insuring objectivity. Researchers are also expected to report their findings honestly and accurately. The social scientific community generally adopts a teleological position with respect to the dissemination and use of scientific knowledge. Thus, research is to be used to promote the general welfare rather than ideology.

NOTES

1. The National Research Service Award Act (Public Law 93-348), signed into law in 1974, created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. As stipulated by this legislation, sponsored research that involves human subjects and DHHS (Department of Health and Human Subjects) funding (all federally funded research) must establish an IRB to assure that ethical standards and research protocols are satisfactorily carried out. Almost every college and university in the United States and most tax-exempt private research foundations have IRBs. Over 90 percent of these IRBs have ''mandated the routine review of ALL proposals, not just those that are, or hope to be, funded (Ceci et al., 1985). In discussing IRB procedures, Gillespie (1987) notes that the codification of ethical principles for research serves to delineate researchers' obligations which spell out one's responsibilities to participants, colleagues and professional audiences, and sponsoring agencies, the public at large, and society (Gillespie, 1987: 503).
2. Citing concerns about the legal implications of using campus computers to distribute obscene material, CMU banned adult oriented bulletin boards from campus computers.
3. The basic elements of informed consent include: an explanation of the procedures used in the research and their purposes; a description of any reasonably foreseeable

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

risks and discomforts to participants; a description of any benefits that may reasonably be expected; a disclosure of any alternative procedures that might be advantageous to the subject; an offer to answer any questions concerning the procedures; and a statement that participation is voluntary and that the participant can withdraw from the study at any time.

REFERENCES

Adair, J.G., T.W. Dushenko, and R.C. Lindsay (1985). ''Ethical Regulations and their Impact on Research Practice,'' *American Psychologist*, 40: 59–72.

American Anthropological Association (1971). *Principles of Professional Responsibility*, Washington, DC: Author.

American Psychological Association (1981). *Ethical Principles of Psychologists*, Washington, DC.

American Sociology Association. (1981). *Code of Ethics*, Washington, DC.

Baumrind, D. (1964). ''Some Thoughts on Ethics of Research: After Reading Milgram's 'Behavioral Study of Obedience,''' *American Psychologist*, 19: 421–423.

Babbage, C. (1969). *Reflections on the Decline of Science in England and on Some of Its Causes*, London: Gregg International.

Baumrind, D. (1985). Research Using Intentional Deception: Ethical Issues Revisited. *American Psychologist*, 40, 165–174.

Becker, H. (1967). ''Outsiders: Studies in the sociology of deviance, New York: The Free Press.

Bermel, J. (1985). ''Prior Publication: Two Research Studies, Two Views,'' *Hastings Center Report*, 15, 3–4.

Diener, E. and R. Crandall (1978). *Ethics in Social and Behavioral Research*, Chicago: University of Chicago Press.

Fisher, K. (1982). The Spreading Stain of Fraud. *APA Monitor*, November, pp. 7–8.

Gelles, R. (1978). Methods for Studying Sensitive Family Topics. *American Journal of Orthopsychiatry*, 48: 408–424.

Gilbert, N. (1992). Realities and Mythologies of Rape. *Society*, 29 (4), 4–10.

Gillespie, D. (1987). ''Ethical Issues in Research,'' in *Encyclopedia of Social Work*, 18th ed., Washington, DC: National Association of Social Workers, pp. 503–512.

Huff, D. (1954). *How to Lie With Statistics*, NY: Norton Publishing.

Humphreys, L. (1970). *Tearoom Trade*. Chicago: Aldine.

Inciardi, J., D. Lockwood, and A. Pottieger (1993). *Women and Crack-Cocaine*, New York: Macmillan Publishing Co.

Katz, J. (1972). *Experimentation With Human Beings*, New York: Russell Sage.

Kelman, H.C. (1967). ''Manipulation of Human Behavior: An Ethical Dilemma for the Social Scientist,'' *Journal of Social Issues*, 21: 31–46.

Kimmel, A.J. (1988). *Ethics and Values in Applied Social Research*, Newbury Park, CA. Sage Publications.

Milgram, S. (1974). *Obedience to Authority*, New York: Harper & Row.

Reese, H.W. and W.J. Fremouw (1984). ''Normal and Normative Ethics in Behavioral Sciences, *American Psychologist*, 28: 134–139.

Sieber, J.E. (ed.) (1982) *The Ethics of Social Research: Surveys and Experiments*, New York: Springer-Verlag.

Singer, E. (1978). ''Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys,'' *American Sociological Review*, 43: 144–162.

Weiss, C. (1972). *Evaluation Research*, Englewood Cliffs, NJ: Prentice-Hall.

Westin, A.F. (1968). *Privacy and Freedom*, NY: Atherneum.

Whyte, W.F. (1979). ''On Making the Most of Participant Observation,'' *The American Sociologist*, 14, 56–66.

Vidich, A.J. and J. Bensman (1958). *Small Town in Mass Society: Class, Power, and Religion in a Rural Community*, Princeton, NJ: Princeton University Press.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

3
Levels of Data, Variables, Hypotheses, and Theory

Marcia L. Whicker and Gerald J. Miller
Rutgers University, Newark, New Jersey

This chapter briefly examines levels of data, variables, hypotheses, and their linkage to theory. While these concepts are not statistical themselves, they are crucial to effective use of statistics and research methods.

I. THE RESEARCH PROCESS

Suppose you were going to build a house. What would you need? In many ways the empirical research process is analogous to house building (see Table 1).

To start, you would need a plan—a *blue print* of what to put where so that when the house was finished, doors would shut, closets would be in place, plumbing would be located in crucial areas of the house, the floors would be level, windows would be in the proper places, a heating and cooling system would be installed, and stairs would connect floors. Without proper architectural plans, the location of key features of the house would be haphazard and key features may be jerry-rigged after the fact. The process of developing the house plans causes the builder and prospective home owner to think through what kind of house is desired and how it should look and function before the building starts.

Similarly, in the research process, planning is a key aspect of the research outcome. The plan is called a *research design*, and it is as crucial to the quality of the final study that is produced as are architectural blue prints to house building. Without a research design, steps may not be taken that are necessary to assure that controls have been put into place extraneous factors and, when possible, spurious relationships have been addressed. Proper sampling, randomization, and the development of control groups may be specified in a well done research design to allow the researcher to test for causation as well as for correlations. Each of these steps, just as carefully thought out placement of architectural features in a blue print, increases the quality of the final product.

But the architectural blue prints alone are insufficient to create an aesthetically pleasing, strong, and functionally useful house. The *builder* must know how to negotiate all the snags and pitfalls that may occur in the house building process, from labor issues, subcontractors, broken machinery, weather delays, and choices about the actual building. When the builder is knowledgeable, reliable, professional, competent, and trustworthy, the likelihood that the final

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 1 House Building Analogy for the Empirical Research Process

House building	Research process
Architectural blue print	Research design
Competent builder	Qualified researcher
Building tools	Statistics
Building materials	Data

outcome of the house building process will be good goes up. Similarly, the *researcher* is an important aspect of the research process. Plainly the skill and competence of the researcher affects how well a research study is designed and implemented, as well as its usefulness once it is completed. Just as the builder makes many decisions that involves an element of discretion, each with bearing on the final outcome, so too does the researcher.

Another element in how quickly and effectively a house is built is the *tools* with which the builder works. Someone constructing a house with a pick, shovel, and hammer will take much longer and likely have a much rougher product than someone using a backhoe, earth moving equipment, and power tools. For the research process, statistics and forms of analysis are the equivalent of the builder's tools. Just as powerful tools for the builder facilitate the building process and typically improve the outcome, high-powered multivariate *statistics* may facilitate a research study and improve the findings by making the outcomes more clear. Because powerful multivariate statistics allow for both competing and complimentary influences to be considered simultaneously, using them may also make the research results stronger.

Finally, the quality of a house is highly dependent upon the quality of the *materials* used to build it. If higher grade building materials are used, the resulting house will be superior to one where inferior low-quality materials are used. Thus, high grade lumber, stone, tiles, marble, plaster, durable and attractive fixtures, and other high quality materials result in a better product than do cheaper clapboard, linoleum, plasterboard, plywood paneling, and inexpensive roofing, heating and cooling systems, and plumbing. Similarly, the quality of *data* used in a research project is equivalent to the quality of materials in building a house. Not all data are created equally. High level data is of better quality in many ways than lower level data. One advantage of high level data is that more powerful multivariate tools may be used upon it, while lower level data require more awkward and less powerful statistical tools.

Just as the quality of each of the above house building elements affects the quality of the final finished house, so does the quality of each of the equivalent research process components affect the final research product. Various aspects of research design, including the differences between experimental, quasi-experimental, will be discussed elsewhere. This chapter will discuss levels of data.

II. LEVELS OF DATA

Data are the basic material of empirical research. Data result from observations of real world phenomena. Data are measurements that represent the operationalization of a concept. Recording repeated observations of the same concept across different subjects or cases is how a variable is created. Data range in level from low to high. The lowest level of data is nominal or categorical data. Other levels in ascending order are ordinal or ranked data, interval data, and ratio data. The properties of levels of data are cumulative, so that each higher level of data has the character-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 2 Important Characteristics of Data

Data are empirical observations of real world phenomenon.
Data represent the operationalization of variables.
The properties of levels of data are cumulative.
High level data are preferred to low level data.
Record data at the highest possible level.
Low level statistics can be used on collapsed high level data but high level statistics cannot be used on low level data.

istics of the level of data immediately below it, plus some additional characteristics. Generally high level data are preferred to low level data. More high-powered statistics can be used on high levels of data, but not on lower levels of data. Further, high levels of data can typically be collapsed to lower levels of data after the measurement has been recorded but the reverse is not true; that is, low levels of data cannot be elevated to high levels of data after the observation has been recorded. Researchers then are typically encouraged to initially record data at the highest possible level to retain the greatest flexibility and power (see Table 2).

A. Nominal or Categorical Data

The lowest level of data is nominal or categorical data (Blalock, 1979). Nominal data consists of classification of observation and subsequent placing each observation into an unambiguously defined category. The observations in a category are homogenous with respect to each other. Observations in different categories are heterogeneous. Categories should be constructed to be both mutually exclusive (each observation can clearly and unambiguously be placed in one category or another, not two categories at once) and exhaustive (cover the entire set of possible categories into which an observation may be placed).

Nominal data have the mathematical principles of symmetry and transitivity. Symmetry implies that if A = B (A is in the same category as B) then B = A (B is in the same category as A). Transitivity means that if A = B and B = C, the A = C. (If A is in the same category as B, and B is in the same category as C, then A is in the same category as C). The mathematics of elementary set operations may be applied to properly constructed nominal data (see Table 3).

Examples of nominal data include placing employees (observations) in categories based on support or lack of support for a management innovation (the employee supports or does not support the innovation), the agency division in which an employee works (personnel, budgeting, field operations, etc.), or the personal background characteristics of the employee (gender, race,

TABLE 3 Levels of Data

Level of data	Key characteristic	Additional mathematical properties	Mathematical operation permitted (cumulative)
Nominal	Categories	Symmetry transitivity	Set operations
Ordinal	Ranking	Direction	Mathematics of inequality
Interval	Equal intervals	Distance	Addition/subtraction
Ratio	True zero	Magnitude	Multiplication/division
Dichotomous	All of the above	All of the above	All of the above

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

marital status and religion). Even with relatively simple categorizations, however, sometimes the researcher must make judgment calls. Should a separate category be added to marital status for people who are legally married but separated from their spouses? What about people of mixed racial backgrounds who do not identify only with the race of either parent? Or what about people from mixed religious upbringing who practiced both religions? And how shall the employee who shifted from field operations to the budget office during the study period be classified? Avoiding such judgment calls is not possible. It is important for the researcher, however, to be consistent in whatever decision he or she makes about how to deal with such cases, treating all cases consistently. As always, honesty in research is a good idea. The researcher should specify how any judgment calls in classification were made and the rationale for the process used.

B. Ordinal or Ranked Data

Ordinal data is ranked data. The data have an order to them and fall along an underlying dimension. The rankings of ordinal data may be so precise that each case has its own unique rank. An example of this would be class rank for graduating seniors, or a listing of the top 25 national universities by individual rank. The top ten songs or best selling books in any given week ranging from most popular to 10th most popular are also examples of individually ranked observations. Alternatively, ordinal data may also consist of ranked categories. An example would be classifying individuals by social class as lower, lower middle, middle, upper middle and upper class. Survey responses along an ordinarily ranked scale also constitute ordinarily ranked categorical data. An illustration would be coding respondents by their answer to a question where the possible answers are strongly disagree, disagree, neutral, agree, strongly agree.

In addition to the mathematical properties of symmetry within categories attributed to nominal data, ordinal data is asymmetric in relation to the underlying dimension. Asymmetry implies some relationships, especially relationships of inequality, hold for A that do not hold for B. For example, if $A > B$ (A is greater B), then it is not true that $B > A$ (B is not greater than A). Transitivity holds for ordinal data as well as nominal data, so that if $A > B$, and $B > C$, then $A > C$. The mathematics of inequality, then, apply to ordinal data, as well as some principles of set mathematics that apply to nominal data.

C. Interval Data

Interval data or measurement adds the concept of distance to that of direction that ordinal data embody. Interval scales have equal intervals or distances between measurable points on the scale, making addition and subtraction possible. Length has meaning, so that units may be added or subtracted to a starting point. Man-made scales such as IQ and some temperature scales (Fahrenheit, etc.) are examples of interval data.

D. Ratio Data

Ratio data has a true zero point, as well as the characteristics of interval data. A true zero point embodies the concept of magnitude and allows the mathematics of multiplication and division. Examples include income and weight. In practice, distinctions between interval and ratio data are more theoretical than practical, and in terms of picking statistical tools to use, interval and ratio data are often treated as the same.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

E. The Special Case of Dichotomous Data

Dichotomous data has two categories, 0 and 1. Often dichotomous data is generated by coding an observation 0 if it lacks a particular characteristic, and 1 if it manifests that characteristic. Hence, a state may be coded 1 if it has a particular law, such as term limits on state politicians, and 0 if it does not have that law. Similarly, an individual may be coded as 0 if he or she is not a college graduate, and 1 if he or she is a college graduate. Dichotomous data is obviously nominal level data—a variable with two categories—so that some set operations apply. Technically, however, dichotomous data also meet the requirements of all levels of data, as well as nominal data. Since a code of 1 implies more of the characteristic in question than a 0, the coding scheme embodies direction and therefore meets the requirements of ordinal data, so that the mathematics of inequality apply. The distance between 0 and 1 is an interval. Because there is only one interval, the requirement of interval data for equal intervals incorporated into the measuring scheme is met. Hence, the mathematics of addition and subtraction apply. And finally, with one category, that of 0, implying the total absence of the characteristic in question, dichotomous data has a true zero point—none of characteristic, so the requirements of ratio data are met.

This versatility of a dichotomous coding scheme has led researchers to try to convert nominal variables with more than two categories to a series of dichotomous variables. Once dichotomous data are obtained, many (but not all) higher level statistics that require interval/ratio data can be used without causing bias or violating the assumptions of the statistical tool. The resulting dichotomous variables are called ''dummy variables.''

One common application of the dummy variable creation process is in multiple regression, when one of the independent variables of interest to the researcher is a multi category nominal variable. In such instances, converting the nominal variable with several categories to a series of dichotomous dummy variables allows the latter to be used in regression equations as independent variables in the regression model. Hence, variable of religion, with categories for various religions would be converted from a single variable called ''religion'' with several categories (Protestant, Catholic, Jewish, Islamic, Buddhist, Mormon, etc.) to a series of dichotomous variables, each named for the former category of a particular religion. Cases would be coded 0 or 1 on a variable named Protestant, based on whether or not the individual was a Protestant. Similarly, each case would be coded 0 or 1 on a variable called Catholic; 0 or 1 on a variable called Jewish; 0 or 1 on a variable called Islamic, etc.

The advantage of this conversion is that higher level statistics can now be used without having to abandon the concept of religion, merely because it is a multicategory nominal variable. The disadvantage is that with dummy variable creation, the single concept of religion has been converted to a series of variables that ''get at'' the concept of religion, but where no single variable contains as much information as the previous multicategory nominal variable. Interpreting the results of analysis using dummy variables is also less compact and often more ''messy'' than using a single variable that contains all the relevant information.

Another example of when dichotomous dummy variables would be created is the concept of race/ethnicity, which normally would be a multicategory nominal variable. Indeed, many demographic variables often lend themselves to dummy variable conversion. In the case of gender, (male or female), no conversion of the variable is needed, since the initial variable is already dichotomous. Thus, male may be coded as 0 and female as 1, so the interpretation of the variable becomes the presence or absence of the characteristic of female. If the reverse coding scheme had been used—0 for female and 1 for male—the interpretation would be the presence or absence of the characteristic of male. Which coding scheme is used depends, in part, on the hypotheses being tested, although neither scheme is technically incorrect.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 4 Types of Variables

Variable	Characteristics
Dependent variable	The primary concept the researcher wants to describe, explain, and predict. Values ''depend on'' independent variables. Usually a study has one primary dependent variable. Symbolized by Y and placed on the Y axis of the Cartesian coordinate graph.
Independent variable	Influence and impact the dependent variable. A study may include several independent variables. Symbolized by X, and placed on the X axis of the Cartesian coordinate graph. If there are several independent variables, differentiated by subscripts and symbolized by X_1, X_2, X_3.....X_k
Control variable	A subset of independent variables that impact or influence the dependent variable. Not the primary focus of the researcher. How the researcher addresses control variables depends on whether the research design is observational or experimental. In experimental research designs, may be demographic variables the researcher cannot manipulate. Are not the same thing as control groups which represent an absence of (zero level of) the primary independent variable.

III. TYPES OF VARIABLES

Variables are concepts that have been operationalized. Operationalization is the specification of unambiguous measurement procedures that, when applied, result in a numerical value for the concept for each case or observation in the study. These values are data that are either nominal, ordinal, interval, or ratio level.

The terms independent and dependent variables have already been used. These are two big categories of variables (see Table 4). Unlike levels of data, a concept is not inherently one type of variable or the other. Rather, whether a variable is independent or dependent depends on the research question and hypotheses. In one study, a concept may be an independent variable, while in another study, that same concept may become a dependent variable.

A. Dependent Variable

A dependent variable is a concept that is impacted or influenced by other variables in the study. Those other values are independent variables. The values of the dependent variable depend upon the values of the relevant independent variables. The goal of science is to describe, explain, and predict an important concept. The dependent variable is typically the concept the researcher is trying to describe, explain and predict. Mathematically, the dependent variable is often symbolized by the letter Y, and is displayed graphically on the Y axis of a Cartesian coordinate system.

B. Independent Variables

Independent variables may be symbolized by X and displayed on the X axis of a Cartesian coordinate system. If more than one independent variable is used, each X variable may get its

own subscript, so that X_1 becomes the first independent variable, X_2 the second independent variable, X_3 the 3rd independent variable, and so forth up to X_K, the kth independent variable. If a researcher is trying to explain student achievement, student achievement becomes the dependent variable for that study. The researcher will posit that various independent variables are likely to impact or influence student achievement, so that the level of student achievement obtained ''depends on'' those factors.

Factors impacting student achievement that may become independent variables in the study include demographic characteristics of the student, such as social class, gender, race or ethnicity, and family income. Other factors that may impact student achievement and therefore may also be independent variables in the study are characteristics of the school the student attends. School characteristic variables may include such things as class size, school organization, the presence or absence of programs for gifted students, the presence or absence of academic tracking, and teacher salaries. Yet other factors may be characteristics of the classroom to which the student is assigned, including amount of homework, type of classroom interaction, and teacher expectations about student achievement.

C. Research Questions Asked and Real World Complexity Determine Dependent and Independent Variables

Plainly, many factors potentially affect student achievement. Similarly, in the typical research study, the number of independent variables may be quite large, while the study may have one major dependent variable. In yet another study, what was an dependent variable in the first study may become an independent variable. Suppose a researcher is interested in predicting whether or not a college student completes four years of college. Student achievement in high school may be one independent variable that is examined for its influence on college completion. Or, perhaps a researcher would like to predict salaries in the first ten years of employment after graduation. Again, student achievement may become an independent variable tested for its impact on salaries in early career years.

D. Control Variables

Control variables are yet another type of variable in a research study, although control variables are actually a subset of the independent variables. Control variables are factors that the researcher suspects may be linked to and impact the dependent variable. In most instances, especially when an experimental research design is used, the researcher may not be primarily interested in exploring this influence or impact. Rather, the researcher is interested in the linkages of the main independent variables to the dependent variable. If the researcher did not address control variables in the study, however, their influence on the dependent variable may confound the inquiry into the main independent variable-dependent variable linkage, and may even cause the researcher to make misleading and/or erroneous conclusions about those linkages. The skilled researcher, then, somehow includes the control variables in the study.

1. Incorporating Control Variables into Observational Research Design Through Measurement and Statistical Analysis

How control variables are included depends on the type of research design (see Table 5). If the research design used is an observational design, where the researcher has little ability to manipulate the independent variables, control variables may be measured, and their impact ascertained statistically. In such cases, the distinction between primary independent variables and

TABLE 5 Approaches for Addressing Control Variables

Type of research design	Approach
Observational design	Measure and include control variables as additional independent variables.
Experimental design	Distribute values of control variable evenly between experimental and control groups. Either: Use randomization to distribute control variable by randomly selecting and randomly assigning cases to experimental and control groups. Match cases in experimental group with cases in control group on key values of control variable. Restrict cases used in the study to one category of the control variable. Build control variable into the study as an additional independent variable.

control variables is very blurred. Both are measured for each case or observation, and the impact of both on the dependent variable is explored using multivariate statistics. The primary distinction between the two is the interest of the researcher and the emphasis the study places on each.

2. Incorporating Control Variables into Experimental Research Designs

If, by contrast, the research design is an experimental design, the researcher has more options for how to address control variables.

E. Equalizing the Distribution of Control Variables for all Groups in the Experiment

One approach is to try to assure the distribution of the control variables is the same for all of the groups in an experiment. The researcher may assume that random selection and random assignment of cases to the experimental and control groups in the study will cause any significant control variables to be evenly dispersed between the two groups, causing the impact of the control variables on the dependent variable values in the experimental group to be the same as the impact of the control variables on the dependent variable values in the control group. Notice that control variables and control groups are not the same thing! Control variables are a subset of independent variables that likely impact the dependent variable. Control groups occur in experimental designs. Control groups are those that do not get the ''treatment variable''—that is are exposed to a zero level (absence of) the primary independent (treatment) variable. A control group receives a zero level of the primary independent variable. Control variables may occur in both experimental and nonexperimental research designs. Control groups occur only in experimental designs, since only in experimental designs are groups used to structure independent variable levels.

A version of this approach is matching in the selection process when cases are picked for the control group versus the experimental group. Cases are matched on key control variable values. If, for example height is city size is a control variable, every time a large city is picked to be assigned to the experimental group, a similarly large-sized city is picked for assignment to the control group. When a small city is picked for the experimental group, a similarly small city is picked for the control group. The even distribution of the key control variable values between experimental and control groups is not left to random chance, but is explicitly matched by the researcher. This approach to assuring an even distribution of the control variable to both

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

the experimental and control groups is particularly appropriate when the population from which cases for participation in the experiment is small, and when the values for the control variable are lumpy, or not a smooth continuous distribution.

F. Eliminating Some Categories of Control Variables from the Experiment

A second way of handling control variables in an experimental design is to eliminate cases that exhibit one or more categories of the control variable from the research study. Suppose a researcher thought that gender was a control variable for a study examining drug effectiveness, and that without controlling for gender, some outcomes the researcher would otherwise attribute to particular regimens of drug therapy would actually be caused by gender differences. The researcher may choose to eliminate this confounding effect by eliminating one of the gender categories from the study. For example the study may be conducted on only women, or only on men, so that all the cases or observations in the study were of the same gender. This description of how various drug therapies have been tested is not far from reality as many medical studies have been conducted on men only. Women's groups in the United States have made this a political issue. While conducting studies on only one gender may make for good science, it has political ramifications in that conclusions that are valid for men may not hold for women. This method of addressing control variables, then enhances the internal validity of a research study (ability to address causal questions), while decreasing the external validity (ability to generalize beyond the study to broader populations).

G. Including Control Variables as Additional Independent Variables

A third approach for dealing with control variables in an experimental design is the same approach used for observational studies: build the control variable into the research study as another independent variable. The main distinction between the primary independent variable and the control variable, then, is the researcher's interest and the hypothesis. In such cases, control variables are often demographic characteristics the researcher cannot manipulate in the context of the research study, while the primary independent variable in an experiment is subject to manipulation by the researcher. The research can administer different levels of the primary independent variable (the treatment variable) to study participants (cases), but can only measure, not manipulate the control variable or variables. This is most explicit method for dealing with control variables and directly tests the impact of the control variable on the dependent variable as well as the impact of the primary independent variable, this approach has the disadvantage of increasing the number of groups that need to be in the experiment, and implicitly, increasing the number of cases needed and the overall research design complexity. The experiment becomes more expensive and difficult to manage.

The number of groups needed in an experiment is a function of both the number of independent and explicitly built-in control variables, as well as the number of values (categories) each of those variables may take on. Assume there is only one primary independent variable, a drug, that is to be administered or not administered. Then two groups are needed: an experimental and a control group. Often tests of program effectiveness have two groups. Either a case is in the program and therefore the experimental group, or else not in the program and in the control group.

If the number of values the primary independent variable can take on increases from two to three, the number of groups increases from two to three as well. Suppose in the drug test experiment, the researcher wishes to test a high dosage level and a low dosage level. Now two experimental groups are needed, one for each dosage level, as well as a third control group that

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

gets no drug at all. Similarly, for the government program, two different versions of the program may be tested, increasing the need to three groups: the intensive program group, the regular program group, and the control group.

What if a control group of gender is added? Gender may be one of two categories—male or female. In the simple version of the drug test experiment, the number of groups needed has now gone from two to four (calculated by multiplying the number of categories of each independent and control variable in the experiment—or two treatment levels . . . drug/no drug, (2) times two genders (male/female). $2 \times 2 = 4$ groups: One group of males that gets the drugs, one group of females that gets the drugs, one group of males that gets no drugs, and one group of females that gets no drugs. Similarly, for the simple version of the program effectiveness study, the number of groups needed increases to four: one group of males in the program, one group of females in the program, one group of males not in the program, and one group of females not in the program.

If we shift to the version of the experiments that has two dosage or program levels as well as a control group (three initial groups), adding in the control variable of gender increases the number of groups needed to 6: (3 levels of the main independent variable—high level, low or moderate level, and no level in the control group), and 2 genders (male and female). $3 \times 2 - 6$ groups needed: males in the intensive program, females in the intensive program, males in the regular program, females in the regular program, males in the control group getting no program, and females in the control group getting no program.

What if we add in a second control variable? The number of groups needed rises rapidly. Suppose the second control variable is race, categorized as white or nonwhite (i.e. two categories). Now, for the program effectiveness test, the number of groups increases to 12. This is based on program levels (intensive, regular, none), two genders (male, female), and two racial categories (white, non-white): $3 \times 2 \times 2 = 12$ groups needed. If we wished to refine our racial categories to white, black, hispanic, and other (four categories), the number of groups needed in our experiment balloons even further: $3 \times 2 \times 4 = 24$ groups needed. Adding in yet another control variable similarly causes the number of groups needed to increase greatly. Since each group must have a number of participants, very soon, the experimental design becomes unwieldy.

IV. TYPES OF HYPOTHESES

Hypotheses are empirically testable statements about relationships between concepts. When tests of the hypotheses are implemented, the concepts are operationalized into variables. Data for the variables for the observations (cases) in the study are collected and organized into a data set. Statistical tests are then conducted to ascertain whether or not the relationships posited in the hypotheses are observed in the particular data set being examined. If the posited relationship is observed, the hypothesis is supported. If the posited relationship is not observed, the hypothesized relationship is not supported.

Hypotheses may be distinguished on several dimensions (see Table 6).

A. Correlational vs. Causal Hypotheses

The type of research design employed determines whether or not a hypothesis is correlational or causal (King, Koehane, and Verba, 1994). This aspect of hypotheses has to do with the degree of causation the hypothesis imputes. Observational designs may have high external validity (the capacity to generalize from the study results to a broader population, because the selection of

TABLE 6 Types of Hypotheses

Aspect of hypotheses	Type of hypotheses	Characteristics of hypotheses
Testing causation	Correlational hypotheses	Used with observational research designs. Can only specify covariation between variables.
	Causal hypotheses	Used with experimental research designs. Can test causal linkages between independent and dependent variables.
Specifying direction	Nondirectional hypotheses	Only specifies that a linkage exists between X and Y.
	Directional hypotheses	Specifies the nature of the linkage between X and Y: For nominal data, specifies which categories of X will be disproportionately linked to which categories of Y; For higher level data, specifies either a positive/direct relationship (X and Y covary in the same direction) or a negative/inverse relationship (X and Y covary in opposite directions).
Formal statement of hypotheses	Statistical hypotheses	Formal statements to test for significance: The null hypothesis, H_O, is always that there is no relationship between X and Y The alternative hypothesis, H_A or H_1, is always that there is a relationship between X and Y. H_A may be either nondirectional or directional.
	Research hypotheses	Stated as the main focus of the research study. Stated that there is a relationship between X and Y. Not restricted to just tests of significance, but also used to test for association/correlation and when appropriate, causation.

study participants has been random, and the study participants are representative of the broader population). Experimental designs have high internal validity (the capacity to prove causation and conclude that changes in the independent variables cause changes in the dependent variable). If the research design is observational, the researcher cannot test for causation. Only correlational hypotheses that posit that concepts covary (move at the same time) can be tested. If the hypothesis is supported by the data, the researcher can only conclude he or she has found evidence of covariation. If the research design is experimental, however, causal hypotheses can be tested. Finding support for the research hypothesis allows the researcher to conclude he or she has found evidence that changes in the independent variables may cause changes in the dependent variable.

The capacity to test causal hypotheses and prove causation must be built into the research design (Campbell and Stanley, 1963). Three conditions must be present for a research design to prove causation (see Table 7). First, the researcher must have the capacity to manipulate (administer) the primary independent variable, so that there is no doubt about when the independent or ''treatment'' variable is given to participants, and in what intensity levels. In proving that changes in X caused changes in Y, the changes in X must precede the changes in Y in time. If the researcher manipulates X, this condition can be met. Second, there must be a control group that does not receive the primary independent variable as well as the experimental group that does. And third, the researcher must randomly select participants, and once selected, randomly assign them to the groups in the experiment to make the effect of control variables has

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 7 Prototype Research Designs

Research design prototype	Research design characteristics
Observational designs	Random sampling provides *high external validity*: Ability to use inferential statistics and significance tests to generalize beyond the study to the larger population. Can only test correlational hypotheses, where variables are tested for covariation, not causation.
Experimental designs	Has *high internal validity*: Ability to test causal hypotheses that changes in X caused changes in Y. Must have 3 features to test causation: 1. Researcher must manipulate X to assure X precedes Y in time and to know the intensity levels of X; 2. Presence of a control group that is not exposed to X as well as an experimental group that is exposed to X; 3. Random selection and random assignment of cases to the experimental and control groups to eliminate spurious relationships by controlling for other factors that impact Y.

been removed/addressed by evenly distributing the various levels of control variables to both the experimental and control groups. This allows the researcher to remove the possibility of ''spurious relationships.''

A spurious relationship occurs when a researcher observes covariation, and erroneously imputes causation into the relationship, when none, in fact exists in that relationship. For example, a researcher may observe covariation between X and Y (that they vary or move together). A spurious relationship would exist if the researcher erroneously leaped to the conclusion that changes in X caused changes in Y, when in reality, that did not occur. Rather, in reality, both the changes in X and the changes in Y may be caused by changes in some third variable, Z. What random selection and random assignment of cases to the experimental and control groups do is assure that values for Z will be more or less evenly distributed in the control group, and more or less evenly distributed in the experimental group. This removes the effect of Z from observations about the impact of X on Y. If, after the experiment, Ys for the cases in the experimental group have a different mean than Ys for the cases in the control group, the researcher knows these differences were not caused by Z, since both groups had the same distribution of Z before and during the experiment.

B. Nondirectional vs. Directional Hypotheses

Direction has to do with how specific is the hypothesis about the character of the relationship. If the hypothesis is not specific, it will merely assert that one variable, X, is linked to Y. This is a nondirectional hypotheses since no direction is implied. An example of a non-directional hypothesis for nominal data would be to hypothesize that gender is linked to political party preference. An example of a non-directional hypothesis for higher level data would be to hypothesize that family income is linked to levels of educational attainment. In each case, a relationship is posited, but not the particular character or nature of the relationship.

Directional hypotheses are more specific, and not only specify that X and Y are linked (co-vary), but how that covariation occurs. For nominal data, a directional hypothesis will specify which category of the independent variable (X) is expected to be linked disproportionately to which category of the dependent variable (Y). To make the above hypothesis linking gender

and political party preference directional, we would hypothesize that women are more likely to prefer the Democratic party, while men are more likely to prefer the Republican party.

For higher level data, directional hypotheses may specify a positive or direct relationship where X and Y both change in the same direction. With a positive or direct relationship, as X increases, Y increases. Similarly, with a positive or direct relationship, as X decreases, Y decreases. Alternatively, a directional hypothesis with variables that are higher level data may also be negative or inverse, so that X and Y move in opposite directions. With a negative or inverse relationship, as X increases, Y decreases. Similarly, with a negative or inverse relationship, as X decreases, Y increases.

To convert the above hypothesis linking family income to levels of educational attainment from a nondirectional hypothesis to a directional hypothesis, we might hypothesize a positive relationship: that as family income increases, levels of educational attainment also increase. An example of a directional hypothesis specifying a negative or inverse relationship would be to hypothesize that drug use is negatively related to levels of educational attainment—specifically, as drug use increases, levels of educational attainment declines.

C. Statistical vs. Research Hypotheses

Whether a hypothesis is stated for a formal test of significance between two variables, or is stated as the major research hypothesis is also germane. Significance testing in inferential statistics requires a formal statement of hypothesis. The null hypothesis (H_0) is always that there is no linkage or relationship between X and Y. The alternative hypothesis (H_A or H_1) is always that there is a linkage or relationship between X and Y. Significance tests allow the researcher to conclude whether, given the number of subjects and what is known about the research setting, including the estimate of the standard error, is the observed relationship between X and Y big (strong) enough that it is not likely to be caused by sampling error? The probability of Type I error (α) and the probability of a Type II error (β) in making that conclusion are associated with significance testing. Accurate estimates of these probabilities requires that the formal hypotheses be set up so that the null is always that there is no relationship between X and Y, and the alternative hypothesis that there is a relationship between X and Y. With this setup, a Type I error is rejecting a true null hypothesis, and concluding that there is a relationship between X and Y, when there is not, and the observed relationship is just caused by sampling error. A Type II error is accepting a false null hypothesis, and concluding that there is no relationship between X and Y, when the lack of a strong relationship in the data is just sampling error, and there is, indeed a real relationship between X and Y. Normally, there is a trade-off between lowering the probability of Type I error (α) and the probability of a Type II error (β). Setting the chosen acceptable level for α automatically results in an associated level for β. When a small (stringent) α is chosen so that the researcher will accept only a small probability of making a Type I error, automatically a large β is set, causing a high probability of making a Type II error. Similarly, when a large (less stringent) α is chosen so that the researcher will accept only a large probability of making a Type I error, automatically a small β is set, causing a low probability of making a Type II error. The only way to lower both α and β simultaneously is to increase the sample size.

By contrast to this formal hypothesis statement in significance testing which follows the formal structure of hypothesis testing so as to retain accurate estimates of α and β, the research study hypothesis is about the substance of the research study. The research study hypothesis is almost always about some type of linkage between X and Y (rather than the absence of a linkage), and therefore usually follows the structure of the alternative hypothesis in formal significance testing. Usually, in scientific journals and other scientific reports, researchers assume

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 8 Types of Statistics

Basis for categorization	Type of statistics	Characteristics
Number of variables	Univariate statistics	Use one variable, X.
	Bivariate statistics	Use two variables, X and Y.
	Multivariate statistics	Use several variables, typically X_1, X_2, X_3 . . . X_k and Y.
Generalizability	Descriptive statistics	Do not generalize; describe a population.
	Inferential statistics	Generalize from sample to population.
Underlying distributions	Non-parametric statistics	Do not assume the normal distribution. Less restrictive assumptions. Used less commonly.
	Parametric statistics	Assume the normal distribution. More restrictive assumptions. Used commonly.
Questions answered	Measurement	Proportions, percentages, ratios Measures of central tendency: Mode, median, mean. Measures of dispersion: Range, variance, standard deviation. Multivariate measurement: Scaling; cluster techniques, factor analysis.
	Statistical significance	Z, t, f, chi-square, binomial distributions.
	Association	Yule's Q, Goodman and Kruskal's tau. Correlation: Spearman's rho, Pearson's R. Regression coefficients.
	Direction	Pearson's R, regression coefficients.
	Prediction	Regression.

readers will understand the formal logic of hypothesis testing for statistical significance, and do not bother to state the formal hypotheses associated with significance testing. Rather, most journals and scientific reports state the research study hypothesis or hypotheses and dwell on the substantive implications of supporting or not supporting that. Research study hypotheses may be used in testing correlation and prediction and larger models, as well as statistical significance between X and Y.

V. SELECTING APPROPRIATE STATISTICS:

How does a researcher select the appropriate statistics or research tools to use in a study? The level of data at which the variables in the research study are measured, as well as the questions the researcher is trying to address determine which statistic or set of statistics is most appropriate. Statistics may be differentiated or categorized by several characteristics (see Table 8).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A. Number of Variables

One way to distinguish groups of statistics is by the number of variables each handles. Simple statistics accommodate fewer variables. More complex statistics accommodate several variables simultaneously. Univariate statistics are applied to only variable. Univariate statistics have the advantage of being simple, and are usually easily understood, even by those who do not have substantial formal training in research methods and statistics. Bivariate statistics are applied to two variables. They have the advantage of allowing for an independent as well as a dependent variable in the formal analysis. Multivariate statistics accommodate several variables at once. Multivariate statistics allow for the impact of several independent variables on the dependent variable to be examined simultaneously, or for clustering patterns across several variables to be explored. Rarely do multivariate statistical models accommodate several dependent variables simultaneously. Since multivariate statistics allow for greater complexity as well as statistical controls to be used, researchers favor their use whenever possible (Babbie, 1990).

B. Generalizability of Results

Another way to distinguish statistics is by whether or not the purpose of the statistic is to generalize the analytic results beyond the data at hand. Descriptive statistics do not generalize, and are used to describe an entire population. Inferential statistics are developed to generalize from a sample to a larger population (Babbie, 1995). The statistical models for descriptive statistics and inferential statistics are often very similar, and at times, even identical. A primary difference is that inferential statistics assume random sampling, and therefore a knowable and calculable standard error to measure sampling error. If sampling error can be reasonably and reliably calculated, then the researcher can make inferences from a sample to a larger population, knowing the probability of making a Type I and Type II error, and able to create a band of confidence around any point estimates. To accommodate this, inferential statistics use degrees of freedom in the calculation of variance and standard error, rather than the total number of cases in the population as descriptive statistics use. Degrees of freedom refer to the number of independent pieces of information used in calculating the statistic, which often is the number of cases minus the number of other statistics that must be estimated from sample data (and therefore depend on the sample itself) to derive the statistical estimate in question.

C. Underlying Statistical Distributions

Statistics also vary according to their assumptions about underlying distributions. Nonparametric statistics do not assume a normal distribution. They have less restrictive assumptions, but are used less commonly. By contrast, parametric statistics do assume the normal distribution. This assumption is more restrictive but due to the law of large numbers as well as the ability to apply parametric statistics to higher level data, parametric statistics are commonly used (Blalock, 1979).

D. Questions Answered

Statistics may be used to answer one of five questions. Some statistics answer only one problem or question, while a few more complex parametric statistics applied to higher level data may answer or yield results for several of the questions.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1. Measurement

One question statistics will address is measurement. Univariate statistics that perform the task of measurement include proportions, percentages, and ratios. Univariate measuring statistics also include measures of central tendency (mode, median, mean), and measures of dispersion (range, variance, and standard deviation). Some multivariate statistics also are used primarily for measurement, including scaling and clustering techniques, such as factor analysis.

2. Statistical Significance

A second question statistics address is statistical significance. Statistical distributions that do this include the Z, t, f, chi-square, and binomial distributions. Statistical significance is the point of the formal hypothesis testing discussed earlier, and answers whether or not an observed relationship in sampling data is strong enough, given the size of the sample and other assumptions, to conclude with an acceptable probability for a Type I error that the observed relationship is real in the larger population from which the sample was drawn and is not caused by random sampling error. Whether statistical significance is found depends, in part, on the sample size. With a very large (random) sample, sampling error is smaller. Even weak observed relationships may be reasonably concluded to exist in the larger population. With a much smaller sample, however, sampling error is much greater. Any observed relationship in the sample must be much stronger for the researcher to reliably conclude that it exists in the larger population from which the sample was randomly drawn.

3. Association

Statistical significance asks whether or not a real relationship exists in the larger population. Statistics that address association ask how strong a relationship is. Usually, the larger the statistic measuring association, the stronger is the association. Correlation coefficients are among statistics that measure association. Yule's Q, Goodman and Kruskal's tau, spearman's rho, Pearson's R, and regression coefficients are all measures of association.

4. Direction

Some statistics, in addition to addressing association, also address the direction of the relationship. If the statistic has a positive sign, the relationship between the observed variables is assumed to be positive. If the statistic has a negative sign, the relationship between the observed values is assumed to be negative. Pearson's R and regression coefficients are statistics that address both the association between two variables and the direction of the relationship.

5. Prediction

A final question or issue statistics will address is prediction. Some statistical tools, including and especially regression analysis in all its many variants are used for prediction. Standard errors become a primary criterion for determining whether or not a statistic is performing well in its predictive capacity (Kleinbaum, Kupper, and Muller, 1988).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

VI. UNIT OF ANALYSIS

The unit of analysis for a research project is the level of social organization at which hypotheses are formed, data are collected, and conclusions are made. The cases or observations are indicative of the unit of analysis. Individuals are often the unit of analysis in social inquiry. Other units include groups, programs or projects, organizations, and levels of government (local, state, national). For both observational and experimental research designs, multiple cases or observations must be included in the study at the unit of analysis. In both observational and experimental designs, cases must be randomly selected to assure external validity (the ability to generalize to the larger population) and randomly assigned to groups in experimental designs to assure internal validity (the capacity to test causal hypotheses).

An example of individual as the unit of analysis would be a study to examine the impact of personality type on achievement. Both the independent variable (personality type) and the dependent variable (achievement) can be observed, measured, and data collected at the individual level. The hypothesis, that personality type impacts achievement, is most appropriate at the individual level, since organizations and higher units of analysis do not have personality types. Personality is a characteristic of individuals. Hence, any conclusion resulting from a study of personality type on achievement would also occur at the individual level. A hypothesis that informal groups have different leadership patterns than formal groups would need to be tested at the group level. If we hypothesized that large cities had program budgets, while smaller cities used only line item budgets, city or municipality would be the unit of analysis. A hypothesis that social unrest was linked to government suppression of civil liberties would most likely be tested at the national level, with data coming from various countries.

Sometimes, in time series analysis, time is the unit of analysis. In a time series study, only one case may be used, but that case may be observed at multiple points in time. Conclusions apply only to that case. For example, the US economy constitutes one case of a national economy. A hypothesis about the growth of that economy across time would likely be tested with time series data, for the US alone.

The ecological fallacy refers to the difficulty of making conclusions about the relationship between the independent and dependent variable at some level other than the unit of analysis. If, for example, a researcher has election data only at the precinct level, as well as data about the racial and ethnic composition of each precinct, to make conclusions about the voting proclivities of various individuals from different ethnic groups would be an ecological fallacy. Ideally, one would collect *individual* level voting data by standing outside the polls with exit interviews or through some other method if one wished to make conclusions about individual voting behavior.

Across time, statisticians have developed procedures to minimize the dangers of making erroneous conclusions when limits to the unit of analysis are violated and the ecological fallacy occurs. The standard statistical process has been to use a technique called Goodman's regression model. More recently, the King approach (1997) has attempted to minimize estimation bias with the ecological fallacy occurs. Such approaches, however, do not eliminate biases that occur from collecting data on observations at one level and inferring from that data conclusions made at a different (usually smaller or lower) unit of analysis. The best advice remains to conduct all data collection and analysis, as well as hypothesis formulation and conclusions at a single unit of analysis.

VII. WHERE DO HYPOTHESES COME FROM?

Defining an appropriate and testable hypothesis is key to a successful research study. Often neophyte researchers wonder: where do hypotheses come from? How does the researcher know

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 9 Criteria for Judging Theory

Criterion	Characteristic
Predictability (effectiveness)	Allows researchers to accurately anticipate and predict dependent concept (variable) outcomes.
Pervasiveness (scope)	Has a broad scope, more generalizability, and wide applicability to a large heterogeneous population.
Parsimony (efficiency)	Uses as few hypotheses, concepts, and variables as can be to attain a particular level of robustness.

which hypotheses to test? Some hypotheses are implicit. For example, in a program evaluation study, the implicit question being addressed is whether or not the program is effective. How effectiveness is measured will vary from program to program and will depend upon the program goals and objectives. Why the program was created in the first place—the underlying rationale for its structure, expenditure, and activities—is presumably linked to some social theory, and testing the program's effectiveness is an indirect test of the underlying theory.

A. Criteria for Judging Theory

Successful hypotheses are not just isolated, but are linked to a larger social theory (Kuhn, 1970). A theory is a set of coherent and consistent propositions (hypotheses) pertaining to a particular phenomenon of interest (dependent variable). Not all theories are created alike. Some theories are better than others. Three criteria, sometimes called the three ''Ps,'' are standard for judging the usefulness of theories: Predictability, pervasiveness, and parsimony (see Table 9).

1. Predictability

Predictability refers to how well a theory predicts the behavior of the primary dependent variable. This criterion is essentially a measure of the effectiveness of the theory. Does it predict well? The theory scores high on this major criterion. If it predicts less well or only sometimes, sometimes allowing researchers and others to anticipate outcomes and other times causing them to predict outcomes that do not materialize, the theory scores lower on predictability.

2. Pervasiveness

Pervasiveness, the second criterion for judging theories as better or worse, refers to the scope of the study. A widely pervasive theory robustly pervades or applies to a large and heterogeneous population. A less pervasive theory is less robust and would apply only to a smaller and more homogenous population. A pervasive theory has a wide scope and is more general, and is therefore judged to be better.

3. Parsimony

The third criterion for judging theories is parsimony. Parsimony refers to the complexity of the theory itself and is an implicit measure of its efficiency. Parsimonious theories have as few hypotheses, concepts, and variables as needed. Nonparsimonious theories are less efficient and contain more hypotheses, concepts, and variables than are needed to attain a particular level of robustness. Everything else being equal, a parsimonious or efficient theory is preferred to one that is not. Theories that become particularly ''jerry-rigged'' with modification and addition upon modification and addition as new data are collected may indicate that a field is particularly

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 10 Approaches to Theory and Hypotheses Generation

Approach	Characteristics
Deduction	From the general to the specific. Applies operational rules to broad assumptions to derive particular hypotheses. May use formal tools of mathematics and computer simulations.
Induction	From the specific to the general. Derives particular hypotheses from observing the behavior of specific cases. Uses the tools of careful watching, listening, and informal observation.

ripe for a ''paradigm shift'' where the fundamental underlying principles of the theoretical foundation of a field are challenged (Kuhn, 1973). A new theory that is more sleek, eloquent and simple with equal or greater pervasiveness and predictability but a radically different perspective and assumptions may arise to quickly sweep away the old theory creaking under its own weight. This rapid alteration in theoretical foundations and core is an episodic and even rare event, but is the part of the large process by which science, including social science, progresses.

Ideally, a theory can be improved on all three criteria simultaneously. In reality, researchers must sometimes make tradeoffs between the three criteria. Improvements in predictability, for example, by the introduction of controls, may lessen the pervasiveness or scope of a theory. Similarly, excessive improvements in parsimony may lessen predictability.

B. Approaches to Theory and Hypotheses Generation

Two approaches to generating theories and their associated hypotheses are induction and deduction (Table 10). Deduction is said to go from the general to the specific. Deductive theorists begin with broad assumptions about human behavior and certain operations or rules of behavior. The operations are applied to the broad assumptions to derive particular propositions or hypotheses. These hypotheses in turn are operationalized and tested in specific settings. The techniques of formal mathematics and computer simulations may be used in deduction.

Induction is said to go from the specific to the general. It involves the development of generalizations (hypotheses) from specific observations. These observations may come out of a researcher's own personal or work experience, or from hearing about the experiences and observations of others. The techniques of careful listening, watching, and observation are the tools of induction.

Science progresses through an alternation of deduction (deriving specific hypotheses from more general theories) and induction (deriving hypotheses from particular observations). Similarly, the skilled researcher also uses both approaches to derive research study questions of appropriate magnitude, realism, and importance.

VIII. CONCLUSION

The complexities of the research process are enormous, yet at its fundamentals, it remains the equivalent of building a house of knowledge instead of physical materials. Once we are sheltered by proven theories from the storms of ignorance, unknown, and uncertainties, life improves.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

REFERENCES

Babbie, E. (1995). *The Practice of Social Research*, 7th ed., Belmont, CA: Wadsworth Publishing Co.
Babbie, E. (1990). *Survey Research Methods*, 2nd ed., Belmont, CA: Wadsworth Publishing Co.
Blalock, H.M., Jr. (1979). *Social Statistics*, Rev. ed., New York, NY: McGraw-Hill, Inc.
Campbell, D.T. and J.C. Stanley. (1963). *Experimental and Quasi-Experimental Designs for Research*, Boston, MA: Houghton Mifflin Co.
King, G. (1997). *A Solution to the Ecological Inference Problem*, Princeton, NJ: Princeton University Press.
King, G., R.O. Keohane, and S. Verba (1994). *Designing Social Inquiry*, Princeton, NJ: Princeton University Press.
Kleinbaum, D.G., L.L. Kupper, and K.E. Muller (1988). *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed., Belmont, CA: Duxbury Press.
Kuhn, T.S. (1970). *The Structure of Scientific Revolutions*, 2nd ed., Chicago, IL: University of Chicago Press.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

4

Univariate Measures for Directly Measurable Phenomena

Changhwan Mo
Rutgers University, Newark, New Jersey

This chapter begins, at first, introducing several tabular and graphical formats that can be used for organizing, summarizing, and presenting numerical data. After that, we briefly examine univariate measures, such as central tendency and dispersion which deal with one variable at a time. After we have collected data for analysis, we have several options for addressing them. We will investigate those options and see the aspects of them.

I. FREQUENCY DISTRIBUTION

Organizing and presenting a set of numeric information are among the first tasks in understanding a problem. As a typical situation, consider the values which represent the travel time to work of 57 employees in a large downtown office building. The times are given in minutes and each value represents an employee's average time over ten consecutive work days. The mere gathering of this data together is no small task, but it still needs further work for utilizing them as useful information. These raw numbers should be organized in a systematic way.

The easiest way to organize a set of data is to construct an array, which is a list of the numerical data ordered from low to high (or high to low). Arrays are often used to make the overall pattern of the data clear. However, the construction of array demands tedious works when the number of values is too large, and its output may turn out to be incomprehensible.

A more systematic way to summarize a large set of data is to construct a frequency distribution. A *frequency distribution* is a summarizing table form that shows the number of items that fall in each class of a data set. A class is an interval of values within the overall range of values in a data set. Generally, this frequency distribution makes us easily see the overall pattern of the data.

A frequency distribution is also known as a frequency table. To construct a frequency distribution, we must follow these four steps:

1. Select the number of classes.
2. Choose the class interval or width of the classes.
3. Count the number of data that falls into each of these classes.
4. Display the results in the form of a chart or table.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 1 Frequency Distribution of Commuting Time

Class (time in minutes)	Frequency (persons)	Relative frequency	Percentage frequency
20–29	10	0.175	17.5
30–39	12	0.211	21.1
40–49	17	0.298	29.8
50–59	15	0.263	26.3
60–69	3	0.053	5.3
Total	57	1.00	100.0

There are no best rules for constructing frequency distributions because no one can fit all situations. Table 1 shows an example of frequency distribution which summarizes the travel time to work of 57 employees in an office. Its class interval is all equal ten minutes and there are five classes.

The number of observations in any class is the class frequency. The total number in all classes is the sum of individual class frequencies. Sometimes, a relative frequency is useful to summarize a set of data. The relative class frequencies, or proportions, are found by dividing the class frequencies by the total number of data. A *percentage distribution* is calculated by multiplying the relative class frequencies by 100 to convert them to percentages. For example, when a class frequency is 17 and total number of frequencies is 57 as in Table 1, the relative frequency is 17/57, or 0.298, and the percentage frequency is (0.298)(100), or 29.8%.

Frequency distributions are useful tools for organizing and summarizing sets of data and for presenting characteristics of data clearly. Sometimes, however, we need information on the number of observations whose numerical value is ''less than'' or ''more than'' a given value. As you show at Table 2, this information is contained in the *cumulative frequency distribution*. We can convert a percentage frequency into a cumulative frequency distribution by adding the percentages from the top or the bottom of the frequency distribution.

Graphics are an effective tool to help people understand the characteristics of data, and they are essential for the presentation and analysis of data. The statistical graphic forms are as follows: line charts, bar charts, histograms, combination charts, and pie charts. *Line charts* use lines between data points to show the magnitudes of data for two variables or for one variable over time. *Bar charts* are often used to show the sizes of data for different qualitative or categorical data. *Histograms* are similar to bar charts, but they are mostly used for quantitative or numerical data and there is no empty space between bars. Usually the horizontal axis denotes class interval and the vertical axis shows class frequency according to each class interval. *Combination charts* use lines and bars, or use other charts together, to show the dimensions of two or more

Table 2 Cumulative Frequency Distribution

Time (minutes)	Percentage frequency	Cumulative frequency
Less than 30	17.5	17.5
Less than 40	21.1	38.6
Less than 50	29.8	68.4
Less than 60	26.3	94.7
Less than 70	5.3	100.0

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

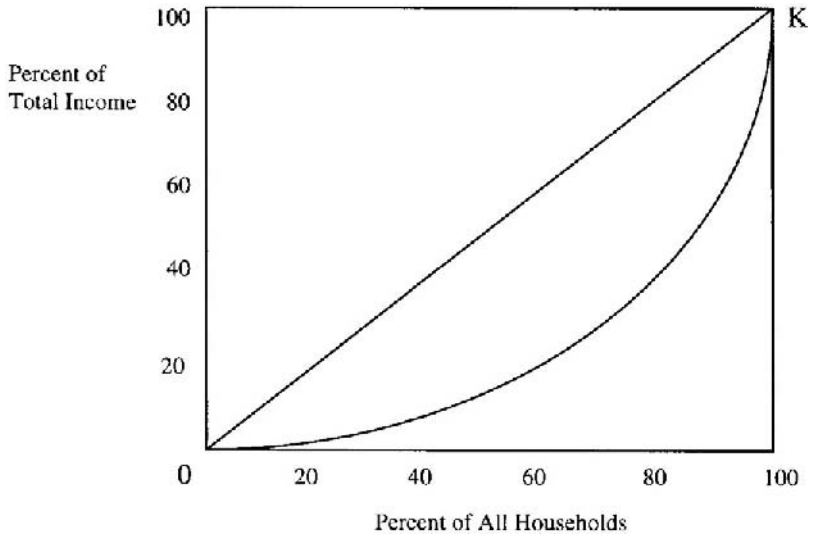


FIGURE 1 The Lorenz curve.

data values for different categories or for different times. *Pie charts* can be used effectively to show the relative proportions or percentages of the total number of measurements in qualitative data. It is recommended to be less than five.

In addition, we introduce a useful graphic form: the *Lorenz curve* (Figure 1). It is usually used for highlighting the extent of equality or inequality of income distribution in economics (Kohler, 1977). Consider the distribution of money income among U.S. households. We draw a square which is measuring percentage of total money income received on the vertical axis and the percentage of households on the horizontal axis. According to each income level from the lowest to the highest, households are arranged from left to right.

Consider a straight line from the bottom left corner at 0 to the top right corner at K. This diagonal line means perfect equality, since it represents the position the Lorenz curve would hold if the same portion of money income went equally to each household. If all households in the country shared total income equally, it would be true that 40 percent of the households shared 40 percent of total income, that 60 percent of the households shared 60 percent of total income, and so on. In fact, the differences of income between the poor and the rich seem to become larger and larger. Thus, the line of actual inequality exists lower than that of perfect equality. The difference between actual inequality and perfect equality determines the Lorenz curve, in other words, the curved line of inequality from 0 to K. Someone may argue that this curve is for bivariate relationship, but we introduce it here because it represents one concept: the inequality of income distribution.

We saw how tabular and graphical forms of presentation may be used to summarize and describe quantitative and qualitative data. These techniques help us to distinguish important features of the distribution of data, but most statistical methods require numerical expressions. We can get these numerical forms through arithmetic calculations on the data, which produce descriptive statistics. The descriptive statistics are measures of central tendency and measures of dispersion. The mode, median, mean, and weighted mean are presented as measures of central tendency. The range, mean deviation, variance, standard deviation, and coefficient of variation are explained as measures of dispersion.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

II. MEASURES OF LOCATION (CENTRAL TENDENCY)

In most sets of data, we usually see that there is a particular tendency for the observed values to cluster themselves around some specific values. Some central values seem to have the characteristic of the whole data, and central tendency refers to this phenomenon. We may use these values to represent the whole set of data because the central values usually position the middle point of distribution.

The *mode* is the most frequently occurring value in a data set. The mode is generally not a useful measure of location. For example, assume that we collect the temperature data (Fahrenheit) of six winter days in New York City: 49, 7, 11, 18, 22, and 49. Although one value (49) does occur more than once, there is no guarantee that this value shows the central tendency of the data set.

The *median* is a number that divides an ordered set of data in half. We can find this value when the values in a set of data have been arranged in a numerical order from the lowest to the highest. If there is an odd number of values in the data set, then the median (Md) is the value in the middle position. In the case of an even number of values in the data set, it is the average of the two values in the central positions. Consider the temperature data in New York City which have six values. When you wish to know the median of this data, it is calculated like this:

$$M_d = \frac{18 + 22}{2} = 20.$$

The most frequently used measure of central tendency is what laymen call an average. The word ''average'' in life has all kinds of different meanings such as a baseball player's batting average, a student's grade point average, and a man's appearance as average. Generally the term average in a set of quantitative data refers to their arithmetic *mean*. Simply, the mean of n numbers is their sum divided by n. Since it is desirable to have a formula which is always applicable, we state it with formal expression. For a given population of N values, X_1, X_2, X_3, . . . , X_N, the *population mean* is denoted by μ and the mean for a population of N data values is their sum divided by N.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

However, we often have to use sample values for estimating the mean of a larger population because of the time and cost involved in using the entire population. For instance, suppose we are interested in estimating the mean temperature of all winter days (population) in New York City by using the sample of six winter days that we already used for calculating the mode and the median. We perform the same calculation as the mean for a population data, but we divide the sum of sample values by the sample size n (as opposed to the population size N), and we call it *sample mean* which is denoted by $\overline{X}$.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Applying the equation of sample mean to the temperature data, we find: $\overline{X} = 1/6(49 + 7 + 11 + 18 + 22 + 49) = 26$. It means that the average winter temperature in NYC may be 26°, but we are not sure this sample mean can be regarded as a population mean because the sample number is only six.

When we compute the mean of a set of data, we assume that each value has equal importance. In the situation where the numbers are not equally important or not equally proportioned,

we can assign each a weight that is proportional to its relative importance and calculates the *weighted mean* ($\overline{X}_w$). Let $X_1, X_2, X_3, \ldots, X_N$ be a set of data values, and let $w_1, w_2, w_3, \ldots, w_N$ be the weights assigned to them. We can find the weighted mean by dividing the sum of the multiplication of the values and their weights by the sum of the weights:

$$\overline{X}_w = \frac{\sum X_i w_i}{\sum w_i}$$

For example, the average salaries of elementary school teachers in Oregon and Alaska were $23,000 and $20,000, and there were 1000 and 200 elementary school teachers in these states. When we want to find the average salary of elementary school teachers in these two states, we should calculate a weighted mean because there are not equally many school teachers in the two states. The solution is as follows:

$$\overline{X}_w = \frac{(23{,}000)(1000) + (20{,}000)(200)}{1000 + 200} = 22{,}500$$

Thus, the average salary of elementary school teachers in these two states is $22,500.

III. MEASURE OF DISPERSION

When we wish to know about the variation or scatter among the values, we calculate a measure of dispersion. Suppose that in a hospital each patient's pulse rate is taken four times a day and that on a certain day the records of two patients show the same mean of pulse rates. Whereas patient A's pulse rate is quite stable, however, that of patient B varies widely. Patient A's records show 71, 73, 73, and 75, while those of patient B are 48, 70, 81, and 93. When we calculate the means of both patients' rates, they are the same (73). Although they have the same mean of pulse rates, it does not necessarily mean that their conditions are identical. Thus, a doctor might pay more attention to patient B than patient A. This example illustrates the importance of measuring dispersion in descriptive statistics. In this section, we will deal with four measures of variation: range, mean deviation, variance, and standard deviation.

The *range* is the difference between the largest and smallest values in a data set.

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

When we apply the formula of range, we can see that the temperature data set of NYC has a range of $49 - 7 = 42$. It is not a satisfactory measure of variation for several reasons. First, its calculation uses only two of the observed values regardless of the size of sample. In this sense, the range is inefficient in that it ''wastes'' or ''ignores'' data. Second, the range is sensitive to sample size. As the size of sample is larger, the range generally tends to become larger. Third, the range may vary widely. It is the least stable of our measures of dispersion for all but the smallest sample sizes.

The *mean deviation* measures variation using distances between each data point and the population mean without considering the algebraic signs. When a data set is tightly clustered around the mean, the distances will be small. When the data set is spread out widely, the distances will be large. When we have a population of N number, $X_1, X_2, X_3, \ldots, X_N$, whose mean is μ, then we might be tempted to think that the average, or mean, of these distances should provide a good measure of dispersion (Watson et al., 1993). If we just add the distances without addressing the fact that about half of the distances will be positive and half will be negative, we will always get one answer: zero. By eliminating the signs of these distances, we

can solve this problem in two ways: first, we may simply ignore the signs by taking the absolute value, or we may square the distances. If we ignore the signs of the distances ($X_i - \mu$) and divide the sum of these absolute values by N, we have the mean deviation.

$$\text{Mean deviation} = \frac{1}{N}\sum_{i=1}^{N} |X_i - \mu|$$

In the above formula, $|X_i - \mu|$ is the absolute value of $X_i - \mu$, that is just $X_i - \mu$ with the sign converted to + (positive) if it happens to be − (negative). Since the mean deviation does not have the mathematical properties because of artificially ignoring the signs of the distances, we are looking for the better procedure to eliminate the sign on the deviation. It is to square the distances of each $X_i - \mu$.

Suppose that we use the square of the deviations instead of the absolute value as a measure of deviation. In the squaring process the negative signs will disappear; hence, the sum of the squares of the deviations from the mean will always be a positive number. Although this sum of squares provides a measure of dispersion, the mean of the squared deviations is more often used as a dispersion measure because of its conciseness. To calculate this measure, we divide the sum of square deviations by N, the size of population. This mean of squared deviations for population data is called the *population variance* (σ^2).

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2$$

The *population standard deviation* σ of the numbers in a population of size N is the square root of the variance. The standard deviation for a population of N data values, X_1, X_2, X_3, . . . , X_N, is the square root of the population variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2}$$

Earlier, we made a distinction between μ, the mean of population, and $\bar{X}$, the mean of sample. The different notations are used to distinguish whether they came from a population or a sample selected to represent a population.

The same type of symbol distinction is made between the population standard deviation σ and the sample standard deviation S. In addition, we must change the formula to divide by degrees of freedom (n − 1) for the sample data rather than the population size (N). When dealing with a sample of size n, we lose a degree of freedom for each parameter in the formula since we must estimate from sample data. If our data set is a sample and we wish to estimate a *sample variance* S^2, we can find it after the sample mean $\bar{X}$ is calculated at first. The variance for a sample of n data values is calculated by dividing the sum of the squared deviations for the values from their mean $\bar{X}$ by the degrees of freedom, n − 1.

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

Applying this formula to the temperature data set of NYC again, we can calculate the sample variance since we already know the sample mean ($\bar{X}$) is 26:

$$S^2 = \frac{1}{6-1}\{(49-26)^2 + (7-26)^2 + (11-26)^2 + (18-26)^2 + (22-26)^2 + (49-26)^2\} = 1724/5 = 344.8.$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

It means that an average squared distance of any observation in the data set from the mean is 344.8.

The standard deviation is always the square root of the variance. Thus, we define the *sample standard deviation* S is the square root of the sample variance.

$$S = \sqrt{\frac{1}{n-1}\sum (X_i - \overline{X})^2}$$

When we apply this equation to the temperature data, we find: $S = \sqrt{344.8} = 18.57$. It means that an average distance from the mean is 18.57.

There are only two differences between the formulas for the population and sample standard deviation. First, in the equation of population standard deviation, we use the mean μ, while in the equation of sample standard deviation we use $\overline{X}$. Second, in the population equation for σ we divide the sum of squared deviations by N, but in the sample equation we divide it by $n - 1$. Why do we have to use $n - 1$ instead of n? In addition to the formal rationale of adjusting for degrees of freedom lost by estimating μ with $\overline{X}$, we can intuitively say that the spread of values in a sample will typically be less than the spread in the population (Watson et al., 1993). In the case of estimating the population standard deviation by using a sample data set, it is desirable to adjust our calculations to complement the smaller spread in the sample. In other words, the sample standard deviation s^2 becomes a better estimator of the population variance σ^2 when we use $n - 1$ rather than n. There are n squared deviations from the mean in a sample of n data values, but only $n - 1$ of the deviations are free because of the limit that the sum of the deviations from the mean is zero as explained in the earlier discussion of the mean deviation (Watson et al., 1993). In general we use s as the estimator of σ because the standard deviation is the square root of the variance.

What does the standard deviation tell us? A data set with a large standard deviation has much dispersion with values widely scattered around its mean and a data set with a small standard deviation has little dispersion with the values tightly clustered around its mean. If the histogram for a set of data values is shaped like a bell or shows normal distribution, we can say that:

1. About 68 percent of the values in the population will lie within $\pm$ 1 standard deviation from the mean.
2. About 95 percent of the values will fall within $\pm$ 2 standard deviations from the mean, which means that about 95 percent of values will be in an interval ranging from 2 standard deviations below the mean to 2 standard deviation above the mean.
3. About 99 percent of the values will lie within $\pm$ 3 standard deviations from the mean.

For instance, when a stock traded on the New York Stock Exchange has a mean price of \$50 and a standard deviation of \$3 for one year, we are sure that 95% of the prices lies between \$44 and \$56 because this interval is $\mu \pm 2\sigma$. This conclusion is based on an assumption that the distribution of prices is approximately symmetrical.

IV. MEASURES OF CENTRAL TENDENCY AND DISPERSION: APPLICATION

We have presented both measures of central tendency and dispersion. In this section, we will briefly investigate which measure is most appropriate for a specific situation and see a descriptive statistic that combines both measures: the coefficient of variation.

How can we select measures of central tendency and dispersion? If the distribution is equally symmetric, then the $\overline{X}$, M_o, and M_d will all coincide. When the distribution is not symmetric or is skewed, the mean, median, and mode will not match together. It is not unusual that some frequency distributions are skewed to the left or to the right. The mean is sensitive to outliers, a few extreme values, but outliers typically have little or no effect on the median. To the temperature data set of NYC, if we add one extreme value (89), the new data set is: 49, 7, 11, 18, 22, 49, and 89. By using it, when we calculate the median and the mean, they are, respectively, 22 and 35. As a result of adding an outlier, the mean has affected a lot from 26 to 35, while the median does not change so much from 20 to 22. Therefore, when the data are skewed or contain extreme values, we can say that the median provides a better measure of central tendency. In the case of dispersion measures, the range is particularly sensitive to outliers. We generally use the variance and standard deviation for representing the dispersion in a set of data values.

A descriptive statistic that combines the standard deviation and the mean is called the *coefficient of variation*. The coefficient of variation (CV) is useful for comparing two number sets of rather different magnitudes. Its formulas are as follows:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100 \quad \mu > 0 \text{ [for a population]}$$

$$CV = \left(\frac{s}{\overline{X}}\right) \times 100 \quad s > 0 \text{ [for a sample]}$$

While the standard deviation depends on the original units of measurement, CV is a unitless figure that expresses the standard deviation as a percentage of the mean (Freund and Simon, 1995). For instance, the lengths of certain distances may have a standard deviation of 1000 meters or 1 kilometer, which is the same, but neither value really tells us whether it reflects a great deal of variation or very little variation. Let's see another example for further understanding. At a hospital, patient A's blood pressure, measured daily over several weeks, averaged 199 with a standard deviation of 12.7, while that of patient B averaged 110 with a standard deviation of 9.5. When we want to find which patient's blood pressure is more consistent, we calculate the coefficient of variation because their means are different.

$$CV_A = \frac{12.7}{199} \times 100 = 6.38 \quad CV_B = \frac{9.5}{110} \times 100 = 8.6$$

At first glance, it appears that patient B's blood pressure is relatively consistent because its standard deviation is smaller than that of patient A. When we compare CV_A and CV_B, however, we can conclude that patient A's blood pressure is relatively more consistent than that of patient B since CV_A is smaller than CV_B.

V. CONCLUSION

The univariate measures refer to measures of central tendency and dispersion. When we summarize data by using univariate analysis, it should be noted that it has a disadvantage of losing critical information. To minimize the loss of information, analysts often use different univariate measures together. As is seen in Table 3, we can generally say that these descriptive statistics are the best fit for different levels of data. However, it does not necessarily mean that only these different levels of data are applicable to specified descriptive statistics.

TABLE 3 Measures of Descriptive Statistics and Level of Data

Measures of central tendency	Measures of dispersion	Minimum level of data required
μ, $\bar{X}$	σ^2, σ, S^2, S	Interval/ratio
M_d	Range	Ordinal
Mode	Frequency distribution	Nominal

In addition, univariate analysis is important because multivariate analysis, which examines several variables such as factor analysis and multiple regression, starts from the basic logic of these descriptive statistics.

REFERENCES

Babbie, E. (1983). *The Practice of Social Research*, 3rd ed., Belmont, CA: Wadsworth Publishing Co.

Freund, J.E. and G.A. Simon (1995). *Statistics: A Firsts Course*, 6th ed., Englewood Cliffs, NJ: Prentice-Hall.

Johnson, J.B. and R.A. Joslyn (1986). *Political Science Research Method*, Washington, DC: Congressional Quarterly Inc.

Kohler, H. (1977). *Scarcity and Freedom: An Introduction to Economics*, Lexington, MA: Heath and Company.

Levin, R.I. and D.S. Rubin (1994). *Statistics of Management*, 6th ed., Upper Saddle River, NJ: Prentice-Hall.

Watson, C.J., P. Billingsley, D.J. Croft, and D.V. Huntsberger (1993). *Statistics for Management and Economics*, 5th ed., Boston: Allyn and Bacon.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

5

Typologies, Indexing, Content Analysis, Meta-Analysis, and Scaling as Measurement Techniques

William M. Bowen and Chieh-Chen Bowen
Cleveland State University, Cleveland, Ohio

I. INTRODUCTION

The techniques discussed in this chapter may all be related to measurement and data analysis. Their tremendous potential usefulness in public administration is due in large part to the strength of the method of which they are all broadly a part. They have roots in scientific method and more specifically in the logic of measurement. To one degree or another they all enable an investigator to extend the logic of measurement beyond the physical realm of concrete and material objects and into the realm of abstract and intangible entities.

The term ''entities'' is used advisedly in this chapter to refer to the units of empirical investigation. Of all the possible terms for these units, the term ''entity'' seems to meet two criteria the best. First, it is consistent with the traditional logical empiricist conception that the units of investigation must be rooted in physical or biological realities or, alternatively, social or psychological realities from a behavioral perspective. Second, it also seems consistent with the possibility of empirical investigation of irreducibly subjective magnitudes and other nontraditional analytic units (Bowen, Chang, and Huang, 1996).

The various conceptual and theoretical aspects in the extensive body of literature dealing with the twin topics of conceptualization and measurement are too multifaceted, technical and contentious to be summarized simply and concisely (Blalock, 1982; Kyburg, 1984; Nagel, 1931; Roberts, 1979). Moreover, the conventional treatments and courses on data analysis in public administration neglect to raise some of the philosophical and theoretical underpinnings of these topics sufficiently for beginners at empirical investigation to fully grasp the elementary ideas involved, much less the relationships between them. As a consequence there is a widespread lack of appreciation among both scholars and practitioners in public administration with respect to the enormous potential for innovation and practical application of measurement and empirical data analysis. Even today many hold the belief that scientific method is only useful in public administration insofar as administrative systems are the concrete and material or directly observable, behaving entities of the sort postulated by logical empiricists. This chapter tends to belie this belief by introducing a conceptual framework. It offers some of the techniques that may be used to extend the boundaries of scientific method to far more abstract and intangible realms

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

such as decision premises, preference and utility judgments, cognitive processes, and linguistic constructs.

To help ensure adequate background knowledge, we begin by considering a few basic ideas about empiricism and measurement. Because the foundations of scaling technique come appreciably close to the status of fundamental theory, we emphasize some of the ideas that are prerequisite to a proper understanding of such foundations.[1] While we include a very few carefully selected philosophical and theoretical ideas, primarily at the beginning, as a rule our coverage of the techniques reflects the assumptions of the practicing investigator, not the finely-detailed formulations of the professional philosopher.

II. SCIENCE AND THE LOGIC OF MEASUREMENT IN PUBLIC ADMINISTRATION

One may distinguish between categories of knowledge about physical, biological, and social realities. Knowledge is not an innate possession, so regardless of the category in question a method is always required to obtain it. The objective of such a method is always to find out, from consideration of what we already know, something else which we do not know. That is, the objective of such a method is to make inferences. Moreover, we may say that method is a ''good'' one if it leads to true inferences from true premises and not otherwise. If the method is good then the question of the validity of the inference is purely one of fact and not of thinking. This is important in public administration largely because inferences about administrative systems often serve as premises for decisions (Simon, 1976).

Of the available methods, scientific method is preeminent insofar as it is the only one which demands evaluation of one's inferences against future experience (Peirce, 1877). Measurement in public administration may be considered to be a technical aspect of scientific method in which ultimately abstract, mathematical and quantitative symbols are translated into the existential qualities or empirical traits of selected and well-defined aspects of administrative systems. Measurement fortifies scientific method, which in turn provides a sort of control on the quality of one's inferences about the world. The analytic techniques in this chapter are thus useful in public administration primarily when motivated by the desire to hold up one's inferences about certain aspects of an administrative system to the standards of social science. Their value increases directly with the importance of the validity of the administrator's inferences, as for example when they serve as the premises for important decisions.

To philosophers the term ''scientific method'' tends to include various speculative activities such as generalizing from observed facts to scientific ''laws,'' or developing logical systems called ''theories.'' The elements of scientific method that stem from measurement, however, focus more narrowly on the processes of empirical observation and description. While our concern here is primarily with the techniques stated in the chapter title, an adequate appreciation of them requires recognition of their conceptual and theoretical roots in mathematics and quantitative reasoning.

In the physical and biological sciences, at least, the ultimate expression of knowledge tends to take a mathematical and quantitative form. In physics, for example, mathematical formulae are used to express the four dimensional reality of the special theory of relativity. In physical chemistry, electrons are specified statistically. In population biology and genetics, the structure of populations are described quantitatively. In all of these examples, and countless more, mathematics and quantitative reasoning, augmented by measurement, appear to offer the final and plenary test of the quality of human thought. It is as if such reasoning provides a proving ground for our thinking; a testing of the truth of our reason.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

In this regard the commonality between the physical and biological sciences on one hand, and the social and administrative sciences on the other extends much further than is often realized by those uninitiated in empirical investigation. Firstly, we have noted that all science, including social science, demands that one evaluate its conclusions against future experience. Normally we observe this by noting that social science demands that we test the validity of our empirical inferences. Mathematics and quantitative reasoning are instrumental in these tests. Measurement extends mathematics and quantitative reasoning to the empirical world and, in doing so, assists the public administrator on the proving ground of his or her thinking. Secondly, most if not all of the basic concepts of scientific method are shared in common between the various realms of science. Scientific method, whether physical, biological, or social, deals with the basic concepts of variables, parameters, assumptions, causes and effects, theories, laws, and research designs, among many others. Beyond this point, if there are any meaningful differences between the social sciences on the one hand and the physical and biological sciences on the other, they are that the former poses greater complexities, difficulties, and challenges (Machlup, 1994). Indeed it may be argued that there is no good reason to conclude that the logic or rationale of social scientific method is essentially any different than in physical or biological science (Rescher, 1970; Salmon, 1984; Simon and Burstein, 1985). Of course, by extension, there is no good reason to conclude that scientific method is any different in public administration.

A. Definition and Measurement

In scientific method, concepts are defined in the process known as concept formation. This process is, by-in-large, a matter of semantic maneuvering to obtain the maximum congruence of categories. Its product, a concept, represents a relevant set of empirical entities by stipulating the relationships between their attributes, characteristics or qualities. Measurements are statements of the interstices of this representation.

The definition of measurement most commonly used in pubic administration may be stated as ''the unique assignment of a range of numerals to a domain of magnitudes according to rules'' (Stevens, 1957). Partially, this definition is used because it is conducive to the study of decisions and other abstract social entities. It implies that measurement is essentially a systematic activity, not necessarily limited in application to concrete and material realities. While numbers themselves may be created by humans, we assume that the magnitudes reflect a definite quality of the entities, events, or objects under investigation.

Measurements begin with operational definitions. Operational definitions are instructions or descriptions of sets of actions or operations an investigator can follow exactly, designed to link the concepts to magnitudes in the world. They enable replicability, a basic requirement of scientific method. They refer to attributes or characteristics of the empirical entities that the investigator is representing with the concept. If these attributes or characteristics have two or more levels then they are known as ''empirical variables.''

One of the basic principles of all empirical research is that before any measurements may be taken one must provide suitable operational definitions of all the major concepts in one's investigation. The reason for this is that all measurements logically presuppose definition. That is, operational definition is required because the numerals have no meaning in and of themselves. Rather their meaning originates, as does all meaning, in the abstract replacement of one symbol (or set of symbols) by another. Specifically in creating an operational definition, the word used to label the concept to be defined, a symbol, is replaced by another set of words, themselves also symbols, this time representing the operational definition. Thus the numerals used in measurement acquire their meaning, at least in part, from the creation of an operational definition. Without operational definition the numerals have no meaning in relation to the concept.

Measurements not only logically presuppose operational definitions. They also presuppose the rules of measurement. That is, the numerals get meaning not only from the creation of operational definitions, but also from the abstract replacement of the entity to be measured, a symbol, by a numeral, itself also a symbol. Let us call this replacement the ''assignment of a numeral.'' The rules of measurement are required to govern and constrain the assignment of a numeral so as to ensure that certain psychological and logico-mathematical antecedents are fulfilled in the process. The relevant psychological and logico-mathematical details of these rules are discussed in a following section.

If the assignment of a numeral conforms with both the operational definition and rules of measurement then the numerals may be logically linked to the concept. Moreover, because numerals are subject to the laws of mathematics, investigators may, with reference to the concept, organize their thoughts and observations with a degree of logical precision and clarity that would otherwise not be possible.

Because assignment of numerals in the measurement process links the numerals with a concept, and because numerals are logically linked to mathematics and quantitative reasoning, measurement links concepts with mathematics and quantitative reasoning. In doing so it enhances the investigator's ability to think logically about the relationships between the attributes or characteristics of the empirical entities represented by the concept. In a word, measurement, as opposed to definition in the absence of numerical assignments, improves the investigator's ability to reason through the relationships with reference to which the world is represented by the concept.

The primary advantage of the measurement process as conceived in our definition may thus be construed to derive not from its ability to somehow put the investigator in direct contact with the actual world, or to link his or her concepts directly to the actual world. Neither the concepts nor the numerals need in any sense be reified for the measurement process to be advantageous. Rather, its primary advantage stems from the systematic linkages it enables between the investigator's concepts and the laws of mathematics and quantitative reasoning. In other words it derives in the first instance not from the information in the numerals as they relate to the world as it actually exists—the facts at issue in the items of information at our disposal—but rather more directly and simply from the enhanced ability one obtains with respect to *how one proceeds in organizing one's knowledge about it* (Rescher, 1979).

B. The Concept of Unique Assignments

The definition stipulates that the assignment of the range of numerals to the domain of magnitudes is ''unique.'' The range of numerals used in measurement symbolically represent or correspond to a domain of empirical entities, and their meaning is derived from representation or correspondence. Figure 1 illustrates this idea using the range of integers from one to three in the left column to symbolize certain qualities of the domain of empirical entities, represented by the asterisks on the right. The asterisks could represent employees or utilities or any empirical entities that may be of interest. Of course, in the world of measurement practice the range usually contains more than three numerals, along with a greater variety of entities in the domain. The principle, however, remains the same. In case A, each numeral in the range maps uniquely, vis-a-vis the mapping function, to the empirical entities in the domain. The uniqueness of these assignments assures the meaningfulness of the numerals. In case B, however, the mapping is not unique. The mappings in B are not necessarily incorrect because there are no things that are necessarily in and of themselves the qualities which we attribute to numbers or equations. Indeed there is nothing logically or inherently wrong with the assignments in case B; the relationships between range and domain are merely ambiguous. They simply lack meaning. This quality

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

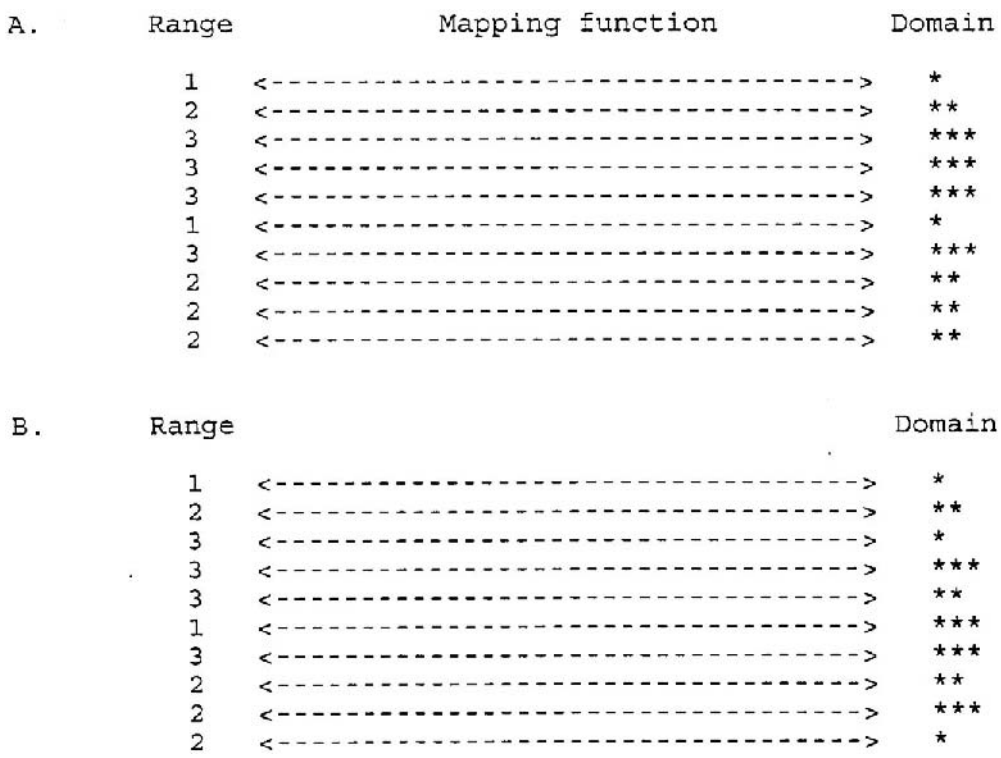


FIGURE 1 The numerals in the range, in the left column, are assigned to the magnitudes in the domain, in the right column. A unique assignment is one in which one and only one numeral is used to represent one and only one magnitude.

of ''uniqueness'' is required to give meaning to the numerals used in measurement. When the assignments are not unique, the information contained in the numerals either lacks integrity or else the amount of information contained in them is less and the ambiguity is greater than it would be were they unique.

C. Mathematical and Psychological Assignment Rules

This definition of measurement also stipulates that numerals are assigned to magnitudes *according to rules*. These rules stimulate minimum conditions for measurement reliability. They have both psychological and mathematical aspects. Unless one conforms meticulously to both of these aspects, one's measurements will not be reliable. Measurement reliability and validity are especially important when dealing with the techniques in this chapter, as will be clarified in the following sections.

Several alternative sets of rules are available, each with different logico-mathematical properties. The decision on which set rules to use constitutes the selection of a level of measurement. Four levels of measurement are traditionally distinguished by the logico-mathematical properties of the numerals used, and at least five may be identified (Stevens, 1957).

The quality of any analysis based upon measurement depends upon the perceptions of the investigator and the laws of mathematics. More specifically, measurement reliability depends upon the ability of the investigator to accurately perform the required perceptual processes. The

different levels of measurement correspond to different perceptual processes. Once the numerals have been assigned to the magnitudes, the quality of any further analysis depends upon whether or not the investigator complies with the pertinent laws of mathematics in his or her treatment of the numerals obtained. We now briefly review these rules in both their psychological and mathematical aspects.

Nominal measurements only have the basic geometric property of dimensionality (or spatial extension). They only contain information about membership in a well-defined subclass of observations. This property requires only that the investigator be able to perceive the divisions that define the subclasses in his or her observations, and locate each entity being measured uniquely in one of the subclassifications using only one numeral. That is, to perform nominal level measurement the investigator need simply be able to discern whether or not an observation is ''equal to'' the standard or criterion used to define the relevant subclass.

Once the investigator has nominally measured all of the empirical entities of interest, he or she may want to transform the numerals mathematically. This might be done, for example to summarize a large number of measurements or to manipulate them into a form that is suitable for making statistical inferences. The information contained in numerals with nominal properties will retain its integrity so long as all the values are transformed similarly, by any one-to-one substitution. So long as one uses a single formulae to transform all of the observed values, and so long as it is applied in a consistent fashion across all observations in all subclasses, one may, without loss of information, add an arbitrarily large constant to each value, multiply it by either a positive or negative number, or exponentiate it, all without loss of information.

Ordinal measurements contain information about rank order. They represent a logical extension of the nominal level, obtained by adding the property of rank order to that of dimensionality. Ordinal measurements thus presuppose not only that the investigator has the ability to determine the equality of the observations, as in nominal measurement, but also the ability to determine, with respect to the location of any two of them on the attribute of interest, whether either one is greater than or less than the other. There is no requirement with ordinal measurements to be able to determine *how much* one such entity is greater or less than another, but only their rank order. It is enough that one is able to psychologically compare two of the empirical entities and identify, with respect to whatever empirical dimension one is measuring, whether or not one of the empirical entities represents more or less of that dimension. Such comparison is often used when one is doing scaling.

Once the investigator has ordinally measured all of the empirical entities of interest in the study, the values or numerals may be mathematically transformed so long as the rank order information they contain is preserved. In comparison to nominal measurements, the set of permissible transformations of the values of ordinal measurements is more restricted. Technically, one may transform ordinal values by any increasing monotonic function without loss of information. However transformation by a decreasing function or a nonmonotonic function may lead to loss of information. This means, for example, one may add an arbitrarily large positive constant to each of the values of an ordinal variable, multiply each of the values by an arbitrarily large constant, or take their logarithms without loss of information.

Interval level measurements contain all of the information contained in the nominal and ordinal levels, plus information about equal intervals. They logically extend the properties of ordinal measurements through the addition of the property of distance to the properties of dimensionality and rank order. When taking interval measurements the investigator must be able to accurately perceive and attribute the additional mathematical property of equal distances between successive numerals to the numerals one uses to represent one's observations. Such attribution depends for its accuracy upon the psychological ability of the investigator to identify the equality of intervals or distances between observations. In other words, to get interval level

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

measurements, the investigator must be able to determine not only whether two observations are equal and whether one observation is quantitatively greater or less than another, as in ordinal measurement, but also whether or not the distance between any two sequential observations is equal to the distance between any two such others on the dimension. If the investigator can perceive the exact distance between successive observations, and if this distance is constant, then interval level measurements may be achieved. Interval level properties are often attributed to the numerals obtained in Likert Scaling, which will be discussed in a following section.

As is the case with nominal and ordinal measurements, once the investigator has obtained interval measurements he or she may want to transform them for some reason. Permissible transformations are those which preserve not only rank order information but also equality of intervals. These include any transformations in which each value is multiplied by an arbitrarily large positive constant, and an arbitrarily large constant is added to the product. Technically, interval level measurements are said to be unique up to a positive linear transformation.

Ratio level measurements contain additional information about a natural zero. At this level of measurement, in mathematical terms, the ratio of two numerals is assumed to be independent of the unit of measurement. Practically speaking, this means that the investigator who attributes ratio level properties to his or her measurements implicitly assumes the ability to perceive whether two ratios of measurements are equal. This requires a ''natural zero.''

Transformations of ratio-level values are feasible, however the permissible set of such transformations is the most restricted of the four levels of measurement. Technically, ratio level measurements are said to be unique up to a similarity transformation. This means that one may multiply ratio level measurements by an arbitrarily large positive constant without loss of information. However, the addition of a constant to the values or the multiplication by a constant equal to or less than zero will compromise the information they contain.

The important point here is that the mathematical properties of these different levels of measurement govern and constrain both the psychological abilities that the investigator must assume and the transformations he or she may perform on the numerals obtained if meaningful measurement is to occur. So long as these constraints are met in the measurement process, the only restrictions on the empirical entities to which it may be reasonably applied are set by limits on the availability of suitable concepts. Conversely, so long as the investigator has the requisite psychological abilities, and the analysis of the numerals proceeds in accordance with the laws of mathematics, the concepts to which measurement may be usefully applied extend as far as the human imagination can take them. In comparison to investigations not based upon measurement, those so based have the distinct advantage that the investigator's inferences are stated in terms of a systematic mathematical foundation, presumably within the larger setting of a rationale-providing framework of conceptual order in administrative systems.

D. Measurement Validity, Reliability, and Error

Measurements have been defined as the assignments of numerals to magnitudes representing empirical entities. The principles of measurement require the selection of a level of measurement, the details of which are, as noted above, reasonably clear and unequivocal. In practice, however, investigators are far from infallible. The actual use of these principles in the conduct of an investigation is often fraught with obstacles that compromise the quality of the answer the investigation provides to the question the investigator asks. The related concepts of *measurement validity* and *measurement reliability* are concerned with whether or not certain aspects of the measurement process compromise this quality.

The concept of measurement validity has to do with whether or not measurements accurately reflect what the investigator intends them to measure. This has a couple of aspects. First, a

valid measurement must represent what it is intended to measure. This requires a well-conceived operational definition of the concept. Take for example the concept of job performance. A measure of the job performance of say, police officers, may be valid only if it really does measure the efficiency and effectiveness with which the officers do their job. One of the authors is familiar with an instance in which an urban police department measured the job performance of its patrol officers by the numbers of miles on the odometer of their police cars. The effect was that rather than going in to the dark corners of the city to control crime, as they ought to have been doing, the officers took to the highway to chalk up miles on their cars as a means of getting good performance ratings. The resulting performance measurements lacked validity.

Secondly, assuming that the measurement represents the concept it is intended to measure, measurement validity requires that the correct numerals are assigned to the empirical entities. Differences between the correct numeral and the numeral assigned to an empirical entity are known as *measurement error*. There are numerous common sources of measurement error. Among them, different people may perceive the empirical entity differently, or interpret the measuring instrument differently. Contextual differences due to factors such as the age, race or gender of the investigator may bias the measurements. So may the investigator's prior disposition due to level of native intelligence, education, or moral development. Temporary conditions such as disease, emotional distress, poor lighting, or high levels of noise may alter the investigator's perception, thus leading to measurement error. Some of these factors arise idiosyncratically and others systematically. All compromise the validity of the measurements.

Steps to mitigate against measurement error may include: (1) a single person taking repeated measurements of the same empirical entity; and taking the average of the observed values; (2) when the measurements require instrumentation, using mechanical devices to fix the reference point of observation; (3) making electronic observations which print automatically whenever possible; and (4) more than one person taking the average of repeated measurements of the same empirical entity. Some errors in measurements are also caused by either carelessness or systematic biases on behalf of whoever is taking the measurements, such as the halo effect, leniency, severity, and similar-to-me effect (Borman, 1991), and some of these may be reduced through training. There are many possible sources of measurement error, and while these and other steps may go a long way toward overcoming the obstacles posed by measurement error, some such error is inevitable.

The concept of measurement reliability applies to both operational definitions and to measurement errors. Measurements are deemed reliable if they are consistent or repeatable. If in repeated applications an operational definition produces a similar result every time, regardless of whether it represents the intended concept, it is reliable but not necessarily valid. Similarly, when measurement error is systematic, the measurements may still be reliable but not valid. All valid measurements are reliable in the sense that they have operational definitions that adequately represent the intended concept and they contain little to no measurement error. But not all reliable measurements are valid.

Take as an example of a reliable but not valid measurement, a hypothetical situation in which a male supervisor is known to secretly prefer male to female employees. Over time, his judgments of the performance of his subordinates may be consistent with their efficiency and effectiveness within the genders but not between them. In other words, he may from year to year systematically bias his evaluations in favor of male employees. The rank order of his judgments of the performance of the females is consistent from year to year, and similarly for the males. But his evaluations of the females are systematically biased downward in consequence of his hidden preferences. In this case, the performance appraisals may be deemed reliable because they are consistent. But they are not valid. Systematically incorrect numerals are assigned to the employees. The appraisals may be reliable, but they are not valid.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

If for whatever reason the operational definitions do not consistently represent the intended concepts, or if idiosyncratic measurement errors occur in an investigation, then the measurements are not reliable. When measurements are unreliable, the investigation is usually considered to be too seriously flawed to be credible to the critical and informed mind. The degree to which unreliable measurements undermine the value of investigations based upon measurement in public administration writ large is open to debate.

E. Validity of Empirical Inferences

A proper understanding of the techniques in this chapter requires extension of the ideas of validity and reliability to include not only measurements, but also empirical inferences. After all, the test of empirical knowledge is not the validity of the measurements but rather the degree to which the inferences they support are consistent with future experience. The decisive point is the validity of the empirical inferences, not the validity of the measurements upon which they are based. Valid measurements are an integral part of valid empirical inferences, since they represent the investigator's perceptions of the relevant segment of the world. In general, however, empirical inferences may not be immediately deduced from measurements or data. That is, a logical or conceptual process is also required. The techniques discussed in this chapter are examples of such processes. Accordingly, the idea of validity is now refined and extended from the realm of measurements to that of the inferences themselves.

One may distinguish between two different broad types of validity: external validity and internal validity. Both deal with the logic through which the measurements are linked to the empirical inferences. External validity refers to the generalizability of the empirical inferences. The question is: to what range of different populations or situations do the inferences pertain? There is no systematic technical device for assessing external validity; it is primarily a matter of judgment. Internal validity, on the other hand, refers to the correspondence between two sets of things, such as concepts, variables, methods and data. A degree of internal validity is the minimum without which a scientific study is uninterpretable. Ideally, empirical inferences have a high degree of both external and internal validity. In reality, however, the two are often at odds with one another (Campbell and Stanley, 1963).

Moreover, there are three forms of internal validity: face validity, criterion-related validity, and construct validity. These three forms are now considered in sequence.

1. Face Validity

The weakest form of validity is face validity, also sometimes called content validity. Face validity is based entirely upon logic, common sense, and subjective judgment. For instance, one may decide that a performance appraisal instrument should gather three types of data: objective, personnel, and judgmental. The objective data, for example, might be the number of letters typed by a secretary, the number of citations issued by a police officer, or the number of claims processed by a claims processor. Personnel data might include records of absenteeism, letters of commendation or discipline, and other information found in the employee's personnel file. Judgmental data could supplement the objective and personnel data by, for example, explicit numerical judgments of how well the employee performs. The test of the face validity of the combined measures would require only that the inferences made on the basis of such measures are plausible and consistent with other information about the performance of the employees being evaluated. Such plausibility and consistency is often evaluated by expert judges.

Face validity may be adequate if the purpose of the investigation is purely descriptive. However, even when expert judges are used to evaluate face validity, it is still the weakest form

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

of validity. Face validity is ultimately based on a single variable, which is to say, informed judgment, and stronger forms of validity all involve more than one variable.

2. *Criterion Validity*

Criterion-related validity refers to the correspondence among predictor variables and some criterion measure. It is established by analytically evaluating the strength of the relationship between the predictor variable and another variable—the criterion—to which it is expected to be related if it is valid. The criterion is a score, rating, or some other value of a variable that either is available at the time of the measurements of the predictor variable, or will be available at a later time.

In the preceding example of performance evaluation, to establish criterion validity would require specification of a criterion against which to compare the results of the application of the performance evaluation instrument. One would use the instrument to evaluate a set of employees and then compare their performance ratings with the criterion. For example, the instrument might not include information about achievement awards from professional organizations or grades received in evening classes voluntarily attended at the local university. One might expect that employees who receive achievement awards from professional organizations, or who receive high grades in their classes, might receive higher ratings. If one were to observe a significant positive relationship between an employee's performance rating, according to the instrument, and the receipt of achievement awards or high grades, then one might be said to have obtained a degree of criterion validity. For many instruments, especially those measuring complex concepts such as job performance, it is difficult if not impossible to decide what criterion to use to validate the instrument.

Two types of criterion validity may be distinguished. When the criterion measure is available at the same time as scores on the predictor, then concurrent validity is being assessed. When the criterion measure will not be available until some time after the predictor scores are obtained, then predictive validity is being assessed. The difference between concurrent and predictive validity is a function of the time when the criterion measure becomes available. Concurrent validity is oriented toward the present and reflects only the status quo at a particular time. Predictive validity is oriented toward the future and involves a time interval during which events take place.

Predictive validity is usually considered to be the more powerful of the two because the inferences from predictor variables are successfully generalized beyond the current study to situations not under the direct control of the investigator. For example, Namenwirth's (1973) analysis of party platforms in presidential campaigns, written in the late 1960s, suggested that America would experience severe economic difficulties that would peak about 1980. Events since seem to confirm this prediction (Namenwirth, 1973).

3. *Construct Validity*

Construct validity is the strongest and most complicated form of validity. It is concerned not only with validating the particular measurements and analysis used in any given application, but also the theory underlying them. To establish construct validity one must have repeated studies, with different measures and analysis; it may not be achieved by only one measure.

Though efforts to establish construct validity are rare in public administration, if one seeks to thoroughly understand the elements of empirical analysis it is important to understand the concept. The concept of construct validity provides an ideal form to which efforts to establish validity in empirical investigation may aspire. This ideal pertains not only to content analysis, meta-analysis, and scaling, but many other forms of empirical analysis as well.

TABLE 1 Example of a Multi-Trait-Multi-Technique Matrix with two Techniques (1 and 2) for Measuring Two Traits (A and B)

		Technique 1		Technique 2	
	Traits	A	B	A	B
Technique 1	A	A_1A_1	A_1B_1	A_1A_2	A_1B_2
	B	A_1B_1	B_1B_1	A_2B_1	B_1B_2
Technique 2	A	A_1A_2	A_2B_1	A_2A_2	A_2B_2
	B	A_1B_2	B_1B_2	A_2B_2	B_2B_2

Construct validity is established if, in measuring abstract concepts, the results are related to other analyses in the ways one would expect them to be on the basis of theory. Thus, one starts with a set of concepts and some hypothesis about how they are related. Take for example Herbert Kaufman's provocative thesis that the reason organizations ''die'' is that ''their engines stop'' (Kaufman, 1991). In light of evolutionary theory it may be deduced that if this thesis is correct then one will find an increase in the ''thickness'' of the organizational medium. The concept of ''organizational thickness'' is defined in terms of a set of measurable traits of organizations including degree of specialization, literacy and educational levels, volume and speed of communication, energy consumption per capita, and organizational density. Kaufman's thesis contains expectations about the relationships between these traits. The traits, in turn, may be measured using any number of different techniques. For a couple of the traits, physical measurements may be obtained. For others, indexes may be constructed. For still others, content analysis or scaling may be useful.

Systematic procedures are required to establish construct validity. These procedures culminate in a ''multi-trait-multi-technique matrix,'' such as is illustrated in general form in Table 1 (Campbell and Fiske, 1959). Each value in the matrix is simply a correlation coefficient. Conceptually, these measure four different characteristics. These characteristics are:

1. the same trait measured by the same technique, such as A_1A_1, B_1B_1, A_2A_2 and B_2B_2 (usually referred to as reliability of the measurement),
2. the same trait measured by different techniques, such as A_1A_2 and B_1B_2,
3. different traits measured by the same technique, such as A_1B_1 and A_2B_2, and
4. different traits measured by different techniques, such as A_1B_2 and A_2B_1.

Satisfactory construct validity is said to occur only under a certain condition regarding the relationships between these characteristics. Specifically when the correlation coefficient is statistically significant for the same trait measured by different techniques (characteristic 2), one has ''convergent validity.'' When different traits measured by the same technique (characteristic 3) are uncorrelated, one has ''discriminant validity.'' Only when convergent validity is significantly higher than discriminant validity can we say with confidence that the study has satisfactory construct validity.

III. THE TECHNIQUES

Having clarified some of the relevant elements of empirical investigation in general and measurement in particular, our attention now turns explicitly to typologies, indexing, content analysis,

meta-analysis, and scaling. As was mentioned earlier, these techniques all enable the investigator to extend the application of scientific methods in the realm of essentially abstract and intangible empirical entities such as decision premises, preferences and utility judgements, cognitive processes, and linguistic constructs.

Typologies are a form of classification. The other four are primarily techniques of data analysis. All are related in one way or another to measurement. While the following descriptions provide some basic insights into the respective techniques, if one wants to produce a study using one or more of them one will first want to prepare further by going beyond this chapter. References for such preparation are provided where required.

A. Typologies

A typology is a special form of classification (Mukherjee, 1983). Before focusing directly upon typologies, it is a good idea to consider classification more generally.

In simple terms, the reason for classification in public administration is that administrative systems may contain immense variation. Organizational structures may vary, for example, as may their goals, functions, communication patterns, and the roles people assume within them, among an unfathomable number of other attributes. Faced with all of this variation, an investigator attempting to consciously predict and control some aspect of an administrative system must first abstract from it, replicating the relevant attributes of it in his or her own mind. Ideally, postulational-deductive theory would serve to guide and organize the process of making these abstractions. Such theory would enable him or her to deduce the full set of variables needed to organize the variation in the system. One would first identify the parameters of the class of administrative systems under investigation, then define the relationships between them as precisely as possible, and finally construct models to relentlessly extend them and to test the postulates. Theory of this kind is either quantitative or at least cleanly qualitative in the sense that it leads to easily recognized inequalities. In public administration, however, most of what passes for theory is better described as concept formation. It tends not to identify the parameters of administrative systems or the relationships between them clearly enough to allow an investigator to deduce the set of variables that are necessary or sufficient to systematically organize and understand the variation in administrative systems. Therefore scholars and practitioners in public administration tend to begin with *classification* of the variation. Classification is also useful to facilitate the routinization of responses to individual cases, aid in summarization, and make others aware of differences between subclasses.

Classification as a quantitative technique abstracts from, formalizes and generalizes the processes of human reasoning. In the classical view, human reasoning is a process through which people obtain knowledge on the basis of abstract propositions that can be objectively either true or false (if not meaningless). The capacity for such reasoning has traditionally been posited to be something transcendental in the sense that it goes beyond the physical limitations of the person. In other words it does not have any bodily, organismic or natural basis. Accordingly classification, broadly construed, is considered to be the main way people make sense out of their experience: it is integral with the human capacity for meaningful thought. It is the process whereby subclasses of experience are characterized by the person according to his or her perceptions and understandings of the attributes shared by their members. Other credible views of classification are feasible (Gardner, 1985).

Operationally, classification is the act of distinguishing between subclasses of empirical entities. A subclass is formed of a number of such entities, each of which exhibits a definite characteristic in a constant manner. In the classical view, the definite characteristic may be any arbitrary division, normally determined culturally and linguistically. The operation involves the

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

mutually exclusive and collectively exhaustive assignment of empirical entities to subclasses, according to this definite characteristic. In other words, each entity is placed into one and only one subclass, such that the definite characteristic is exhibited (a) in a constant manner by all members of the subclass and (b) in a manner that is different than is exhibited by entities not in that class. There may be no overlap between members and nonmembers of a subclass. According to our definition, measurement occurs when numerals are assigned to designate the subclasses.

The overt use of classification is often criticized due to the fact that when one places an empirical entity in to a subclass, one inevitably loses some information about it. If, for example, one classifies a person as an ''executive,'' a ''manager,'' a ''technician,'' or a ''street level bureaucrat'' one asserts that everyone in that subclass is somehow essentially similar. Moreover, the information lost may be significant from some points of view. For example, the subclass of executives may include a young female from Taiwan with a Ph.D. in psychology and an elderly male from the United States with a bachelors degree in engineering. The two may be very different in all respects except that both are classified as executives. Thus, not classifying entities together may certainly avoid erroneous generalizations. But unless one classifies entities one cannot handle them in a small enough set of groups to enable generalization, making science impossible. Therefore it is worth emphasizing in this light that classification is only a conceptual device to facilitate the handling of information in a scientific or coherent manner. Classification says nothing about whether or not entities can ''really'' be considered equal with respect to many of the important characteristics not included in the definitions of the subclasses.

The product of the classification process may be termed a ''classification scheme.'' The classification scheme may be viewed in either one of two ways, depending upon whether it is conceived to group the (1) the empirical entities or (2) their characteristics (Kendall and Stuart, 1966). In the first view, the entire scheme may be considered one nonordered polytomous variable that measures the empirical entities themselves. In this view the value assigned to an observation designates its subclass. Numerals with nominal properties are assigned to designate the subclass for an empirical entity. Alternatively the scheme may be viewed as an amalgamation of a set of related variables used to specify the characteristics of the entities. In this view, each variable is seen to correspond with one of the characteristics used to evaluate the empirical entities. The value assigned to an observation designates a magnitude for the characteristic, as expressed by the entity in question. The measurement properties of the assigned numerals thus depend upon the type of gradation to which the particular characteristic admits. It is entirely conceivable that the numerals might have nominal, ordinal, or even interval properties, depending upon the nature of the characteristic.

There are two basic tasks in creating a classification scheme. One is to construct the categories or scheme of characteristics to be used in distinguishing between empirical entities. The other is to assign each empirical entity to the appropriate category. Strictly speaking, a typology may be considered to be the product of a deductive approach to connecting these two tasks (Mukherjee, 1983). In this strict sense the typological approach to classification starts with the categories and then deduces the appropriate subclass for any given entity from there.

With respect to public personnel systems, a position-classification scheme may be considered a typology. Such a typology is usually an abstract organization of job-types, arranged according to the nature of the work performed. Accordingly, a job in any given agency is assigned a classification based upon comparisons between statements of the nature of the work to be performed in that job and statements about ''typical'' jobs that are grouped according to the typology. The assignment of a classification to a particular job is typically *deduced* from the classification scheme.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

In contrast, a typology may be distinguished from other classification schemes created by an approach in which one starts by enumerating the variations in the relevant characteristics of the entities and then proceeds to work *inductively* to the scheme. Mukherjee (1983) refers to this later form of classification as the ''population approach.''

There is some doubt as to whether the distinction between an inductive and deductive approach to the creation of a classification scheme is useful in the field of public administration. The distinction may be too finely-detailed to fit within the range of practical ideas in such an applied field. In any case, for better or for worse, in public administration the term ''typology'' tends to be used to refer to essentially any reasonably-definite classification scheme, regardless of whether an inductive or deductive approach is taken to its creation.

B. Indexing

The term ''index'' is commonly used in a number of different ways. In situations in which one wants to compare a given value of a time series with an earlier ''benchmark'' or reference value, the term ''index'' may refer to a ratio of the form:

$$\text{Index number} = \frac{\text{Comparison number}}{\text{Base number}} \cdot 100 \tag{1}$$

For example, one may assume that the number of violent crimes in a given city in a given base year, say 1995, was 624. Furthermore that the following year the number rose to 714. This form of the index value would be approximately equal to 116, meaning that the number of crimes in 1996 was up 16% in comparison to 1995. The ratio of the two quantities is multiplied by 100 so that when the comparison number equals the base period number, the resulting index value will have a value of 100. This, however, is *not* what the term ''indexing'' refers to in this chapter.

1. Index Construction

In proper use, as examined in this chapter, the term ''index'' refers in general to any value, I, which contains a set of empirical variables, $x_1\ x_2\ x_3\ \ldots\ x_n$, combined in such a way as to represent the concept of interest. That is to say, the index is some function of the empirical variables, such that $I = f(x_1, x_2, x_3, \ldots, x_n)$. When constructing an index, one must concern oneself with deciding upon what empirical variables to use, how to measure the variables, how to weight them, and how to combine them.

One may take as a typical example of an index the concept called ''cost of living.'' The concept refers to the expenditures required to maintain a constant level of satisfaction (Mansfield, 1982). Various cost of living indexes have been constructed, all of them closely associated with the measurement and problems of inflation. They are used to measure changes in the purchasing power of the dollar for a wide variety of purposes. Probably the most famous cost of living index is the Consumer Price Index, which the Bureau of Labor Statistics has been constructing for over sixty years. It is one of a number of possible proxies for the concept labeled ''cost of living.'' It includes a set of empirical variables reflecting practically everything people buy for living—food, clothing, homes, automobiles, household supplies, house furnishings, fuel, drugs, doctors fees, rent, and transportation among other things. It refers to data gathered by personal visits to about 25,000 retail stores and service establishments in urban areas. The tremendous influence it carries is attributable to the fact that it reduces the complexity inhering in the concept of cost of living, and indicates its value for a particular time and place in an intuitively plausible manner. Other examples of indexes which may be put together in areas related to public adminis-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

tration include job performance; agency performance; gubernatorial power; fiscal capacity; industrial production, concentration or association; legislative professionalism; small business optimism; voter or consumer confidence; and sustainable economic welfare, among many others.

a. Deciding Upon Empirical Variables for an Index The first concern in constructing an index is that one select variables that measure what one wants to measure, given the purpose to which the index will be put. Good empirical investigation always starts with a clear statement of purpose. The empirical variables one selects should adequately represent the universe to which the concept is to be applied.

For example one may construct an index of the risk faced by a municipal government for purposes of selecting between insurance portfolios. One cannot predict all of the possible mishaps faced by the city. The fire department is familiar with causes of fires and how to minimize their outbreak, but they cannot precisely predict the time, location, or cause of particular fires. The police department deals constantly with burglaries and vandalism, but they cannot precisely predict them. Similarly, the accountant knows how to prevent defalcations; public works personnel know about the construction and maintenance of buildings and infrastructure; and building custodians know about dangerous conditions and practices. But no matter how much care is taken to avoid mishaps, some accidents are bound to occur. Property damage and adverse liability claims may result. Because some of the relevant probabilities, outcomes or costs always remain undetermined, one cannot precisely measure the risk faced by the city. Accordingly, the risk manager who is constructing an index to help select between portfolios will want to be sure to include a set of empirical variables that adequately represents at least the highest risks. These may include estimated damage to real and personal property, property loss, loss of income or increased costs that ensue from property losses, and liability associated with various possible mishaps in each of the major branches in the municipal government.

b. Deciding How to Measure the Variables for an Index Assuming that one has decided upon which empirical variables to include in an index, the decision may arise as to how to measure them. Not all variables in an index are necessarily measured the same way, using the same level of measurement. While aside from common sense there are no hard and fast rules to use at this point, some rough guiding principles are available.

First, all else equal, subject to the psychological and mathematical constraints noted above, constituent variables with levels of measurement containing more information are normally to be preferred to those containing less. Higher levels of measurement contain more information than lower levels. For example, interval measurements contain more information than do nominal measurements. While the degree of gradation to which an empirical variable admits may at some point limit the feasibility of a higher level of measurement, the information content of higher levels of measurement is richer in comparison to lower levels. Consider the simplest possible case, in which one nominal measurement is compared with a single interval measurement selected from along a point in a gradient. The information contained in the nominal measurement may signify only the existence or the non-existence of the characteristic. It conveys at most one bit of information.[2] The interval measurement, in contrast, signifies the existence or the nonexistence of the characteristic, and when it exists it further designates a point on the gradient. The additional number of bits yielded is a function of the logarithm of the total number of points on the gradient that can be discriminated. So long as the value to the investigator from the increase of information is larger than the costs associated with the extra effort required to obtain the higher level measurements, the higher level is to be preferred.

This raises the second guiding principle. If two potential constituent variables equally meet the investigator's purpose, those obtained with greater ease and less expense are to be preferred to those obtained with more difficulty and cost. As a rule, measurements con-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

taining more and better information cost more to obtain than do measurements containing less. The key is for the investigator to consider the purpose for which the index will be used. On one hand, there is no point in spending time, energy, and resources for increased information without adequate purpose. On the other hand, if the index serves an important enough purpose then the added resource expenditures required for higher quality information may be justified.

Another related principle is that measurements containing relatively little error are, all else equal, to be preferred to those containing more. Measurements containing less error tend to require more time and energy to obtain. Measurement error was introduced above. Error increases uncertainty and, in turn, uncertainty may exact a price. In general measurements that contain less error cost more than ones that contain more error. If the additional cost of better measurements can be justified then they are to be preferred.

Finally, there are a couple of guiding principles for measuring categorical variables. Categorical variables are those involving either nominal or ordinal measurements in several categories. For example, if one assigns numerals to individuals to represent their department in an organization, one obtains categorical measurements at a nominal level. If one assigns one of five numerals to individual working adults to represent their degree of educational attainment, one obtains categorical measurements at an ordinal level. There are a couple of rough guiding principles useful in the construction of categories. First, those that adequately represent all of the variation in the variable are to be preferred to those that do not. The number of categories should be small enough to be manageable. Seven, plus or minus two, is a reasonable rule of thumb. Each category should also contain some of the variation. For example, if one has seven categories and all of the variation is contained in two of them, something is probably wrong. It may also be a good idea to reconstruct the categories so as to have some of the observations represented in each category.

c. Deciding How to Weight the Variables for an Index Not all of the variables for an index are necessarily of equal importance in representing a concept. In this context, a weight is a numerical value that is presumed to reflect the importance of a particular empirical variable for an index. Whether the weights in question are the same or different across all such variables, multiplying the weight by each value of the variable renders the appropriate importance for that variable in terms of the index. Multiplication of weights by the values of a variable assumes that the weights are measured at least at a ratio level and that the values of the variable are measured at least at an interval level.

Weights may be obtained through various techniques, all of which to some extent involve the subjective judgments of an expert or judge. The most common approach is through direct assessment, in which a judge directly produces the numerical values for the weights subjectively, on the basis of his or her experience and capacity for judgment. A less common but often more sophisticated approach is to use indirect assessment in which an analytical tool such as the Analytic Hierarchy Process (Saaty, 1988), regression analysis, or mathematical programming is used to determine the weights mathematically. At times, remarkable structural similarities may be found between the subjective and objective elements of some of these techniques (Bowen, 1990). In any case, the investigator must designate some weights to the variables, even if they are all equal to unity. Though the element of subjectivity invariably raises the suspicions of many scientists, some scholars argue that such subjective judgments are an inevitable part of every index (Rescher, 1970).

d. Deciding How to Combine Variables for an Index Indexes involving more than one variable all assume a functional form with which to aggregate variables. The functional form is the overt form of the functional relationship between the variables in the index. That is, given an index, I, such that $I = f(x_1, x_2, x_3, \ldots, x_n)$, a decision must be made regarding how to operation-

ally combine the variables. There is no hard and fast rule, however, a couple of guiding principles are available.

First, a functional form theoretically rooted in mathematics and quantitative reasoning is clearly preferable to one without such roots. Take for example the assumption known as ''additive independence.'' If two variables are not additively independent, to add them together is mathematically incoherent. That is, the assumption of additive independence requires that the two variables do not interact. If for a specific example an investigator is attempting to construct an index of personal expenditures on clothing by using variables that reflect the person's sex and whether or not he or she is a college graduate then it is coherent to add the two variables together only if expenditures on clothing for, say, females relative to males is not affected by whether or not they are college graduates. If graduating from college differentially effects the clothing expenditures of females relative to males, then the assumption of additive independence is untenable. In this case, to the extent that the two variables interact, the functional form of the relationship between them is not additive but rather multiplicative. Thus we say that a mathematically coherent functional form is preferable to a merely expedient one.

Secondly, all else equal, it is preferable to postulate a simpler functional form rather than a more complicated one. More basically while it is a good idea to simplify the world as much as is reasonable, it is not a good idea to simplify it more than that. Again, the important thing is to bear in mind the purpose of one's investigation. *All* indexes abstract from and simplify the world. Without direct knowledge of the world from which the concept is abstracted, there is no way to know for sure whether the more complicated functional form is a better description of the true relationships between the variables in question in the actual world. Thus there is no final basis from which to compare the indexes under the simple and more complicated functional forms. The better question therefore is whether the increased complexity associated with the more complicated functional form sufficiently enhances ones ability to achieve the purpose of the investigation. Unless the more complicated functional form is somehow demonstrably superior to the simpler one, the simpler functional form is preferable.

Finally, one does well to bear in mind that it is prudent to respect the measurement properties of one's measurements when constructing an index. The integrity of the index depends upon preserving and accurately expressing the information content of each constituent variable through the aggregation process. The measurement properties of an index are determined by the measurement properties of it's lowest level constituent variable. This variable restricts the permissible forms of aggregation. If for example the index contains a nominal variable then addition is mathematically incoherent. Neither addition nor multiplication for any level of measurement below an interval scale is coherent unless the operation involves the addition or multiplication of the values of one's variable by a constant (in which case the measurement is meaningful and the properties are those of the lower order measurements). When the measurement properties of the constituent variables are ignored in the process of combining the variables for an index, the index no longer contains the force of the logic of measurement. While the index may in this case have the appearance of being a meaningful measurement, such appearance is, strictly speaking, illusory.

2. Differences Between Indexes and Scales

Although the term ''index'' is commonly used interchangeably with the term ''scale,'' the two may be construed to have clearly distinct meanings. Moreover considerable confusion may be easily avoided by bearing this distinction in mind.

First, the central concern in constructing an index is to simplify reality enough to allow the investigator to more-or-less match it to his or her concept. In contrast, the central concern

of scaling is to validate the empirical characteristic of interest. Scaling will be discussed at length in a following section.

Second, the type of theory used to construct an index is basically different than that used to construct a scale. The theory used to make an index is primarily phenomenological. Accordingly, the word ''scale'' is properly used to refer to the mathematical process and quantitative reasoning techniques employed to discern and substantiate the existence of one or more defined characteristics of an empirical entity and to establish operational indices of the relative magnitudes of those characteristics. The term ''index,'' on the other hand, refers to an empirical variable or set of empirical variables used as an indicator or proxy for an abstract concept. An index in this sense is constructed to represent the concept of interest in relation to a definite segment of the empirical world, without regard to its dimensionality. This is accomplished by measuring many seemingly different empirical variables, and somehow combining them so as to reduce the real world's complexity enough to represent the concept of interest using a single number.

C. Content Analysis

Content analysis is a dynamic technique for making inferences about the content of recorded text. Such content may be referred to as ''sign-vehicles.'' The term ''sign-vehicle'' refers to whatever units of content, document or form of recorded text contains the particular information or signal of interest in the investigation (word, theme, story, article, and the like). The technique is dynamic in the sense that the definitions of content analysis have changed over time with technical innovations and application of the tool itself to new problems and types of materials. A couple of representative definitions are as follows:

> ''Content analysis'' may be defined as referring to any technique a) for the classification of the sign-vehicles, b) which relies solely upon the judgments (which theoretically may range from perceptual discriminations to sheer guesses) of an analyst or group of analysts as to which sign-vehicles fall into which categories, c) on the basis of explicitly formulated rules, d) provided that the analyst's judgments are regarded as the reports of a scientific observer (Janis, 1949, p. 55).

> Content analysis is a phase of information-processing in which communication content is transformed, through objective and systematic application of categorization rules, into data that can be summarized and compared (Paisley, 1969).

In the early stages of development, content analysis was considered to be a simple descriptive tool. Later, it was developed into an inferential tool through the creation of techniques that transformed the sign-vehicles into comparable data. In this chapter, content analysis is defined as a scientific data analysis technique which meets the following requirements:

a. systematic inclusion and exclusion of relevant sign-vehicles regardless of the researcher's personal preference,
b. each step in the research process must be carried out on the basis of explicitly formulated rules, and
c. the findings of the content analysis must have theoretical relevance.

When these requirements are met, content analysis may, among other things, be used to generate cultural indicators that point to the state of beliefs, values, ideologies, or other aspects of cultural or linguistic systems (Weber, 1985).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1. The Requirements of System, Objectivity, and Generality

We noted earlier that measurement enhances one's ability to reason systematically through the relationships between one's concepts and the empirical entities of one's interest. In content analysis the systematic inclusion and exclusion of relevant sign-vehicles is designed to ensure that the analysis of the content of text is done according to consistently applied rules, so as to ensure as much validity as possible (Holsti, 1969). This is particularly valuable in an era of information overflow, in which a subjectively biased investigator could find enough written materials to conduct a quantitative study to support his or her beliefs about most anything.

Content analysis is based upon sampling rules which, when properly applied, clearly eliminate those analyses in which only materials supporting the investigator's predispositions are admitted as evidence. The sampling rules achieve this purpose by guiding the investigator's decisions in the process of delimiting the analysis. Often the first such decision is how to take a potentially tremendous volume of text related to any given topic and reduce it to an analytically manageable one. There is often no clear and universally applicable normative criteria with which to systematically identify the most important sources of text in a way that avoids the subjective prejudices of the investigator. The sampling rules prescribe that one way to avoid subjective prejudice of the investigator is to use pooled experts' judgments about the relevant material. Another way is to use some quantitative criterion to select sources of text. For example, in Bowen's (1996) content analysis of classified ads in Taiwan, she selected the two Taiwanese newspapers with largest circulation as the sources of text and conducted a content analysis of the personnel classified ads during the same month over two years.

The requirement of objectivity in content analysis stipulates that each step in the information analysis process must be carried out on the basis of explicitly formulated rules and procedures. What categories are to be used? How is category A to be distinguished from category B? What criteria are to be used to decide that a sign-vehicle should be placed in one category rather than another? Objectivity implies that these and other decisions are guided by a clearly stated and explicit set of rules designed to minimize the possibility that the findings reflect more the investigator's subjective predispositions rather than the content of the text under analysis. One important fact to bear in mind in this regard is that objectivity can be replicated. In other words, any other investigator who is interested in testing the findings of an investigation should be able to come up with similar results when following the identical procedures with the same data.

The requirement of generality stipulates that the findings must have theoretical or general relevance. The requirement of theoretical relevance, for example, was met by Bowen's (1996) content analysis of personnel classified ads. The purpose was to test dual labor market theory by examining the employment opportunities for men in comparison to women. The requirement of general relevance may be met by comparing the results of a content analysis with other attributes of the documents analyzed, with documents produced by other sources, with characteristics of the persons who produced the documents, or the times in which they lived, or the audience for which they are intended. Examples of content analyses that met the requirement of general relevance include one of a sample of fifty years worth of articles from *Public Administration Review*. Bingham and Bowen (1994) did a content analysis of PAR as a means of characterizing the boundaries of mainstream public administration. Another content analysis analyzed 50 messages from 900 number services. The results were used to provide policy implications for the US Federal Communications Commission (Glascock and LaRose, 1992).

2. Reliability of the Content Coding Process

Content analysis entails reducing data by classifying many words into far fewer categories. The degree of difficulty of this data reduction process depends largely on the content unit chosen by the investigator. It is usually easier to classify smaller content units (i.e. words or phrases),

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

into categories than larger ones (i.e. themes, paragraphs or articles). This is because larger content units contain more information and greater topical variety. They therefore afford a greater chance of providing conflicting or uncertain cues.

An accurate coding process is the first step toward successfully dealing with this difficulty. As is the case with all analyses of data, the accuracy of the results depends upon the reliability and validity of the measurements. Inconsistencies in coding constitute a form of unreliability, therefore, the content coding process is critically important to a successful content analysis. A high degree of reliability is a minimum requirement for the coherence and believability of a content analysis.

The coding process, which is to say the process of classifying specified content units into categories, usually involves some degree of subjective and idiosyncratic judgment. Unless one properly checks the reliability of this process, the results of a content analysis will remain, at best, questionable. Appropriately trained coders, clear and well-defined content units, and clear, well-defined, theory-guided categories all tend to increase the accuracy of the coding process.

Three types of reliability are pertinent to evaluating the coding process: stability, reproducibility, and accuracy (Krippendorff, 1980). Among them, stability and reproducibility are used more frequently than accuracy.

Stability refers to the extent to which the results of content classification are consistent over time. This is the most lenient indicator of reliability. It can be calculated when the same content is coded by the same coder two or more times. Because the coder and the content stay the same, this type of reliability contains the fewest possible sources of uncontrolled variation. Such sources include inconsistencies in the written material, ambiguities in the coding rules, emotional changes within the coder or simple marking errors.

While stability measures the consistency of one person's understanding or interpretation of certain material over time, intercoder reliability measures the consistency of shared understanding or meaning of the text. Intercoder reliability, also called reproducibility, refers to the degree to which two or more coders replicated each other's results. The coding process is said to be reproducible if the coders coded the same text in the same way. Intercoder reliability is a more objective indication of reliability than stability. Inconsistent codings usually result from ambiguities in the text, cognitive differences among the coders, ambiguous coding rules or from random recording errors.

Accuracy refers to the extent to which the coding of text corresponds to a standard or norm. However, such a standard or norm seldom exists in the field of public administration. It more often pertains in situations such as for training purposes, when it is used to test the performance of human coders against preestablished standard for coding some text.

The type of reliability one selects to evaluate one's analysis depends on the criterion one uses to check the consistency of the coding. When the criterion is from the same coder but only at a later time, it is stability. When the criterion is from another coder, it is reproducibility, intercoder reliability. When the criterion is a previously established standard or norm, it is accuracy. The calculations of reliability prescribed specifically for categorical data, the most common form of data in content analysis, are available (Cohen, 1960).

Content analysis may involve nominal data. When it does, the agreement between coders may be computed using a Kappa coefficient. Its formula is:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

Where p_o = the proportion of units for which the judges agree
p_e = the proportion of units for which agreement is expected at random

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

When the observed agreement equals the agreement expected at random, k = 0. When the observed agreement is less than the agreement expected at random, k becomes negative values. When there is perfect agreement between two judges, k = + 1.00.

3. Necessary Steps for Designing a Coding Scheme

A well designed coding scheme is prerequisite to a successful content analysis. After the investigator has identified the relevant theories, found the important questions and made sampling decisions, the next step is to design a coding scheme. There are a series of necessary steps for designing a coding scheme (Weber, 1985).

First, one defines the coding units. There are six coding units commonly employed:

a. word, simply recording every word,
b. word sense usually referred as a semantic unit such as idioms or proper nouns,
c. sentence, recording meaning of the entire sentence,
d. theme, the definition of a theme as a unit of text has no more than one each of the following elements: the perceiver, the agent of action, the action, the target of the action and the situation,
e. paragraph, and
f. whole text.

Larger units contain more information or potential conflicting information than smaller units and may require more subjective judgment of individual coders, so it is usually more difficult to achieve high reliability when coding larger units than when coding smaller units. There is a trade-off, however. Larger units require less effort in the coding process and make the size of the coding load more manageable. No one coding unit is necessararily better than another in every case. Rather, the investigator needs to consider the purpose of the study, the available time and resources to make the most suitable choice of coding units.

Having defined the coding units, the investigator must next define the categories. The most important consideration in defining categories is to make sure that the definitions of the categories are exhaustive and mutually exclusive. In other words, each coding unit should be able to be assigned to one and only one category. The choices of category should be theory-guided. One way to create a satisfactory coding scheme is to make sure that the investigator is already familiar with results of previous studies, currently sampled materials, and relevant theories when creating the coding scheme.

The next step is to conduct a pilot test on the coding scheme. This involves selecting a small proportion of the text and carefully going through the coding scheme. Pilot testing not only provides a chance to clear any potential ambiguity in the category definitions but it also leads to insights in terms of revising the classification rules.

The pilot test allows the investigator to assess the reliability of the coding before doing any further analysis on the data. Before the investigator actually starts analyzing the data, the reliability of the coding process should be assessed. If the coding scheme is found to be unreliable then the results of content analysis will not be credible.

The pilot test also enables the investigator to revise the coding rules as needed. If the reliability is low, the coding rules must be revised. Studies show that clarity on the coding scheme increases measurement reliability in content analysis even more than does coder training. In other words, the reliability of untrained coders using clearly defined coding rules is higher than trained coders using ambiguous coding rules. After the revisions are made, the next step for the investigator is to do another pilot test and make further revisions until the coders reach

sufficient reliability. Once coder reliability is deemed on the basis of the pilot tests to be adequate, the investigator is ready to code all the text.

The final step is to assess the final reliability. After all the text has been coded, the final reliability should be assessed. The first steps do not guarantee a high reliability. Factors such as coder fatigue or subtle cognitive or mood changes of coders may still lead to unreliability. The rule is to never assume high reliability of all coded text until statistical assessment has been performed and sufficient evidence has been gathered. The advance of technology allows computers to replace human coders to do the coding. However, the principles apply to human coders still are applicable to the design of computer procedures before computers can do the coding reliably.

5. Analysis Tools for the Coded Data and Interpretation of Content Analysis

Statistical analysis tools for content analysis are similar in many respects to those used for any other types of data. The coded content analysis data is treated as is any other type of data. The criteria for selecting suitable statistical tools for the coded data are defined by the purposes for which the analysis is conducted and the measurement properties for the pertinent level of data (usually nominal, categorical, or ordinal). With this in mind, there is always more than one way to skin a cat. Data do not speak for themselves; the investigator must explain their significance in light of theoretical and substantive concerns. It is incumbent upon the investigator to explain what the data say and how he or she arrives at this understanding. Are there competing interpretations? If so, which interpretation makes the most sense in light of the statistical evidence and whatever theories or other knowledge pertains to the current situations? The answers to these and similar questions all involve some level of idiosyncratic judgment on behalf of the investigator. They ultimately depend upon the investigator's experience, knowledge, and capacity for judgment. Unbiased results from a content analysis results may only be achieved if all of the requisite subjective judgments are backed up with statistical evidence. Beyond this point, whatever statistical techniques are appropriate to establish this evidence, given the type of data one is working with, are suitable.

D. Meta-Analysis

While content analysis examines words, sentences, themes, paragraphs or whole texts for the purpose of making clear and systematic comparisons across different text materials, meta-analysis statistically combines the numerical results of previous studies on a specific topic. Recognizing that statistical research findings are inherently probabilistic (the results of any single study could have occurred by chance), meta-analysis utilizes statistical procedures to combine two or more empirical studies relating one variable to another (Hunter and Schmidt, 1990). The result of a meta-analysis is a more comprehensive and systematic synthesis of previous studies than would be feasible with a narrative review, limited by unaided human cognitive information processing and interpretation. The additional inferential power of a meta-analysis comes from placing all the results of each of the included studies into a single experimental design. This helps draw more precise conclusions about inconsistent findings in a particular area of investigation (Gaugler et al., 1987).

There are seven steps involved in conducting a meta-analysis: (1) conceptualize the relationship under consideration; (2) gather a set of studies that have tested the specified relationship; (3) design a coding sheet to record the characteristics of the conditions under which each study was conducted; (4) examine each study and, using the coding sheet; record the conditions under which it was conducted; (5) compute the ''effect size'' for each study (to be explained below); (6) statistically analyze the characteristics and effect sizes for all of the studies;

and (7) write a research report. The following sections briefly describe and illustrate the seven steps.

Step 1: *Conceptualize the relationship.*
The first step is to provide a detailed specification of the relationship to be examined, giving attention to the major theories and methods important in the literature. The investigator must define the X and Y (independent and dependent) variables in both theoretical and operational terms. This definition may set the boundaries of the literature under consideration.

Moderator variables, or study characteristics (W) are also important. These are variables that can be expected to change the direction or magnitude of the relationship between X and Y. They should be considered as clearly as possible. The greater the clarity given to X, Y, and W before the literature search, the stronger the review is likely to be.

Example. In Chang's (1993) review of the relationship between gender and performance appraisals, the independent variable, X, was defined as the gender of the ratee (male vs. female); the dependent variable, Y, was defined as a performance evaluation given to the ratee in a real work setting. Previous studies had shown inconsistent results in terms of the relationship between these two variables. Several moderators (W) were deemed important including gender stereotype of the job, group composition in terms of percentage of men and women, stereotype of the measurement, purpose of performance appraisal, amount of performance-related information, subjectiveness of measurement, rated position, and type of work setting were all coded as W variables. Theoretically-based expectations were developed to specify the influence of W on the relationship between X and Y for all W variables. For example, one of the W variables was the position of the ratee. Among all the positions in an organization, managerial positions usually hold higher prestige than professional, clerical, technical, or blue-collar positions. Moreover, in managerial positions, job tasks are varied, not predictable and the criteria for performance are relatively unclear. In this situation, nonperformance factors may enter the evaluation. The expectation was that in this sort of a job situation, because of the lack of clear performance-specific criteria, the rater will simplify the rating process by using sex-role stereotypes. Therefore males will receive higher performance evaluations (Auster, 1989).

Step 2: *Gather relevant source reports.*
We noted that clear definitions of X, Y, and W may be expected to set clear boundaries for the relevant literature. The next step is to locate and retrieve all of or at least as many as possible of the pertinent reports.

Not all studies containing the specified relationship between X and Y will be suitable to be included in the meta-analysis. For example, in Chang's meta-analysis, one of the studies used an atypical group of ratees—people with substance abuse records. Such atypical studies are likely to be found in the course of any meta-analysis and it is important for the investigator to spell out the reasons for such exclusions. Whenever possible effort should be made to include unpublished studies such as theses, dissertations, technical reports, and working papers.

The investigator should always thoroughly describe his or her methods of locating articles, along with descriptions of the criteria used for study selection and the reasons for rejection of studies. Guidelines one can use to locate and retrieve all of the pertinent studies include (Johnson, 1993):

1. Computer database searches can be used as a starting point to locate references or abstracts that contain keywords relevant to the topic specified by the investigator. A lot of different databases may be used. Keyword usage in the computer search is part of the key to a successful search. Usually investigators start out by putting the X and Y variables as the keywords. Then other words synonymous with the X or Y variables

should also be included in the keywords search. Different labels are often used to refer the same thing in the field of public administration. For example, gender, sex, men and women, man and woman, male(s) and female(s) are used interchangeably to refer to the same thing, so all of them need to be included in the keyword search.

2. The ancestry approach involves examining the reference lists of previous narrative reviews. One can start with the most recent articles and proceed to older articles.
3. The descendance approach involves identifying a critical piece of study in the literature, and trying to locate all the studies which cite it.
4. Networks may be contacted. This involves writing letters to other investigators who are known to work on the specified topic and asking whether they know of any other unpublished studies.
5. Manual searches of important journals may be conducted. Although manual searches may be old-fashioned, they may still turn up some articles that are overlooked by other techniques.

Step 3: *Design a coding sheet.*
Although each study in the selected set of studies examines a single clearly specified X–Y relationship, the conditions under which the relationship was examined (W) may vary from study to study. These conditions must be considered. The coding sheet is designed to record the characteristics of these conditions. These characteristics may be used later on to explain any inconsistencies in the results of different studies.

Step 4: *Code study characteristics*.
Having gathered the relevant literature, the next step is to record the important characteristics (W) of each study. It is important to record all moderator variables. Since they could alter the direction and/or magnitude of the relationship between X and Y, they become extremely important when the investigation tries to integrate the findings of previous studies.

Study characteristics may be either continuous or categorical. Categorical characteristics reflect qualitative differences among different values of the relevant variable while continuous characteristics reflect real-valued quantitative differences. For example, in Chang's (1993) study, publication form, sex of first author, type of work setting, purpose of the performance appraisal, rater's gender, rated position, rating instrument and type of rater were all categorical characteristics of the various studies. In contrast, year of publication, percentage of male authors, clarity of the presentation, amount of training time on rating scale usage, familiarity of raters with ratees' performance, degree of rater-ratee interdependence, percentage of male incumbents in the organization, sex stereotype of the job, and sex stereotype of the measurement were recorded as continuous characteristics.

Step 5: *Compute effect sizes.*
An effect size computation is a standardization process through which the strength of the X–Y relationship in an individual study is expressed as standard deviation units. These units are defined with reference to the summary statistics used to describe the X and Y variables in that particular study. They may be computed in different ways depending upon the particular summary statistics provided in the individual study source report. The goal of effect size computation is to convert the summary statistics provided in the individual report into standard deviation units that may be statistically integrated across studies. The direction and number of standard deviation units computed for a particular study is termed its ''effect size.'' When the effect sizes are properly computed, they may be used to aggregate or compare the studies for purposes of overall summary description and statistical inference.

The effect size may be referred to as ''**g**.'' The specially designed computer software for meta-analysis, DSTAT, allows the following source report summary statistics to be converted to **g** easily: (a) means and standard deviations; (b) t-tests or F-value from analysis of variance

(ANOVA); (c) correlation coefficient, r-values; (d) Chi-square; (e) proportions or frequencies; (f) exact p-values. Details of DSTAT usage will not be discussed in this chapter. Instead interested readers should refer to Johnson's (1993) DSTAT manual.

Some studies may yield more than one effect size. This occurs when the X or Y variables are operationalized in more than one way in a particular study. For example, in Chang's (1993) study, in concept the Y variable represented performance appraisal. But in several studies the concept of performance appraisal was operationalized in slightly different ways, even within the same study. Some studies operationalized it in terms of both reported productivity and customer satisfaction. Some studies gathered the data in more than one organization. Some used both self-rating and supervisory-rating. In all cases in which the variables are operationalized in more than one way within the same study, if enough data are available then more than one effect size may be computed.

When multiple effect sizes are computed for a study, they may be combined. Combining multiple effect sizes avoids the fallacy of overweighting those studies with multiple effect sizes in the meta-analysis process. To combine multiple effect sizes, one may simply average them or, alternatively, compute Rosenthal and Rubin's (1986) ''Composite g.'' The advantage of computing the Composite g is that it corrects the underestimation bias that inheres in simple averaging. Composite g may be computed if the source reports provided sufficient statistical information about the intercorrelations between the multiple Y variables. Details of calculation composite g will not be discussed in this chapter. Interested readers should refer to Rosenthal and Rubin (1986)

Step 6: *Analyze the data.*

The next step is to combine effect sizes and determine their overall mean and consistency with respect to all of the source studies. Reference to the study characteristics (W) may provide the required explanations of any inconsistencies noted between studies.

Another common reason for inconsistencies between studies is the fact that different studies contain different sample sizes. Specifically, the results of a study with a large sample size are usually more stable than are the results of one with a small sample size. Therefore, when the source studies contain a large variance in sample size, before conducting any further data analysis, the effect size for each study should be weighted. To accomplish this, the reciprocal of the variance for the Y variable is used as a weight. In the process of combining the effect sizes this weight is multiplied by the effect size in a particular study to adjust for the various degrees of stability of the results from the various studies. This process tends to give small weights to studies with large variances, and large weights to studies with small variances. The weighted effect size is referred to as ''**d**'' in this chapter. Once **d** is obtained, the investigator is ready to begin the analysis process.

The analysis process begins with the computation of an average effect size for all of the **d** values. These averages are used to assess the magnitude, direction, 95% confidence intervals, and homogeneity of the overall effect sizes in the combined data set. If all the effect sizes present a homogeneous picture, then the investigator may draw conclusions based on the magnitude, direction and significance of the average effect size. When the 95% confidence interval includes zero, the average effect size is not different from zero. In this situation it may be concluded that there is no relationship between X and Y across all the source studies. When the 95% confidence interval does not include zero, it may be concluded that across all the studies there is a significant relationship between X and Y.

Experience unfortunately shows that most effect sizes are heterogeneous across studies. Heterogeneous effect sizes mean that individual study outcomes are quite different from each other in terms of the magnitude and/or direction of the X–Y relationship. One way to try to attain homogeneity in heterogeneous cases is to identify outliers among the effect sizes and

sequentially remove those that reduce the heterogeneity statistics by the largest amount (Hedges and Olkin, 1985). Usually homogeneity is reached by removing as few as 20% of the largest outliers in the combined data set. When the effect sizes are heterogeneous, the mean effect size does not adequately describe the study outcomes so further work is needed. Another way to try to attain homogeneity is to do statistical model testing, though this approach is fraught with statistical difficulties when sample size is small.

In heterogeneous cases, the study characteristic variables (W) may be used to statistically account for some of the variability. Both categorical and continuous study characteristics may be used for this purpose. With respect to categorical characteristics, categorical analysis such as analysis of variance may show that heterogeneous effect sizes are indeed homogeneous within the subgroups established by dividing the source studies into classes based on the study characteristics, and furthermore that the classes differ in the mean effect size they produce. Such analysis may be used to estimate both a between-class effect and a test of homogeneity of the effect sizes within each class. With respect to continuous characteristics, on the other hand, linear analysis may be used. Ordinary least squares regression is commonly used for this purpose. The goal in such analysis is to use the moderator (W) variables to statistically account for as much as possible of the variation in the effect sizes. Each such linear analysis yields a test of significance of each moderator variable as well as a specification test which evaluates whether significant systematic variation remains unexplained in the analysis.

Oftentimes the W variables are not successful at explaining the variation in the effect sizes. For example Chang (1993) tested all possible study characteristics in her meta-analysis of gender and performance appraisals. These included stereotype of the measurement, subjectiveness of the measurement, number of items in the work performance scale, job stereotype, group composition of men and women, rater training, familiarity of rater with ratees' performance, publication year, and percentage of male authors. She found only one of them, stereotype of the measurement, to be significantly correlated with effect size. When this sort of thing occurs, outlier elimination and statistical analysis may be used iteratively. Meta-analysis is a trial-and-error process, not an exact and prescriptive science. One clear guiding principle is that it is better to record some study characteristics and find one does not need them in the analysis than it is to find oneself needing information about some critical study characteristics that were not recorded in the first place. Beyond this, the exact combination decided upon will depend among other things upon the number and heterogeneity of the effect sizes, the number of W variables, and whether they are categorical or continuous. The best guideline is for the investigator to choose the most parsimonious and convincing possible combination of outlier elimination and statistical analysis.

Step 7: *Write the report.*

The process of writing a research report to describe the meta-analysis process and results is not different from writing any other report. The primary elements of such a report are (a) abstract, (b) introduction, (c) methods, (d) results, and (e) discussion. A well-written guideline can be found in the *Publication Manual of the American Psychological Association* (1994).

One section of such a report requires special attention in a meta-analysis study. The ''methods'' section should include: (a) procedures for locating and retrieving previous studies; (b) criteria for including and excluding studies; (c) coded study characteristics and a measure of the reliability of the coding process (similar to one used in content analysis); (d) effect size calculations; and (e) data analysis tools.

A meta-analysis, should always include an appendix containing the references of the source studies in the sample. The list should contain all the studies *before* the investigator conducts the outlier elimination procedures. Such a list is helpful for future studies and for reviewers to judge the completeness of the sample.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

E. Scaling

Of the techniques discussed in this chapter, scaling is by far the most highly developed. Early research in scaling dates back to at least the psychophysical experiments of Fechner in 1840 (MacKay, 1988). Since then, knowledge about scaling has cumulated progressively. Scaling is the only one of these techniques with foundations that come appreciably close to the status of fundamental theory.

The term ''scaling'' refers to the processes and techniques used to empirically test and validate the existence of the properties or attributes to which a concept refers and to establish operational indices of their relative magnitudes (Gorden, 1977). Though scaling is seldom used in public administration it does have considerable potential use-value in the field, largely as a means of quality control for knowledge claims.

A principle of scientific method holds that one must verify or empirically test and validate one's inferences about the world prior to their acceptance. This principle often makes it difficult, at best, to make scientifically acceptable inferences about many of the abstract and vital concepts in public administration, such as ''risk,'' ''attitude,'' ''efficiency,'' ''effectiveness,'' ''performance,'' and ''leadership'' among many others. Scaling extends the logic of measurement into the realm of many such concepts as these, allowing us to test for the validity of a wide range of inferences that include them. In doing so, the techniques of scaling can go a long way toward overcoming the difficulties of making scientifically acceptable statements about a wide range of things of interest in public administration. Scaling may also be used to (1) graphically represent or otherwise simplify the description of a complex data set or (2) give scores to individual entities in relation to groups of such entities.

1. Scaling as the Mathematical Representation of Behavioral Data

Scaling applications result in mathematical representations of the relationships among the attributes of empirical entities. The empirical entities may be visualized as having a dual nature; which is to say as being composed of both ''objective'' and ''subjective'' aspects or dimensions.

In scaling theory, the empirical entities one scales may be termed ''stimuli.'' For example, in an effort to effectively select the best candidate for a position, a supervisor may desire to scale the candidates according to their expected ability to fulfill the demands of the position. In scaling theory, in this situation the ''stimuli'' would be the candidates. The field of public administration contains many possible sets of stimuli for scaling. The minimum condition for a set of stimuli to be scaled is that they must all be experientially real or otherwise meaningful to the people who provide the data. The people who provide the data are termed ''respondents.'' In this case, the respondents are the members of the selection committee tasked with reviewing the candidates.

It is mathematically and technically feasible to scale most any reasonably well defined set of stimuli. Ideally the scale may be used to operationally define a concept. This is done by using scaling theory and technique to structure the relationships identified between the stimuli thought to be deductively subsumed under the concept. In other words, having identified the stimuli, one then posits the types of relationship one expects between them. One then operationalizes the concept by mathematically representing these relationships in accordance with scaling theory.

The stimuli are presented to the respondents using ''items.'' Scaling data are responses to the items. An item is a single question or statement to which the respondent provides a numerical response or judgment. These judgments reflect the respondent's perception or evaluation of the stimuli. Evaluative judgments always presuppose perceptual judgments. Items may be formulated in different ways depending upon the type of information the investigator is

seeking and the particular scaling technique he or she is using. The measurement properties of a scale depend upon these formulations. Most techniques of scale construction are designed around particular item formats. The entire set of items required to mathematically describe the respondent's judgments is normally called the ''scale.''

Four basic types of scaling data may be identified (Coombs, 1964). The first are *preferential choice data*, which reflect a respondent's ranking of a set of empirical entities (stimuli) according to one or more criteria. An example would be when a group of decision makers select between a set of sites for the location of a new public works facilities (MacKay, Bowen, and Zinnes, 1996). All the members of the group may agree in principle that sites with more utility would be preferred to sites with less. The group member's preferences between the alternative sites may however differ due to differences in how they perceive the criteria as well as how the sites relate to the criteria. The data that are scaled in this case are the group members' stated preferences for the choice set of sites.

The second type are *stimulus comparison data*. Respondents are presented with two or more stimuli at a time and are asked to determine which of them has more or less of whatever dimension is being scaled. An example would be pairwise comparisons of the relative risk associated with various global environmental issues (Bowen and Haynes, 1994). Specifically, global warming and habitat destruction, along with other global environmental issues, may be defined as two such stimuli. The respondent may be presented with these two stimuli and asked to judge which poses the greater risk to our long term security and well-being. The respondent may judge that global warming poses the greater risk. The dimension is relative risk. The numeral used to indicate the relative location of the two stimuli on the dimension is the respondent's subjective numerical judgment of the relative magnitudes of the two risks.

The third type of data are *dissimilarities data*. Dissimilarities data reflect the respondents' judgments of the dissimilarity between two stimuli in terms of the criteria (Bowen, 1995). An example of dissimilarities data might be a judgment of the dissimilarity of the leadership styles of all of the pairs in a group of executives.

The fourth type are *single stimulus data*. As the label for this type implies, the data do not reflect any comparisons between stimuli. Likert scaling, discussed in a following section, is an example of single stimulus data. Though at times the distinctions between these four types of data become blurred, together they give great flexibility to scaling, enabling its use in a wide range of applications.

The dual nature of the mathematical representation of the data may be seen in its objective aspects, as an ''objective space,'' insofar as the dimensions in the data correspond to or reflect the attributes of the stimuli as they actually exist. In other words, the dimensions may be conceived of as being defined by the objective measures used to describe the stimuli. In contrast, in its subjective aspects, the mathematical representation consists of the locations of the stimuli as they relate to the dimensions revealed in the judgments of the respondents. The dimensions in their subjective aspects are posited to correspond to or reflect the attributes of the stimuli as they are assigned to them by the respondents. The objective and subjective aspects together may be termed an ''attribute space'' (Green, 1989).

If the respondents or the stimuli or both may be deductively subsumed under a concept then the attribute space may be considered as an operationalization of that concept. Take for example the use of scaling to operationalize the concept of attitude (Shaw and Wright, 1967). Psychologists define attitudes in relation to classes of objects. These objects comprise the set of stimuli that are perceived to be involved with the concept. The scaling technique provides an exact and replicable set of instructions or descriptions of sets of actions or operations for the investigator. The result of following these instructions is an attribute space that contains information about either the respondent, the stimuli or both the stimuli and the respondent. The

attribute space may be considered an operationalization of the concept. Essentially any reasonably well defined set of stimuli may be scaled, so a similar approach may in principle be followed to operationalize many of the central concepts in public administration.

An attractive feature of scaling is that a scale may yield valuable information even if the objective aspects of the attribute space do not agree with its subjective aspects. This is because the force of the logic of scaling originates in the mathematical and quantitative reasoning about the relationships between the stimuli or respondents, not from the stimuli or respondents themselves. The desiderata for a good scale is conformity of the scaling data with the mathematics and quantitative reasoning as stated in the scaling theory. If in any given instance a lack of such conformity characterizes the attribute space then such a fact tends to be made clearly explicit in various numerical indicators of the internal error-indicative conflicts of discrepancy, inconsistency and disuniformity. These indicators, which are often produced in the scaling procedure, suggest a shortcoming such as underconceptualization, poor operational definition, or a high degree of uncertainty on behalf of the respondents.

2. Some Techniques of Scale Construction

A variety of techniques of scale construction are available. One way that the various techniques may be distinguished is on the basis of whether the entities they scale are persons, stimuli, or both people and stimuli together. Some techniques are designed to locate the respondents in relation to a fixed set of stimuli in the attribute space. Some are designed to locate the stimuli in the attribute space for a fixed set of respondents. Some locate the stimuli in the attribute space over time for a specific respondent. And some locate both respondents and stimuli in the space for a fixed situation. Scaling techniques may also be distinguished on the basis of the ''traces'' or theoretical curves assumed to depict the mathematical relationship between the probability of a specific judgment on a item and the attribute or dimension that the item is intended to measure (McIver and Carmines, 1981). It is important to make sure that the technique one selects in principle enables one to scale the desired entities. Information about exactly which techniques are designed to scale exactly what entities may be found in the many fine scaling texts available in any research library.

Another way that scaling techniques may be distinguished is on the basis of whether the scale is unidimensional or multidimensional. The simplest way to state the difference between unidimensional and multidimensional scaling is with reference to the number of dimensions represented in the attribute space. Both unidimensional and multidimensional scales represent the entities or events one is investigating as relations between data-points in a geometrical space. Unidimensional scaling refers to the set of techniques used to establish the location of a set of entities along a single axis or dimension in the space. Only one coordinate is required to uniquely specify the point associated with the empirical entity in the space. Multidimensional scaling, on the other hand, refers to the set of techniques used to establish the location of the entities in **k**-dimensional space. In multidimensional scaling one requires **k** independent coordinates to uniquely specify the point associated with an empirical entity in the space. There are numerous algorithms available for either type of scaling. One may select between unidimensional scaling and multidimensional scaling and between algorithms within each of them, depending both upon one's purpose and the properties of the concepts and the entities to which they are applied in the investigation.

a. The Unidimensional Scaling Techniques of Thurstone, Likert, and Guttman Unidimensional scaling refers to the techniques designed to locate stimuli and/or respondents along a single dimension. Probably the most frequently used unidimensional scaling techniques are those associated with the names Thurstone (1929), Likert (1932), and Guttman (1944).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Early investigations using unidimensional scaling were unable to empirically test whether a set of items actually belongs on the same single dimension and what position the items occupy on that dimension. Due primarily to a lack of adequate scaling theory it was necessary to merely assume unidimensionality, without performing the analyses required to determine whether the data conform to the pertinent rules of mathematics and quantitative reasoning. For instance, in his investigations of peoples' attitudes regarding the immigration of racial and ethnic groups in the 1920s, Borgardus had to depend upon his own empathy and understanding alone to select and order items for measuring people's attitudes (Bogardus, 1929).

Thurstonian Scaling Over the following decades, Louis L. Thurstone, whose techniques of scaling have probably been used more widely than any of the others, developed the notion of ''equal appearing intervals'' and used it to enhance our ability to test the validity and reliability of scales. Thurstone also invented the method of paired comparisons. The method of paired comparisons may be generalized and applied in a wide variety of decision situations in business, public administration and policy analysis.

The steps involved in constructing a Thurstonian scale are: (1) A large number of items related to the attribute to be scaled are formulated; (2) these items are sorted by a sizable number of judges into eleven piles or categories which appear to the judges to be equally spaced in terms of the degree to which agreement with the item reflects the underlying attribute; (3) the piles are numbered from 1–11; (4) a scale value is computed for each item and taken as the median of the position on the attribute given the item by the group of judges; (5) the interquartile range is computed as a measure of interjudge variability; (6) all of the items for which there is much disagreement are rejected, (7) a small number of items for the final scale are selected so that they are spread more or less evenly along the attribute; and (8) the respondent is asked to check each item with which he agrees (Thurstone, 1929). His score is the median of the scale values of all the items checked. In this manner, theoretically, each individual should agree only with a few contiguous items near his or her actual position on the attribute. Thurstone took the situation in which a large proportion of the respondents checks noncontiguous items to indicate the multidimensionality of the scale.

While Thurstone's methods improved our ability to precisely locate each item on the postulated dimension, they still did not provide the concepts or techniques required to empirically test the assumption of unidimensionality. Perhaps the key contribution made by Thurstone was that he recognized the importance of the processes of selecting and assigning values to statements.

Likert Scaling Rensis Likert invented a widely-used scaling technique in which a large number of items are selected for the characteristic that the more favorable the respondent's evaluation of the stimuli, the higher his or her expected score for the item. Likert used a panel of judges to select an initial set of such items. The initial set was posited as a complete scale. The scale was then given to a sample of the target population, and the sample respondents instructed to indicate their response by means of a five-point rating system. The following is an example of a Likert response item:

__ Strongly Agree __ Agree __ Uncertain __ Disagree __ Strongly Disagree

These five categories are scored by assigning values of 5, 4, 3, 2, and 1 respectively. This scoring is reversed for negatively worded items. An analysis of the responses of the sample respondents was used to eliminate a subset of the initial set, on the basis of the internal consistency of the responses. Item scores are correlated to determine their internal consistency with total scores (the sum of the item scores), and items that correlated highly with the total score are selected for the final scale. Likert assumed that the intercorrelations of the items is attributable to a single

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

common factor or dimension to which all of the items are mutually related. The item score is assumed to be a weighted sum of this common dimension and an error factor specific to the item.

While the Likert method of internal consistency analysis thus provides the investigator with the ability to do a cursory evaluation of the unidimensionality of the scale, it does not enable the investigator to strictly establish whether or not a particular set of items actually belongs to a single dimension. No attempt is made to ensure the equality of units. Unidimensionality is sometimes inferred from high item correlations with the total score, but under certain circumstances an item which correlates highly with the total score does not belong on the same dimension with the other items in the set used to obtain the total score. Likewise, items with low correlations with the total score may belong on the dimension. In consequence, Likert scaling may not legitimately be said to validate a unidimensional scale. On the basis of mathematics and quantitative reasoning alone, contrary to much common practice Likert scales probably should be treated as having ordinal rather than interval properties.

Guttman Scaling It remained for Louis Guttman to devise a unidimensional scaling technique, scalogram analysis, that does legitimately validate a unidimensional scale. The technique assumes that items can be arranged in an order such that a respondent who provides a positive judgment for any particular item also responds positively to all other items having a lower rank. If items can be thus arranged, they may be said to have validity as a unidimensional scale.

To develop a Guttman scale, one starts by formulating a number of initial items posited as monotone along the dimension of interest. The set of items is administered to a group of respondents and their response patterns are analyzed to determine whether or not they are unidimensional. If for example there are **N** initial items requiring only agreement or disagreement, then there are 2^N possible response patterns. If the items are unidimensional then only $\mathbf{N} + 1$ of these patterns will be obtained. The fact that the probability of deviant patterns may be thus exactly computed allows for computation of a coefficient of reproducability, **R**, as follows:

$$R = \frac{\text{total number of errors}}{\text{total number of responses}} \tag{3}$$

where an error is any deviation from the idealized unidimensional pattern. The total number of errors may be counted in different ways (Gorden, 1977). Thus computed, the coefficient of reproducability may be interpreted as the proportion of responses to items that may be correctly reproduced from knowledge of an individual respondent's score. If the value of R is greater than .9 for a given scale then it is normally considered to be unidimensional.

b. Multidimensional Scaling Multidimensional scaling is probably most often considered to be a technique for geometrically representing the relationships within data. The idea may be nicely illustrated by a geographical example (MacKay and Zinnes, 1981). An investigator might define his stimuli as some of the major cities in the continental United States. The relationships he wishes to measure are the distances between the cities. These are dissimilarities data. Consider the eight cities of Seattle, San-Francisco, Los Angeles, Dallas, Atlanta, Miami, Washington D.C., and New York. If the distances between all of the twenty-eight possible pairs are estimated correctly and then the distances are superimposed on a map such that any two of the cities are located correctly, then the estimated locations of all of the cities must of necessity exactly match their locations on the map. This is a two-dimensional representation of the dissimilarity relationships between the eight cities. The same basic idea of establishing the geometric relationships between stimuli is at the root of all multidimensional scaling applications. And this regardless of whether the stimuli are tangible entities such as cities or less tangible ones such as those of more direct interest in public administration.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A generic sequence of steps in multidimensional scaling starts with the selection of the stimuli. One or more of the four types of scaling data noted above are gathered for these stimuli. The various techniques of multidimensional scaling make different assumptions about the measurement properties of the data. The techniques called ''fully nonmetric'' assume ordinal input data and yield ordinal output. The techniques commonly known as ''nonmetric'' assume ordinal input data and yield metric output. ''Metric'' methods assume that the input as well as the output data have at least interval level properties. Regardless of which technique one deals with, the relationships in the data are assumed to be distances in the attribute space. A desired initial number of dimensions is assumed and, using a process developed by Walter Torgerson, the distances and hence the stimuli are configured in the space.

In Torgerson's model, the data are assumed to equal distances in a Euclidean multidimensional space (Torgerson, 1958). Let $\mathbf{D}_{ij}$ be the dissimilarity data between stimuli i and j. Let $\mathbf{x}_{ij}$ and $\mathbf{x}_{jk}$(i = 1, . . . , I; j = 1, . . . , J; I = J; k = 1, ... K) be the coordinates of stimuli i and j along dimension **k**. Torgerson's fundamental assumption is:

$$\mathbf{D}_{ij} = \mathbf{d}_{ij} = \{\sum(\mathbf{x}_{ik} - \mathbf{x}_{jk})^2\}^{1/2} \tag{4}$$

Torgerson showed how one can start with this assumption and derive a matrix of coordinates in the attribute space for the data points. Measures of the goodness of the fit between the data and the interpoint distances in the configuration ($\mathbf{d}_{ij}$) is used to indicate whether the initial number of dimensions posited for the configuration is adequate to represent the data. New configurations are estimated, evaluated, and adjusted until a satisfactory goodness of fit is achieved (Davison, 1983).

Identifying the dimensions in the configuration is often a difficult task. Multidimensional scaling procedures have no built-in mechanisms for labeling the dimensions. The investigator, having developed the configuration under the selected dimensionality, can follow one of several procedures. He or she may (1) directly ask the respondents to subjectively interpret the dimensions once a satisfactory configuration has been achieved; (2) identify the dimensions in terms of objective characteristics of the stimuli; or (3) ask the respondent to identify the dimensions that were the most significant in terms of giving their judgments and infer from their responses to the configuration.

Multidimensional scaling offers considerable promise in public administration outside its role in graphically representing data. For example, a highly innovative use is in supporting complex group decisions, primarily in decision situations characterized by multiple conflicting objectives and high levels of uncertainty (Easley and MacKay, 1995). The key technique here is a recently developed multidimensional scaling technique known as PROSCAL. PROSCAL combines traditional multidimensional scaling procedures with advanced statistical and psychological models. In doing so it allows the investigator to perform formal hypothesis tests for dimensionality, estimate the most likely levels of agreement among respondents in terms of the appropriate priorities for particular stimuli, and estimate dimensional weights. Many other innovative uses of scaling are feasible.

IV. REVIEW OF THE MAIN POINTS

Before bringing this chapter to its conclusion, it is appropriate to briefly review the main points of the discussion.

We concur with Young in not accepting the commonly held position that measurements are characteristics of empirical entities in vacuo (Young, 1987:64). Rather we assume that they depend in the first instance upon the interaction between the empirical entities and the psycholog-

ical processes through which they are observed. We presuppose that data are obtained on the basis of a clearly articulated operational definition in an empirical situation having sufficiently well-known characteristics. We further assume that measurements are a result of a classification process in which, pursuant to the operational definition, two equivalent empirical entities are assigned to the same observation category, whereas two nonequivalent empirical entities are assigned to different categories. Based upon these assumptions, our view of measurement requires that it is always possible, for any two empirical entities, to psychologically determine at least whether they are empirically equivalent with respect to the categories stipulated by the operational definition. Beyond this point, higher measurement levels require finer gradation in the classification scheme as well as more demanding psychological processes of observation.[3] And once a level of measurement is selected, preservation of the integrity of the relationships that may exist among observations within the data set requires that the transformations implied by the appropriate restrictions apply. We hold that this view of measurement may be generalized to the physical and biological sciences as well.

Introductory discussions of classification, typologies, indexes, content analysis, meta-analysis, and scaling are included in the chapter. Classification is, at root, an expression of the logic through which recognition of similarity and difference occurs. A typology is, strictly speaking, a formalized classification scheme from which the appropriate subclass for an empirical entity may be deduced. An index is a combination of a set of empirical variables, used to represent them all simultaneously in a summary fashion. Content analysis, meta-analysis, and scaling are all systems for assigning numerals to abstract empirical entities. Knowledge of the properties of each such system enables one to organize observations and identify critical parameters of the entities one is investigating. Content analysis is a system for making inferences about the empirical content of recorded text. Meta-analysis is a statistical system for numerically estimating parameters that span across individual research projects on a specific topic. Scaling is a system with which to empirically test and (possibly) validate the existence and magnitude of the characteristics associated with a concept. Numerous useful references are provided throughout the chapter for the researcher who wants to use any of these techniques.

Finally, we want to recognize that reliance on numbers is no substitute for reflective thought. At the same time however, reflective thought is no substitute for a basic understanding of measurement and empirical research. If nothing else, such an understanding helps to avoid misleading inferences by recognizing the difference between quantification and measurement. Not everything may be measured. Measurements only reflect those particular descriptive features of things that may be reflected in quantitative terms. That is, to measure something is to assign a numeral to some quantitative parameter that describes a feature of a set of empirical entities. It is by no means the case that every quantity one can specify is a measure of some such descriptive feature. When numerals fail to capture such descriptive features they simply do not *measure* anything. Badly misleading inferences may result. While in the everyday life of most public administrators, genuine understanding of many highly significant and interesting matters may be obtained without the use of any sort of measurement or analytic technique, an understanding of the logic of measurement and how it is applied to improve our inferences may be of considerable value in improving our decisions.

NOTES

1. Fundamental theory in this sense may be contrasted with phenomenological theory. The postulates of phenomenological theory are, at best, determined by the perceptions of communities of scholars who study the relevant segment of the world. It aims at

organizing a mass of data from such segment around a concept. On the other hand, the postulates of fundamental theory are rooted in mathematics and quantitative reasoning. Its aim is not to confront the raw data so much as it is to explain the relatively few parameters of the phenomenological theory in terms of which the data are obtained.

2. The amount of information is measured by a ''bit.'' A bit of information is shorthand for a ''binary digit.'' One bit of information is the amount of information required to control, without error, which of two equiprobable alternatives is to be chosen by the receiver of the information.
3. While the reasoning involved goes beyond the scope of this chapter, we are convinced by Young (1987: 64) that our presentation of four discrete, unique levels of measurement (nominal, ordinal, interval, and ratio) is oversimplified. A more accurate view in our judgment is that there is a measurement continuum rather than four unique levels, and that the four levels we identify in this chapter are roughly-identifiable points on the continuum.

REFERENCES

American Psychological Association (1994). *Publication Manual of the American Psychological Association*, 4th ed. Washington, D.C.: American Psychological Association.

Auster, E. R. (1989). ''Task Characteristics as a Bridge Between Macro- and Microlevel Research on Salary Inequality Between Men and Women,'' *Academy of Management Review, 14* (2): 173–193.

Bingham R. D. and W. M. Bowen (1994). '''Mainstream' Public Administration Over Time: A Topical Content Analysis of the Public Administration Review,'' *Public Administration Review, 54* (2): 204–208.

Blalock, H. M., Jr. (1982). *Conceptualization and Measurement in the Social Sciences*, Beverly Hills, CA, Sage.

Bogardus, E. S. (1929). *Immigration and Race Attitudes*, Boston, MA, Heath.

Borman, W. C. (1991). ''Job Behavior, Performance, and Effectiveness'' in Dunnette, M. D. and L. M. Hough, *Handbook of Industrial and Organizational Psychology*, CA: Consulting Psychologists Press.

Bowen, C. C. (1996). ''Manager Wanted, Male Only; Secretary Wanted, Female Only: A Content Analysis of Classified ads in Taiwan,'' Unpublished paper (available from the author).

Bowen, W. M. (1995). ''A Thurstonian Comparison of the Analytic Hierarchy Process and Probabilistic Multidimensional Scaling Through Application to the Nuclear Waste Site Selection Decision,'' *Socio-Economic and Planning Sciences, 29* (2): 151–163.

Bowen, W. M. (1990). ''Subjective Judgments and Data Envelopment Analysis in Site Selection,'' *Computers, Environment, and Urban Systems, 14*: 133–144.

Bowen, W. M., C. C. Chang, and Y. K. Huang (1996). ''Psychology and Global Environmental Priorities in Taiwan: A Psychometric Test of Two Learning Models,'' *Journal of Environmental Psychology, 16*: 259–268.

Bowen, W. M. and K. E. Haynes (1994). ''Environmental Priorities and Individual Differences: Metropolitan Cleveland,'' *The Environmental Professional, 16*: 303–313.

Campbell, D. T. and D. W. Fiske (1959). ''Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix,'' *Psychological Bulletin, 56*: 81–105.

Campbell, D. T. and J. C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research*, Boston: Houghton Mifflin Company.

Chang, C. C. (1993). *Gender and Performance Appraisals in Work Settings: A Meta-Analysis*, Unpublished doctor's dissertation, The Pennsylvania University, University Park.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Cohen, J. (1960). ''A Coefficient of Agreement for Nominal Scales,'' *Educational and Psychological Measurement, 20* (1): 37–46.
Coombs, C. H. (1964). *A Theory of Data*, New York, NY, John Wiley and Sons.
Davison, M. L. (1983). *Multidimensional Scaling*, New York, John Wiley and Sons.
Easley, R. F. and D. B. MacKay (1995). ''Supporting Complex Group Decisions: A Probabilistic Multi-Dimensional Scaling Approach,'' *Mathematics and Computer Modelling, 21* (12): 25–33.
Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution*, New York, NY, Basic Books.
Gaugler, B. B., D. B. Rosenthal, G. C. Thornton, III, and C. Bentson (1987). Meta-Analysis of Assessment Center Validity, *Journal of Applied Psychology Monograph, 72* (3): 394–511.
Glascock, J. and R. LaRose (1992). ''A Content Analysis of 900 Numbers: Implications for Industry Regulation and Self Regulation,'' *Telecommunications Policy, 16*: 147–155.
Gorden, R. L. (1977). *Unidimensional Scaling of Social Variables* New York, NY, Free.
Green, P. E. (1989). *Multidimensional Scaling: Concepts and Applications*, Boston, MA, Allyn and Bacon.
Guttman, L. (1944). ''A Basis for Scaling Qualitative Data,'' *American Sociological Review 9*: 139–150.
Hedges, L. V. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*, New York, NY, Academic.
Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*, Reading, MA, Addison-Wesley.
Hunter, J. E. and F. L. Schmidt (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Beverly Hill, CA, Sage.
Janis, I. L. (1949). ''The Problem of Validating Content Analysis, in *The Language of Politics, Studies in Quantitative Semantics*, H. D. Lasswell, N. Leites, R. Fadner, J. M. Goldsen, A. Gray, I. L. Janis, A. Kaplan, D. Kaplan, A. Mintz, I. De Sola Pool, and S. Yakobson, (eds.), George Stewart, 49, 55–82.
Johnson, B. T. (1993) *DSTAT: Software for the Meta-Analytic Review of Research Literatures*, NJ, Lawrence Erlbaum Associates.
Kaufman, H. (1991). *Time, Chance, and Organizations*, 2nd Ed., Chatham, NJ, Chatham House.
Kendall, M. G. and A. Stuart (1966). *The Advanced Theory of Statistics Volume 3*, New York, Hafner.
Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*, Beverly Hills, CA, Sage.
Kyburg H. E., Jr. (1984). *Theory and Measurement*, Cambridge, Cambridge University Press.
Likert, R. (1932). ''A Technique for the Measurement of Attitudes,'' *Archives of Psychology, 140*: 1–55.
Machlup, F. (1994). ''Are the Social Sciences Really Inferior?,'' in *Readings in the Philosophy of Social Science*, M. Martin and L. C. McIntyre (eds.), Cambridge: The MIT Press, 5–20.
MacKay, D. B. (1988). ''Thurstone's Theory of Comparative Judgment,'' *Kotz-Johnson Encyclopedia of Statistical Sciences, Volume 9*: 237–241.
MacKay, D. B., W. M. Bowen, and J. L. Zinnes (1996). ''A Thursonian View of the Analytic Hierarchy Process,'' *European Journal of Operational Research, 89*: 427–444.
MacKay, D. B. and J. L. Zinnes (1981). ''Probabilistic Scaling of Spatial Distance Judgments,'' *Geographical Analysis, 13*: 21–37.
Mansfield, E. (1982). *Microeconomics: Theory and Applications*, Fourth Edition, New York, W. W. Norton and Company.
McIver, J. P. and E. G. Carmines (1981). *Unidimensional Scaling*, Newbury Park, California, Sage.
Mukherjee, R. (1983). *Classification in Social Research*, Albany, New York, State University of New York Press.
Nagel, E. (1931). ''On the Logic of Measurement,'' Ph.D. dissertation, Columbia University.
Namenwirth, J. Z. (1973). ''The Wheels of Time and the Interdependence of Value Changes,'' *Journal of Interdisciplinary History 3*: 649–683.
Paisley, W. J. (1969). ''Studying 'Style' as Deviation from Encoding Norms,'' in *The Analysis of Communication Content: Developments in Scientific Theories and Computer Techniques*, G. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisley, and P. J. Stone (eds.), New York: John Wiley and Sons.
Peirce, C. S. (1877). ''The Fixation of Belief,'' *Popular Science Monthly*.
Rescher, N. (1970). *Scientific Explanation*, New York, Free.
Rescher, N. (1979). *Cognitive Systematization: A Systems-Theoretic Approach to a Coherentist Theory of Knowledge*, Totowa, New Jersey, Rowman and Littlefield.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Roberts, F. S. (1979). *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Reading, Massachusetts, Addison-Wesley.
Rosenthal, R. and D. B. Rubin (1986). ''Meta-Analysis Procedures for Combining Studies with Multiple Effect Sizes,'' *Psychological Bulletin*, *99*(3): 400–406.
Saaty, T. L. (1988). *The Analytic Hierarchy Process*, Pittsburgh, PA, University of Pittsburgh Press.
Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*, Princeton, Princeton University Press.
Shaw, M. E. and J. M. Wright (1967). *Scales for Measurements of Attitudes*, New York, NY, McGraw-Hill.
Simon, H. A. (1976). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, Third Edition, New York, NY, Free.
Simon, J. L. and P. Burstein (1985). *Basic Research Methods in Social Science*, Third Edition, New York, NY, McGraw-Hill.
Stevens, S. S. (1957). ''On the Psychophysical Law,'' *Psychological Review 64*: 153–181.
Thurstone, L. L. (1929). *The Measurement of Social Attitudes*, Chicago, University of Chicago Press.
Thurstone, L. L. (1927). ''A Law of Comparative Judgment,'' *Psychological Review*, *34*: 273–286.
Torgerson, W. S. (1958). *Theory and Methods of Scaling*, New York, NY, Wiley.
Weber, R. P. (1985). *Basic Content Analysis*, Beverly Hills, CA, Sage.
Young, F. W. (1987). *Multidimensional Scaling: History, Theory, and Applications*, Hillsdale, New Jersey, Lawrence Erlbaum Associates.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

6
Questionnaire Construction

Donijo Robbins*
Rutgers University, Newark, New Jersey

I. INTRODUCTION

A questionnaire is just one way to collect data about the researcher's objectives and purposes. Questionnaires are used in all types of research such as academic research, public policy, and public relations research. Questionnaires, if constructed carefully with reliable and valid questions, will result in a predictable relationship between the respondents' answers and what the researcher is trying to measure. Moreover, a good questionnaire is one that works and that maximizes this predictable relationship. To achieve a good questionnaire, the questions must be valid and reliable, clear and concise, easily comprehendible by the respondents, coded and entered into machine readable form and analyzed without bias or errors. Questions are reliable when two or more respondents interpret and understand the question the same way. And questions are valid when the respondents' answers are true to what the researcher is attempting to measure.

Unfortunately, there is no set format to construct questionnaires, it just requires knowledge of the field and common sense. This chapter offers guidelines and suggestions about the construction and design of questionnaires. It also suggests ways to improve the validity and the reliability of the overall design.

There are six basic steps involved in the construction process (see Table 1). The first step is the development of the research topic and a statement of purpose. Once the purpose of the research has been stated, the researcher must decide what variables are to be studied and develop questions relevant to the variables and the purpose of the project. These questions must then be constructed and logically ordered in the questionnaire in order for the researcher to get valid and reliable results. Next, the questions are pretested in order to detect any errors. After the pretest, the necessary corrections should be made to the questionnaire and the questionnaire should be tested again. After the second test, the questionnaire is administered to the target population. Once all of the surveys have been administered the data must be coded and entered into machine readable form. Finally, the researcher analyzes and interprets the results and reports the findings.

Although it may seem simple because there are only six steps involved in the process, each step is difficult and complex. And each step must be taken as seriously as the others; each deserving equal weight. If the researcher neglects any one part of the construction process, the questionnaire will fail, not to mention the entire research project. The remainder of this chapter outlines general guidelines and suggestions to each step of the questionnaire process.

Current affiliation: University of Maine, Orono, Maine.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

TABLE 1 Steps to Questionnaire Development

1. Statement of purpose.
2. Define relevant variables.
3. Develop questions.
4. Construct questionnaire.
5. Pretest questionnaire.
6. Administer, code, and report.

II. THE RESEARCH IDEA AND THE VARIABLES TO BE STUDIED

What should the research idea be? What type of individuals should be studied? What phenomena or events should be analyzed? With any research project, whether or not it involves the use of a questionnaire as a way to collect data, the researcher must first find an idea to research. The research idea must be explored and the researcher must submerge himself in the literature in order to gain an ''expert'' understanding of the subject material. The researcher should then narrow down the topic and define the problem or purpose which he wishes to study. At this point, the researcher should write a paragraph or so stating the purpose of the research. This allows the researcher to pinpoint the area of interest. The researcher should attempt to visualize what the results should look like. Visualization will help the researcher develop the appropriate variables that could be used to measure the objectives of the research project.

The researcher should test the variables believed to give the results that were visualized. The data that is selected must reflect the researcher's objectives and purposes behind the study. In other words, the researcher should make a list of the variables that are necessary to measure the relationship that is being posed. The researcher need not make a list of questions at this time, just focus on the variables to be studied. It should be decided what variables will be the independent variable(s), the dependent variable(s), and the control variable(s). The researcher must also decide who the target population is and the appropriate sample size.

At the end of the first two steps, the researcher should have a plan of action stating the purpose of the research, a list of the relevant variables to be measured, and what lies ahead in the project. The researcher should also include things such as development of the questions and the questionnaire, and the types of questions that might be appropriate. Generally, asking one question per variable will suffice, but if the variable or idea is complex, then it is always better to ask multiple questions about the same idea. Asking multiple questions will increase the validity and reliability of the study. This process paves the way for the next step of question development.

III. QUESTION DEVELOPMENT

The developmental stage is by far the most important stage of the whole construction process. The right questions must be asked. And the questions that are asked must be universally understood by all respondents. The answers that are produced will be valuable if and only if the researcher can show a predictable relationship with the researcher's purposes and objectives of the study. Pre-existing questionnaires can be used as a reference to guide researchers composing questions and constructing the questionnaire itself. This section is devoted to the preliminary development of questions and ways to maximize the reliability and validity of the questions that are asked.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A. Designing Questions To Maximize Reliability

Questions are reliable if they are interpreted the same way by all those participating in the study. In other words, the questions mean the same thing to all respondents. To begin, researchers should conduct focused group discussions in order to get a general idea about the backgrounds and the cultural differences of the target population that is going to be studied. The researcher must also decide what type of survey to administer, interview surveys or self-administered surveys, and what question format should be used, open or closed. The questions that are developed should be relevant to the research. The wording should be simple, unambiguous, and universally understood.

1. Focus Groups

The best way to begin the development of questions is to conduct focused discussions with individuals from the target population. The discussions should be focused around the purpose and objectives of the research. Focus groups, essentially, are a reality check for the researcher. The group discussion allows the researcher to compare the actual responses relayed by the participants with the complex ideas the researcher is attempting to measure. Since the researcher only needs to get a general understanding of the respondents' perceptions and interpretations, therefore, these groups are not much larger than six to eight people.

The feedback and results the researcher receives from the focus group will assist the researcher with future decisions. For example, these discussions will help decide what type of survey method to use, interviews or self administered. It will also help decide what type of data to collect. Generally, the data that are collected with surveys is nominal, for example, the gender of the participant. Other types of data that are collected with surveys is ratio data such as annual income, tuition cost per semester, or hourly wage rate and ordinal or categorical data are used to categorize responses such as rating the job of the president as good, fair, or poor. Categorical data is used when it becomes too difficult to measure the actual result.

2. Types of Surveys

There are two different types of surveys, interviews (face to face interviews or phone interviews) and self administered surveys (normally sent through the mail). Deciding what type of survey to use is a difficult task. This section discusses the pros and cons of each type of survey.

The interview process creates the assumption that the respondents will, on average, participate more in the survey because someone is present. Whereas, self administered surveys lack respondent participation. Not to mention, respondents may get bored with the process and skip around within the questionnaire. When respondents lose interest and skip around, distortion is created and the responses become unreliable. Unfortunately, the researcher would be unaware of this distortion and report distorted results. In this sense, self administered surveys lack the control that is more apparent with interview surveys.

Interviews ensure high completion rates whereas self-administered surveys have the lowest response rates. The more work respondents are required to do, the lower the response rate. The more interest bestowed in the respondent, the higher the response rate. Generally, response rates for mail surveys should range between 60 to 70 percent.

Another advantage of interviews is the rapport that can be established between the interviewer and the respondent. This cannot be achieved with phone interviews or self-administered surveys. This personal touch will ease any tension that the respondent may have prior to the interview process. The presence of an interviewer also allows for more flexibility. If the respondent does not completely understand a question, the interviewer can clarify any ambiguities.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The interviewer can repeat the question if necessary, whereas the question may be skipped and left unanswered if the questionnaire is self-administered. The worst case scenario is if the respondent simply guesses at the meaning of the question and answers it incorrectly.

Interviews do not require respondents to have a certain level of education or a specific literacy rate. Instead, interviews depend on the expertise of the interviewer and the interviewer's level of education and training. However, questions in self-administered surveys may go unanswered or answered incorrectly because the respondent had a difficult time reading and interpreting the questions.

If the researcher chooses to use interviews, it must be understood that interviewers are the most important part of the interview process. The interviewers will make or break the research project. Interviewers must be well trained and have complete knowledge about what the research project entails. The researcher should go over every question with the interviewers to clear up any misunderstanding or ambiguities about the questionnaire. Interviewers must also be briefed on who they will be interviewing and the participants' backgrounds.

In order for the interviewer to establish a trusting and understanding rapport with the respondent, the interviewer should be aware of the background and cultural differences of the target population. For example, if blue collar workers are the target population, the interviewer should not dress in a three-piece business suit. The interviewer should never come across as someone who is better than the respondent. If this attitude is portrayed, the respondent's answers may be distorted because they may feel uncomfortable and unacceptable.

After the interview process has been completed, the researcher needs to verify the data that was collected. The researcher should call respondents to verify that they actually participated in the study and thank them for their time. If the researcher questions the validity of the data, the interviewer must be confronted about the issues and dismissed at once. This verification process, although lengthy, is very important and very necessary. Verification only helps the researcher validate the data.

Unfortunately, interviews are very expensive to conduct. And personal questions may be less reliable with interviewers because respondents may be embarrassed to answer the questions honestly. Although the surveys are always confidential, personal questions may still be embarrassing for some individuals. Whereas, personal questions may be answered more truthfully with self-administered questionnaires.

3. Question Formats

There are two general types of question formats. The first format is closed questions which provide respondents with a uniform frame of reference. For example, a Likert Scale is used with closed questions. A Likert Scale is a scale ranking of the respondents preferences or opinions. The other question format is open questions. These questions allow the respondent to answer freely, without being constrained to a supplied frame of reference. This section discusses the advantages and disadvantages of both types of formats.

Open questions are useful because they allow unanticipated answers to be obtained. Respondents are free from any constraints and the answers given represent how respondents interpreted the question. Open questions allow for specific and precise answers. Therefore, if the researcher interpreted the question one way, which differs from the respondent's interpretation, and left the question unconstrained, the respondent's answers would be more precise. Open questions suggest the respondent's level of knowledge about a given topic or idea.

Open questions are useful when the researcher wants to give the respondent a sense of involvement. Respondents like to be involved with the survey process and allowing them to freely answer a few questions gives them a sense of involvement.

However, open questions take much longer to answer than closed questions. Open questions are often difficult to code which makes it difficult to statistically analyze and draw conclusions. There is also less order and lower reliability associated with open questions.

Closed questions are those with a list of given responses to choose from. These questions require less skill and effort of respondents and take less time to answer. Closed questions are easier to answer and easier to code and analyze. The questions will be interpreted the same way because a constant frame of reference is supplied to all respondents. Therefore, the questions are uniform, more reliable and easier to interpret.

However, the frame of reference is difficult to compose. It is difficult for the researcher to develop an exhaustive list of responses. But when it becomes too difficult to develop an exhaustive list or the list is too long, open questions should be used. These lists of answers put words in respondents' mouths and keep the respondents from answering freely. Also, closed questions do not guarantee universal understanding. Although all respondents are exposed to the same frame of reference does not imply that the questions are interpreted the same way by all respondents.

What question format should be used? Generally, the best way to approach this dilemma, is to develop open questions in the early stages for use in the focus groups and pretests. Once the researcher has an idea about the interpretations and the type of responses the questions generate, the wording should be improved and the question should be changed to a closed question with an exhaustive list of responses.

4. Question Wording

There should be one idea per question and the questions must be reliable and valid. Questions must be relevant to the purpose of the study. The language must be simple and unambiguous. Researchers should avoid questions that are double-barrelled, loaded, negative, or biased. Questions are reliable when two or more respondents understand and interpret the question the same way. In other words, the question is universally understood. Questions are valid when respondents' answers are a true measure of what the researcher is trying to measure. This section discusses the do's and don'ts of question wording.

Differences in answers must be attributable to the differences among respondents' personalities, not different interpretations. To achieve universal understanding among respondents, questions should ''rub'' respondents the right way. Questions should be *relevant* to the purpose and objectives of the research project. Questions should be *unambiguous* and straightforward. And questions should be *brief* and to the point. Remember, the more work a respondent has to do, especially with self administered questionnaires, the lower the response rate.

There is no set way to word questions perfectly. For example, consider the following questions. The first is a question asked by the Gallup poll and second by the Harris poll. 1) ''Do you support or disapprove the way President Clinton in handling his job?'' 2) How would you rate the job Clinton is doing as president—excellent, pretty good, only fair, poor?'' The Harris poll then combines the ''excellent'' and ''pretty good'' responses as positive support and combines ''only fair'' and ''poor'' as negative support. Both polls are reliable, but the wording varies in such a way to generate different results.

The Harris poll seems to reflect more reliable results simply because respondents are not constrained to polar extremes, the respondents have a broader spectrum to chose. If a moderate conservative was asked about the president's job, they could support some of the president's actions and positions, but disagree with other things. However, they would probably respond unfavorably to the question asked by the Gallup poll. This same individual may reply to the Harris poll by responding ''pretty good.'' Therefore, the president would have a favorable ranking. Researchers can word questions in such a way to get the outcome they desire.

The researcher must offer enough categories to rank responses, but the researcher should not offer too many. If too many categories are offered, it becomes difficult for respondents to distinguish between the categories. For example, respondents may be asked to rank the foreign policy practices of the current president as excellent, very good, good, fair, or poor. Although two people could feel the exact same way about the foreign policy practices, one respondent may reply ''good'' while the other respondent may respond ''fair.'' Fair and good could have been interpreted as average by both respondents, but each responded differently. Although the question was universally understood, the categories are too close to distinguish.

Having only two choices is a disadvantage to any research project. It limits respondents to only two choices and forces respondents to either agree or disagree; approve or disapprove; excellent or poor. The researcher is only asking about the polar extremes of the continuum and forces respondents who are somewhere in the middle to make a choice. It is better to have more than two choices on the continuum, but remember not to have too many.

It is also important to include a ''don't know'' or ''no opinion'' choice when respondents are asked to rank a response. Sometimes the respondent may lack the knowledge of the topic being studied. Instead of forcing respondents to make choices or decisions they do not understand or have any knowledge of, the ''no opinion'' or ''don't know'' is the best choice. These responses allow researchers to analyze the unavailable knowledge base of the target population. Unfortunately, respondents that do have an opinion may not want to answer specific questions, therefore, these respondents may choose the ''no opinion'' or ''don't know'' option as a way to avoid the question.

a. Questions Should be Relevant Questions that are asked, should be relevant to the researcher's purpose and objectives of the study. Questions should not be asked just to ask and later determine whether or not to use them. This wastes time and effort for the researchers, the respondents, and the interviewers, if interviewers are used. The questionnaire process is so difficult in and of itself, the researcher should not waste time and energy developing more questions than are needed.

Once the researcher has determined what is to be measured, the researcher must decide what questions should be used to measure the variables. At this point, the researcher should refer to previously conducted surveys. For example, the National Opinion Research Center at the University of Chicago conducts the General Social Survey. The researcher can use these surveys as references for question wording and questionnaire construction. Questions that are relevant in other surveys can be used as long as these questions are used in the correct context. However, just because these questions have been used before does not guarantee that the questions are reliable and valid.

b. Questions Should be Universally Understood All the questions asked in the survey must be universally understood. In other words, the questions must mean the same thing to all respondents. If the questions are universally understood, the questions are considered to be reliable. Researchers must remember that abstract thinking is the norm in most research fields, and exposing the average lay person to such abstract wording may make the questions too complex and too difficult to interpret. Therefore, researchers should avoid abstract and complex wording, especially technical jargon that only certain professionally trained individuals have been exposed.

Keep it simple. But do not make the questions so simple that the questionnaire is viewed as talking down to respondents. If respondents perceive this type of behavior, the respondent may be offended and lose interest.

The questions should be relevant to the target population. The researcher must always keep in mind the target population the project is surveying. The researcher must always consider the background and cultural differences of the target population. For example, if the target

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

population is represented by non-high school graduates, the researcher should not word questions that target college graduates. Also, do not ask questions such as ''How old is your husband?'' when the respondent is single or has a wife. Instead, first ask a question that categorizes respondents' marital status, then ask ''If married, how old was your spouse on his/her last birthday?''

The questions should not only be simple, but also clear, specific, and unambiguous. If terms or concepts are ambiguous, then the researcher should define the concept prior to asking the question. For example, if respondents are asked about their last visit to the doctor, responses will vary considerably. What constitutes a doctor to one person may not be classified as a doctor by someone else. A doctor could be a licensed medical doctor (M.D.), an osteopath, a chiropractor, or even a witch doctor. The researcher must therefore define what the term ''doctor'' constitutes. It must be defined in such a way that all respondents understand the term universally. Because different opinions exist about terms and concepts, the researcher should ask multiple questions to clarify the analysis process.

The researcher should avoid asking for information that respondents are likely to have forgotten. The researcher should not ask a question that requires respondents to recollect the past, for example, to think back five years ago. Respondents will more than likely guess or approximate the answers. For example, researchers should not ask respondents what their annual income was six years ago, hospitalizations over the past ten years, or the when their last flu shot was. If the researcher wants the respondent to recollect the past, as a rule of thumb the time frame should be nothing more than six months ago. Always remember to have a narrow time frame.

The researcher should also keep in mind that it is difficult for respondents to answer questions about their opinion. It is easier to answer questions about personal experiences, fact, and/or behavior. If opinion questions are asked, respondents have to think about how they really feel about that particular issue, whereas questions that concern fact will require less thinking, less effort, and less time. With factual questions, the answer is either one way or the other.

Not only should the questions be understood universally by all respondents, the answers given by the respondents should be standardized. For example, if respondents are asked ''When did you have the chicken pox?'' They may respond a variety of ways: ''last year,'' ''when I was in high school,'' ''When I was 10 years old.'' If the researcher wanted respondents' age when they had the chicken pox, the researcher should have asked for their age specifically. Instead, the researcher should have asked ''How old were you when you had the chicken pox?'' Therefore, all respondents will answer with their age at the time of infection and all responses will be standardized.

The more general the question, the wider the range of interpretations and responses. In order to get uniform interpretations and standardized responses, the researcher should make the question as specific as possible. The researcher should not assume the respondent will interpret the question the same as the researcher or even the same way as other respondents. The researcher should not assume anything about respondents when developing questions.

c. Avoid Double-Barreled Questions Researchers should avoid the use of double-barreled questions. Double-barreled questions are those questions that ask two or more questions at the same time. For example, ''when the cost of college tuition increases are you more likely to drop out of school and look for a job?'' In this example, some may drop out of school but not look for a job, while others stay in school and look for a job. Generally, when the word ''and'' is included, the question is probably a double-barreled question and should be avoided. If the word ''and'' appears, the researcher should reword the question such that only one item is asked per question.

d. Avoid Loaded, Negative, and Bias Questions The researcher should avoid questions that are loaded. Loaded questions are those that persuade the respondent to answer a certain way through implication or suggestion. Loaded questions generally include words such as ''forbid,''

''prohibit,'' and ''allow.'' For example, the following is a loaded question: ''Should the United States prohibit it's citizens from carrying handguns?'' The researcher, unknowingly, is suggesting that the United States' government should not allow people to carry guns. These types of questions result in distorted responses and should be avoided.

Negative questions generally include the word ''not'' and should be avoided. Most of the time respondents will overlook the word ''not'' and read the question the opposite way. For example, the following is a negative question and should be avoided: ''United States should not eliminate nuclear testing?'' The respondent may overlook ''not'' and read it as ''the United States should eliminate nuclear testing?'' If the respondent believes that nuclear testing should be eliminated, the respondent would agree with the question the way it was read but actually disagree with the actual wording of the question.

Any question that includes loaded terms or negative words, or if the question is too complex and too ambiguous it is considered to be biased. Bias can be controlled by avoiding these terms and concepts as well as using closed questions. Closed questions help control bias as long as the list of options is completely exhaustive.

e. Ask Multiple Questions The researcher should ask multiple questions about the same idea. Sometimes one question per variable or idea will suffice, for example, gender. But often, relying on just one question makes it difficult for the researcher to interpret the results, especially when the variables or ideas are complex. Asking multiple questions will also increase the accuracy of the overall research project. For example, if social class was the variable of interest, the researcher may ask about annual income or hourly wage, occupation, education, and residence.

B. Designing Questions to Maximize Validity

The previous section suggested ways to make questions more reliable and to make questions mean the same thing to all respondents. This section is devoted to validity and ways to improve the validity of questions and the questionnaire. Recall, that questions are valid when the respondents' answers are a true measure of what the researcher is trying to measure. Hopefully the responses are perfectly accurate. Responses will be accurate if the researcher had access to the information needed to answer questions the same way respondents answered. For example, if the chicken pox question was asked and the respondent answered 10, this response is valid and accurate if the researcher referred the respondent's medical records and found that the respondent was in fact 10 when infected with the chicken pox.

Unfortunately, answers are not always accurate. Inaccurate answers may be given because the respondent does not know the answer; the respondent may know the answer, but cannot recall the answer; the respondent may not fully understand the questions; or the respondent may know the answer and refuse to answer. All of these things result in less accurate responses and make the data less valid.

There are ways the researcher can take specific steps to increase the accuracy of the answers. Some of the things that could be done were discussed in the previous section ''Question Wording to Maximize Reliability.'' Also, the researcher could allow respondents to answer the question the way in which they interpret the question. Allowing respondents to interpret and answer questions accordingly, will help researchers detect faulty wording. For example, if the question is unanswerable by everyone then there is a problem within the design of the question. Therefore, the question should be reworded or dropped from the questionnaire altogether.

IV. QUESTIONNAIRE LAYOUT

The layout of the questionnaire has significant barring on the results of the research project. The results could vary significantly if the position of the question is moved from the beginning

TABLE 2 Characteristics of Questions

The questions should be relevant to the objective of the study.
The questions should be clear and unambiguous; what may seem clear to the researcher may be unclear to the respondent.
Be careful when asking personal questions; do not pry.
Provide definitions to unfamiliar words or words with multiple meanings.
The questions should mean the same thing to all respondents; reliable.
Ask multiple questions with different question form that measure the same idea.
Ask open questions prior to asking closed questions in order to create an exhaustive list of options.

of the questionnaire to the middle or even the end of the questionnaire. The sequence of questions and the overall physical appearance of the questionnaire are also very important. The questions should flow smoothly with a clear and orderly sequence.

Instructions should accompany the questionnaire at the very beginning of the questionnaire explaining who the researcher is, the researcher's affiliation, and the research project itself. For example, the researcher should explain why the research is being done and what will be done with the responses. The researcher should also stress that all responses are strictly confidential. Remember, never assume the respondent is familiar with questionnaires. The researcher must ensure respondents that the survey is strictly confidential and there is absolutely no way to trace the responses back to anyone.

The researcher only needs to explain why the research is being done; the purpose of the research should be explained but nothing else. The researcher should never attempt to explain the relationship that is hypothesized. Attempting to explain this may influence respondents to answer questions a particular way. In one sense, if the researcher states the hypothesized relationships, the instructions could be consider ''loaded.'' For example, if the purpose of the study is to find the effects of increasing college tuition costs on the behaviors of college students, the researcher should state this, the purpose. However, the researcher should not state that what is believed to exist, that college students will drop out of school more rapidly as the rate of tuition increases. State the purpose of the research but not the hypothesized relationships.

The researcher should also include a thank you statement in the instructions. The researcher must always remember to establish a trusting and confident rapport with the respondents. The researcher will achieve better results if respondents are given a sense of involvement and importance. This is not to say that respondents are not important, they are very important. Without respondents, researchers would have no data to analyze.

The researcher should also provide necessary instructions throughout the questionnaire as well. For example, questionnaires often have skip patterns such as ''if you are a dependent, skip to question number . . .'' or ''if you are unemployed, skip to page . . .'' If this type of sequence is used, the researcher should provide clear and precise instructions allowing respondents to move forward smoothly. Skip patterns should be kept to a minimum, and if used, the instructions provided should guide respondents like a road map. There should be no wrong turns or dead ends. Remember, questionnaires are not guessing games for respondents, the more work respondents have to do, the lower the response rate.

Opening questions should be simple, pleasant, interesting, and nonoffensive. The researcher does not want to excite the respondent in such a way that the respondent refuses to answer any more questions. The researcher should try to motivate the respondent and make the respondent feel important.

Sensitive questions should never be placed at the beginning to the questionnaire. There is no perfect place for these questions. Generally, the rule of thumb is to place sensitive questions

TABLE 3 Characteristics of Questionnaires

Questionnaires should be self explanatory.
Questionnaires should start with general, simple, and interesting questions.
Questionnaires should be restricted to closed questions as much as possible.
Questions should be few in number; do not ask more than necessary.
Questionnaires should be typed and laid out in a clear and uncluttered fashion; maximize ''white space.''
Skip patterns should be minimized.
Allow enough space for open question responses.
Set off different sections with lines, bold type face, or spacing.
Arrange the questions logically.
Provide redundant information to all respondents.

toward the middle of the questionnaire, but never in the beginning or at the very end of the questionnaire. Sensitive questions should be placed logically, where the questions are most relevant to the questionnaire and at a point where it is assumed the respondent has become comfortable and confident with the survey.

Boring questions and questions concerning race, gender, and age are normally placed toward the end of the questionnaire. These questions, albeit sensitive, should never be placed at the beginning because they may excite the respondent in such a way causing them to stop participating. If placed at the end, the respondent has had time to become comfortable with the survey and feel less offended by such questions.

The physical appearance must be attractive and pleasing to the eye; convenient to use and easy to follow and read. The printing should be large enough to read, and the researcher should never try to put as many questions on one page as possible. If the survey looks too cluttered it will look too complex and be too difficult to read. The researcher should maximize the ''white space'' to make the questionnaire more attractive and easier to administer. There should also be enough space available for respondents to provide answers to open questions. If skip patterns are used, the research might want to separate these patterns with different type styles, sizes, and shades. The researcher should do everything possible to make the questionnaire as attractive as possible.

V. PRETESTING

After the questions have been developed and constructed logically into a working questionnaire, the next step is to pretest the questionnaire. Like every other step involved in the questionnaire construction process, there is no set way to pretest surveys. Generally, pretests are always conducted. Pretests allow the researcher to weed out any uncertainties and ambiguities that were not apparent prior to the pretest. Pretesting is a way to increase and to reinforce the reliability and the validity of the questions.

The researcher has a number of options or ways to conduct pretests. Generally, two pretests are conducted. The first pretest involves the researcher giving a draft of the questionnaire to colleagues, friends, and relatives to read, to critique and to offer suggestions. Once this has been done, the researcher makes the necessary changes and then pretests the questionnaire again. The second pretest should involve people that mirror the target population. Normally a sample of 25–75 is an acceptable size to pretest. Pretesting a similar population allows the researcher to ensure that the questions are interpreted the same way and mean the same thing to all respon-

dents. Once the second pretest is complete, the researcher should polish the questionnaire by making the necessary changes, cutting items, rearranging questions to fit the questionnaire more logically, clarifying the questions, and making the entire questionnaire flow as smoothly as possible.

VI. PERFORMING THE SURVEY

Although administering the survey, coding the responses and reporting the results are all very important steps in the questionnaire process, they do not receive much attention in this chapter.

Once the questionnaire has been pretested, the survey is ready to be administered. The researcher has already previously decided the type of survey to use (interview or self administered) and the target population. Next, the researcher distributes the survey to the chosen sample of the target population. This could be the most lengthy part of the whole questionnaire process. It takes time to conduct interviews especially if the sample size is large. And it takes more time to administer surveys through the mail. Once the survey is distributed through the mail, it takes time to get enough responses back. Generally, a good return rate for mail surveys is 60–70%. If this percentage is not achieved the first time, the researcher could send a letter to those who have not returned the survey asking them to cooperate and return the questionnaire as soon as possible.

Once all of the interviews are conducted or the self administered surveys are returned, the next step is for the researcher to code the responses. Once again, there is no set way to code responses, especially open questions. This part of the process is solely up to the researcher. Generally, the researcher codes the responses as conveniently and simply as possible; this makes the analysis and interpretation process much less complicated. The researcher's statistical background and knowledge normally guides this process of coding and entering the responses into the desired statistical package or spreadsheet form. After the coding and entering process, the researcher analyzes the data and reports the findings.

VII. CONCLUSION

Questionnaires are not easy to construct. The construction process requires time, common sense, and an understanding of the research and the target population. It is also an advantage if the researcher has artist ability which will contribute to the physical appearance of the questionnaire. In sum, questionnaires must be simple and straightforward. Questionnaires must be universally understood, unbiased, unambiguous, and ethical. They must be valid, reliable, and replicable. And most importantly, questionnaires must accomplish the purpose(s) or objective(s) of the research project.

REFERENCES

Babbie, E. (1990). *Survey Research Methods, Second Edition*, Belmont, California: Wadsworth Publishing Company.

Converse, J.M. and S. Presser (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*, Newbury Park, California: Sage Publications, Inc.

Fallowfield, L. (1995). ''Questionnaire Design,'' *Archives of Disease in Childhood*, *72*: 76–79.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaire Surveys: Theory and Practice in Social Research*, Cambridge, UK: Cambridge University Press.

Fowler, F.J., Jr. (1993). *Survey Research Methods, Second Edition*, Newbury Park, California: Sage Publications, Inc.

Oppenheim, A.N. (1966). *Questionnaire Design and Attitude Measurement*. New York, New York: Basic Books, Inc.

Schuman, H. and S. Presser (1977). ''Question Wording as an Independent Variable in Survey Analysis,'' in *Survey Design and Analysis: Current Issues*, D.F. Alwin (ed.) Beverly Hills, California: Sage Publications, Inc., pp. 27–46.

Stone, D.H. (1993). ''Design a Questionnaire,'' *British Medical Journal*, *307*: 1264–1266.

Warwick, D.P. and C.A. Lininger (1975). *The Sample Survey: Theory and Practice*, Newbury Park, California: McGraw Hill Company.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

7
Sampling and Data Collection

Alana Northrop
California State University at Fullerton, Fullerton, California

One starts with a research topic. Then one develops hypotheses and identifies the variables to be measured. Now it is time to plan the data collection.

First, one needs to decide from whom the data will be collected. Data can come from a wide variety of units of analysis. These units can be people, cities, counties, countries, departments, and corporations.

Second, one needs to decide if one needs to do a sample or a census. A census is information that comes from all the units of analysis in a list. Obviously, if one's list of units is all citizens in a country, that list is very large. Just consider the resources that the US expends every ten years to do a census of its population Census 2000 is expected to cost the government $3.9 million. Given the magnitude of data collection involved in doing many censuses, sampling is a common alternative form of data collection.

Sampling means collecting data from a smaller number than the whole list of units. The need to do a sample instead of a census is driven by the answers to several questions. Does one have the time to collect information from all the units? Does one have the resources to collect information from all the units? And, most importantly, is it necessary to collect information from all the units for what one wants to learn from the data?

When one only collects data from a subset or sample of the complete list, the question arises whether or to what extent does the sample look like the whole universe. The ability to answer this question is the difference between probability samples and nonprobability samples. Probability samples are samples chosen from the universe by random without the researcher having any role in choosing which units are sampled and which are not. Non-probability samples are samples in which the researcher does play a role in choosing which units from the complete list or universe end up in the sample for data collection. The topic of this chapter is sampling and data collection. We will describe the different types of probability and non probability samples, the advantages of each, and the special problems involved in data collection, such as achieving a high response rate.

I. DEFINING THE THEORETICAL POPULATION

Before deciding whether to sample or what kind of sample to do, one must clearly define the theoretical population. To define the theoretical population, one specifies from what units data will be collected in terms of time, territory, and other relevant factors.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A. Unit of Analysis

Data can be collected from individuals, groups, or social artifacts. Individuals are human beings, whether adult citizens or employees in city hall. Groups represent collectivities, such as cities, counties, countries, or departments. If one wants to know how an employee feels about a different work schedule or how a citizen evaluates the delivery of city services, the data are collected from each individual. Thus, the individual is the unit of analysis. If one wants to know the population of a city or the mortality rate of a hospital, the data are collected from each city or hospital. In these cases the unit of analysis is the group and not the individual because only a group can have a population or a mortality rate. To find out whether data collection should be focused on the individual or group, one asks on what variables one wants to collect data. If the variables are characteristics of individual people, then the unit is individuals; and if the variables are characteristics of groups of people, then the unit is groups.

The last kind of unit of analysis is social artifacts. An artifact is any object made by people with a view to subsequent use. Examples of social artifacts are laws, books, buildings, computers, etc. A study of fire risk factors might use buildings as the unit of analysis. Buildings could be evaluated by such characteristics as number of stories, square footage, business use, and type of roofing material.

B. Time

The unit of analysis must be defined in terms of time. Should data be collected as of one point in time or over a period of time? Data that is collected as of one point in time is called cross sectional. For example when the Gallup Poll asks adult Americans to rate the president's performance, it is doing a cross-sectional analysis of public opinion that describes how the public evaluates the president as of a set date. When a news agency compares several of these cross-sectional polls, data are now being compared over more than one point in time and such data are called longitudinal.

Whether to do a cross-sectional or longitudinal study depends on resources and why one is collecting data. The State of California draws cross sectional samples of names on initiative petitions because it only cares if enough legal signatures have been collected as of a certain date. Initiative drives are given 150 days to collect the required number of registered voters' signatures. Enough names are either collected by that date or not. In contrast, a study of the effectiveness of community policing on the crime rate involves looking at the crime rate at more than one point in time, before the introduction of community policing and after.

There are three kinds of longitudinal studies: trend, panel, and cohort. A trend study collects data from different units at more than one point in time. The previously mentioned Gallup poll is an example of a trend study because the same citizens are not interviewed more than once. A panel study collects data from the same units at more than one point in time. If one were doing the community policing evaluation, one would need to do a panel study, collecting data from the same city or cities at more than one point in time. It would only make sense to look at the same city's crime rate before and after the introduction of community policing.

A cohort study falls in between a panel and a trend. In a cohort study different units are studied but the units have something in common. Typically, what the units have in common is age or shared experience in a training program. A study of different police academy classes would be a cohort study. The classes could be compared as to their rates of officer involved shootings or complaints of sexual harassment.

In general, longitudinal data collection produces better quality data than does cross-sectional. Obviously, data that are collected at more than one point in time can indicate whether findings vary over time, which cross sectional cannot. Cross-sectional data are perfectly fine

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

when one needs to know only about one point in time, such as the initiative petitions example or a city surveying households about whether to build a senior citizen center or not. Cross-sectional studies are also quite acceptable when the variables that are being measured are known to be stable, such as the square mileage of a city and population density.

A panel study is better than a trend when the theoretical population is heterogeneous. Studying different units from populations with great variations can give very different results than studying the same units. For example, the poverty rate in the US has stayed fairly stable since the 1960s. Using trend data, we cannot tell whether or not it is the same people who fall below the poverty level. Thus, the data do not allow us to know whether there is a permanent underclass. Using panel data, we could tell that while the poverty level stayed the same, the people who comprised that group changed a lot, so a permanent underclass would be an inaccurate description.

C. Territory

A theoretical population defines the units to be studied in terms of time and also territory. Territory literally refers to governmental boundaries. So if one wanted to study households, one needs to specify households in which city or state. If one wanted to study adult citizens, one needs to specify adult citizens living within distinct territorial boundaries, such as west of the river in the city of Hartford, Connecticut.

D. Other Relevant Factors

Here is the catchall consideration in defining theoretical populations. If one were doing a study for Washington state's highway patrol on drivers who speed, a useful theoretical population would be all licensed drivers in the state as of July 1, 1996. Note we have identified the right unit, which is individual. We have stated a date, so we know we will only collect data from people who lived in the state as of that date. We have also stated a territory, the state of Washington. The other relevant factor specified is that we will only collect data from licensed drivers. If one's unit of analysis is individuals, typically one needs to limit the population by setting a minimum age limit or status, such as licensed driver, or employee. Two year olds are not very helpful survey respondents, even though they can be accident victims. Studies of employees should consider limiting the theoretical population to only full-time employees who have passed their probationary period.

II. WHETHER TO SAMPLE OR NOT

One should now have a well-defined theoretical population. Look at it. Does the theoretical population involve under two hundred employees or does it involve 50,000 households? The rule is if one's population is under 200, one does a census. Essentially, there is no way to do a probability sample on populations under 200 and have any useful error rate. Still, resources may force one to sample when the population is under two hundred but beware of the increase in error.

If one's population is over 200 do not automatically consider a sample. While time and money can be saved by doing a sample, there can be political costs that are too high. For instance, consider studies that want to survey employees about their satisfaction with benefits, work schedules, or training programs. If the list of employees is above 200, those study directors would still be well advised to survey all employees. Probability theory is all fine and good about

drawing conclusions from a sample to the universe. Employees, though, want their individual voices heard on many matters and will not understand why they were not chosen to do so. The same can be said about voters or citizens if we are talking about a local area issue, such as building a new school or fire house in the neighborhood.

There are also theoretical populations above two hundred in size that are rarely sampled because collecting data from all of them is so easy. The case of Congressional districts comes to mind. Data from districts are so readily available that there is negligible time and staff savings gained by using a sample for data collection. The decision comes down to whether the time and staff savings are significantly large enough to outweigh the error and political risk that comes from drawing a sample versus doing a census.

III. PROBABILITY SAMPLING

The theory of probability sampling was first explained by a Swiss mathematician Jacques Bernoulli (1654–1705). He argued that a small randomly chosen sample would look like the entire population. There would be a difference between the characteristics of the sample and the population, but it would be small and calculable. Thus, probability samples are distinguished by the fact that they are chosen randomly from the populations and that how they differ from the populations can be expressed by a calculable error rate.

Many American television viewers have absorbed this argument. Broadcasters frequently report survey results, results based on a random survey of adult Americans. For example, broadcasters report that 62% of Americans support a national health care plan and then go on to say that the margin of error for the survey was $\pm$ three percent. We, the television viewers, interpret the report as saying between 59 and 65% of us support a national health care program. This interpretation is essentially correct. Few viewers could go on to explain the assumptions behind the data, such as respondents to the survey were randomly chosen and that there is another error rate besides the one reported. Still, Bernoulli's description of probability sampling has laid the basis for data collection that is so common in the US that the average citizen cannot escape its effects. From news reports to telephone market surveys to product labeling, Americans are the recipients of data collected from probability samples.

There are four types of probability samples: simple random sample (SRS), systematic sample, stratified sample, and a cluster sample. If one has a list of one's theoretical population to begin with, one can do any of the first three types. If one does not have a list of the theoretical population, then one must consider doing a cluster sample or redefining one's theoretical population so that a list exists. In other words, if one's theoretical population is all households in the city of Fullerton as of October 1, 1996, the city can provide one with such a list because it provides water service to all households. Thus one can do a SRS, stratified, or systematic sample. However, if the city bills landlords for water usage for apartment complexes because each apartment does not have its own meter, then no list of the theoretical population is available from the city. If this is true, consult with the central Post Office in the area to see if they can direct one to a firm that has a list of addresses. If still no luck, then a cluster sample is one's option.

The quality of one's sample rests on the quality of one's list of the theoretical population. The list should be up to date. The list also should describe the population about which one wants to draw conclusions. If apartment renters are left off the list of households, then the conclusions one draws from the sample of households only represents home owners and home renters. This may not be a problem if apartments make up less than five percent of the city's households. The point is one needs to critically evaluate whether a list is available that adequately reflects the theoretical population. The list one uses to draw a sample from is called a sampling frame.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A. Simple Random Sample

Most statistics assume that the data are collected by means of a simple random sampling. Thus, SRS is the ideal type of sample in theory. It may not be the most appropriate one to do in practice. We need to discuss how to do the different samplings before we can expand on this point.

To do a SRS, one must have a sampling frame, which is one's list of the theoretical population. Then, take the following steps:

1. Number every unit on the list. It does not matter whether one starts numbering from one or one thousand. But it is easier if one uses the typical numbering system of 1, 2, 3, etc.
2. Obtain a random number chart. They are in the appendixes of most statistics' books. Some computer software packages also include them. The RAND Corporation also printed a book of them (RAND Corporation, 1955).
3. Decide on how to read the chart. One can start anywhere on the chart. Because it is random, there is no pattern to the appearances of the numbers. One can read rows left to right or right to left. One can read columns down or up. One can also read diagonals, but this way is very hard when one is reading more than one digit.
4. Decide how many digits to read. One reads the number of digits equivalent to the number of digits one used to number one's sampling frame. If one's list was numbered from one to nine, one reads one digit. If one's list was numbered from one to 99, one reads two digits. If one's list was numbered from one to 902, one reads three digits (see Appendix A).
5. Now read the appropriate number of digits on the random number chart. For example, if I was supposed to read three digits and the first numbers I read on the chart were 777, 939, and 961, then the units with those numbers in my sampling frame have made it into the sample. If no one in my sampling frame had one of those numbers, then I ignore the number and keep reading the random number chart (see Appendix A). I read as many numbers from the chart as I need to get the number of units I wanted in my sample. Do not choose extra names to compensate for refusals or failures to respond. Enlarging the sample size for this purpose does not work. Return rate is based on the number of surveys completed as compared to the number attempted.

As one can imagine, if one is reading five digits and needs to get a sample of 1000, reading a random number chart could make one's eyes hurt. The solution is to computerize one's list and use a random number generator. That way the computer chooses one's sample. For instance, the random selection procedure within the widely used SPSS software package can select a SRS.

Entering all the names in one's sampling frame into the computer may not be worth the time trade-off, though. If that is the case, then a systematic sample may be the solution to one's eye strain problem.

B. Systematic Sample

Again, one must begin with a list or sampling frame to do this second type of probability sample. Here is a list of steps that one can follow.

1. Number each unit listed in the sampling frame. This time one must start with the whole number one and continue in normal numbering fashion until one runs out of units to be numbered.
2. Divide the size of the sampling frame by the number of units one wants in one's

sample. For example, if one has 1000 employees and needs 250 in one's sample, divide 1000 by 250. The result is four. This is referred to as the sampling interval. In other words, one out of four units in one's sampling frame will be chosen to be in the sample.

3. Go to a random number chart. One will read as many digits as one's sampling interval. In our example that would be one digit. Start wherever one wants, reading the random number chart. One is looking for the first number between one and one's sampling interval to appear. Ignore numbers on the random chart that do not fall within that range. So in our example we are looking for the first number to appear between one and four. Whatever it is becomes the random start. So if we read a zero and then a three, our random start is three. If we read a six and then a two, our random start is two. Let us assume we got a two.
4. The unit in one's sampling frame with the number two assigned to it is chosen for the sample. Now add the sampling interval to the random start. Four added to two gives a six. Now the unit in the sampling frame with the number six assigned to it is chosen for the sample. Keep adding the sampling interval to the last number and one will select the numbered units in the sampling frame that will be in the sample. In the example, 2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, etc., will be the units chosen from the sampling frame for the sample. When one runs out of numbers in one's sampling frame, one will have exactly the right number of units wanted for the sample. This was accomplished with just using the random number chart once, so no eye strain.

Obviously, a systematic sample is easier to choose than a classic simple random sample. So why ever use a SRS? There is one problem with a systematic sample, but it is not always a problem. If the sampling frame has a cycle to the order of units, then a systematic sample can pick up that cycle and actually increase sampling error compared to a SRS. So one needs to inspect the sampling frame to make sure there is no cycle to the order.

What do we mean by cycle? Let us assume one's list is made up of Boy Scout troops, each with 15 scouts. The first name on the list of each troop is the oldest boy in the troop. If the random start had been a one and the interval a 15, the resulting sample would be made up of the oldest boy in each troop. The first boy to be chosen for the sample would be number one, the oldest boy in the first troop. The second boy to be chosen for the sample would be number 16, random start one plus the interval of fifteen. This means that the second boy to be chosen for the sample would be the oldest boy in the second troop. Continuing with adding the interval of 15, the oldest boy in each troop ends up in the sample. The result is a randomly chosen sample with a marked bias to over representing the characteristics and opinions of older boy scouts. The aim of probability sampling is to reflect the population or sampling frame not to distort it. Thus, if there is a cycle, a repeatable order to how the units' names are listed in the sampling frame, do not use a systematic sampling method. Of course, if the cyclical order of the list, if one does exist, has no relevance to the aims of the study or to any variables being measured, then there is no increase in error rate created by systematic sampling. That assumption, though, may be hard to prove. Hence, if there is a cycle to how units' names are listed in one's sampling frame, refrain from using a systematic sampling method. Opt for doing a SRS.

C. Stratified Sample

To do a stratified sample, one not only needs a list of one's theoretical population but one also needs to know at least one variable about each unit in the list. This information must be available before one begins data collection. So it is not enough just to have a list of the names of the

units. One must initially also know something about them. For example, if one's sampling frame is a list of all current full-time employees of the maintenance department, personnel could provide one with a list of names and also for each name an age, income, position title, how long they had worked for the city, whether they belonged to the union or not, etc. All the latter information are variables that can be used to divide the personnel list into stratas before one draws a sample. If one's sampling frame is a list of all counties in the state of Illinois, information exists in various resource books about the population size of the counties, the median income, political party registration, ethnic make-up, etc. These latter variables or characteristics of the counties can be used to divide the county list into strata before any sample is drawn.

The reason one wants to be able to divide one's sampling frame into strata of units with something in common is that it reduces sampling error. The result is a sample for the same cost as a SRS or systematic but one that is a more accurate representation of the theoretical population. The logic behind this reduction in error is that there can be no sampling error if one is choosing units for the sample from a group in which each unit looks exactly alike. So if one were choosing a sample of police cars from a list of all police cars delivered on August 1 to one's city, no matter which car one selected it would be a 1996 Chevrolet Impala. But if one were randomly choosing a sample of cars from a list of all cars delivered on August 1 to one's city, one might not get one Chevrolet Impala because only the police department ordered that make and model. The resulting sample would not reflect the variation of kinds of cars delivered to the city as of August 1.

Here is how to draw a stratified sample. Begin with a list of units and at least one known variable on each of the units. Let us assume the researcher is a supervisor in the maintenance department and wants to devise a routine maintenance schedule for the city owned vehicles. To do so, the supervisor wants to check on past maintenance records of the vehicles, how often they had a routine inspection and how often they were sent to the yard with problems. Because the city owns over a thousand vehicles, the supervisor decides to do a sample. A stratified sampling technique is possible because one knows which department was assigned each vehicle.

The researcher orders the list of all city vehicles by department. Thus, there is a stratum or group of vehicles assigned to the mayor's office, a stratum of vehicles assigned to refuse, a stratum assigned to police, a stratum assigned to parks and recreation, etc. Then one does a SRS or a systematic sample within each stratum. So if police have one hundred vehicles, one numbers the vehicles and randomly chooses which vehicles' records will be inspected (see Appendix B).

Vehicles chosen by a SRS would be determined by reading three digits in a random number chart. Again ignore any random number that does not match a police car's number. To choose a systematic sample of police cars, one needs to determine the sampling interval (i.e., divide the number of police cars by the number of cars one wants in one's sample from this department). Then find the first number between one and the sampling interval to appear when reading the random number chart. The police car with that number is selected for one's sample. Add the sampling interval to the random start number and select the police car with that number. Continue adding the sampling interval to the last number chosen, and one will get a sample of cars in the police department. To get the sample of all city owned vehicles, one must repeat this procedure for each department or stratum. Note that the sample will likely end up with a vehicle from every department in the city. A department will not be represented only if it has less vehicles assigned to it than the sampling interval. There is no way, though, that a stratified sample does not reflect the population's strata, in this case departments with assigned vehicles.

1. Proportional or Nonproportional

An issue in stratified sampling is how many units to select from each stratum. Because the aim of sampling is to choose a sample that looks like the theoretical population, one normally wants

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

one's stratified sample to look like the sampling frame in terms of the variable on which one stratified. If the police department has 20% of the city owned vehicles and the refuse department has 30%, 20% of one's sample should be chosen from the police strata and 30% from the refuse strata. This is called proportional stratified sampling. One samples each stratum in proportion to its size in the sampling frame. If one wants 100 vehicles in the sample, 20 or 20% would need to be chosen randomly from the police vehicle list and 30 or 30% from the refuse list. In this way the sample would perfectly reflect the distribution of city owned vehicles assigned by department. Only through stratified sampling can one insure this perfect department representation. Using a SRS or systematic method for choosing the vehicles, normally will result in a sample of vehicles that is close to the actual department vehicle assignment but not as close as using a stratified. The stratified sample therefore reduces sampling error on the variable one has used to divide the units into stratas.

A stratified sample also reduces the sampling error on any other variables that may be related to the strata's variable. For example, not only does stratification reduce error on choosing police cars for the study, but it also reduces error on another variable, how many drivers per car. All police cars are driven by different people because of the 24-hour nature of police work. In contrast, in other departments cars are often assigned to an individual; these cars have only one driver. Cars driven by different drivers may experience the need for more frequent maintenance than cars driven by the same driver. As the supervisor, one would want to see if this is true. The suggested stratified sample would allow one to assess more accurately this factor.

The aim of sampling is to choose a sample that accurately reflects the population's characteristics. This is the logic behind proportional stratified sampling. There are instances, though, when it may be worthwhile to do nonproportional sampling. If one or more of the stratas are so small that none or less than five units from that strata would be chosen through proportional sampling, then one may wish to over sample that stratum. To over sample a stratum one just selects more units from that strata then one would through proportional sampling. The presumption is that those very small stratas are of interest to the study's purpose. If only one car is assigned to a department, it may not make sense to make sure that car ends up in the sample. Then again, if that one car is assigned to the mayor, one may want to sample that car to insure that the mayor is never left on the side of the road with a disabled car. Cities with small but politically active Latino populations or senior citizens, may want to over sample the Latino or senior citizen stratas to understand more accurately the concerns of that group.

Of major importance, whenever using the whole sample to state findings, one must restore the over sampled strata to their size in proportion to the population. One uses nonproportional sampling to learn about the strata individually, never to learn about the population as a whole. To draw conclusions about the population, one must use proportional sampling. If one has used nonproportional sampling of a stratum and wishes to also speak about the whole population, one must weight the over sampled stratum back to its proper proportion of the population.

To over sample Latino school children, one randomly selects more of their names from the school provided lists than their proportion of all public school children. These data allow one to talk about Latino school children. When one wants to talk about all public school children, one needs to weight the Latino children back to their proportion of the school population and combine the data with the data from the other stratas. If Latino's are eight percent of the school population but one sampled twice that amount, one multiplies the Latino responses by one-half to weight their responses back to their proper proportion. Now one has a proportional sample again.

2. Choice of Strata Variable(s)

What variable(s) to stratify on is an important consideration. Sampling error is only reduced if the variable on which one stratifies is related to the purpose of the study. If one wants to sample

employees on their benefit packages, choose a variable to stratify on that can affect their opinions of benefits, such as sex, age, or department. Female employees may be more interested in whether they can use sick days for personal business or if child care is available on the premises. Older employees may be more interested in retirement benefits, and safety employees may be more concerned with disability rules and paid survivor insurance.

One can use more than one variable in dividing the sampling frame into the strata. The more variables used, the less sampling error. The addition of a second variable at least doubles the number of stratum and so complicates the choosing of the sample. If one were stratifying on sex and whether or not the employee was safety personnel, one would have four strata: female safety personnel, female nonsafety, male safety, and male non-safety. One reduces sampling error to the extent that the second variable is related to the study's purpose and is unrelated to the first variable. So do not choose a second variable to stratify on if it is highly associated with the first one even if it is related to the study's purpose. Therefore, one probably does not want to stratify on sex and safety personnel status in the example if safety personnel tend to be overwhelmingly male and nonsafety overwhelmingly female. The addition of sex as a second stratifying variable will not reduce one's error much but will increase the effort involved in drawing the sample.

D. Cluster Sample

A cluster sample is one's only choice if one does not have a list of the theoretical population, and it involves too many resources to get such a list. For example, there is no list of all adult Americans, except the US census that very quickly gets out of date and access to actual names is severely limited. There also are not lists of all adults in any city or county or state in the US. To obtain such a list is beyond the resources of any governmental unit besides the federal government. And the federal government only does its population census because it is mandated in the US Constitution. In fact, in the past there has been discussion in the Bureau of the Census to substitute a probability sample for the census. The Census Bureau currently uses a sample to check on the accuracy of the census. It first used a sample in 1850 when 23 counties were sampled to check on marital and educational trends in US society. Clearly, a sample would be less costly. A sample would also be more accurate, especially given the low mail response rate to the 1990 census. The hitch is getting around the wording of the Constitution.

The Census Bureau has come up with a new use for sampling for Census 2000. Instead of using probability sampling to check the accuracy of the census, probability sampling will be used to estimate the last 10% of the population who did not respond by mail or door-to-door interviewing.

Back to our question, what does one do if one cannot get a list of one's theoretical population? First, one can redefine the theoretical population so a list is possible. Change ''all adult citizens'' to ''all registered to vote citizens.'' Now a SRS, systematic, or stratified sample is possible. The problem is that one may have to redefine the theoretical population to such an extent that the then available list is inappropriate for one's purposes. If this happens, consider doing a cluster sample before one turns to considering a non-probability kind of sample.

To do any kind of probability sample one needs a list of units from which to sample. This is also true of a cluster sample. A cluster sample involves drawing at least two samples or, put another way, a cluster sample is drawn in at least two stages.

To illustrate, one wants to draw a sample of all adults living within two miles of a proposed baseball stadium. These would be the people most likely to feel the effects of the stadium in terms of traffic, noise, litter, and lights. No such list exists. One might consider redefining the theoretical population to households within the two mile limit. The city has access to a list of dwelling units. However, someone on the city council objects because she is concerned about

voters' reactions to the stadium and households do not represent potential voters. Back to the drawing board. The list of registered voters is rejected as a sampling frame because it would be unrepresentative of actual voters, especially when a hot issue increases late registration and turnout. Finally, the city council accepts that the only way to find out how adult citizens feel about the stadium is to do a cluster sample. The money is allocated with the proviso that interns are used to do the enumeration.

To carry out this cluster sample, a list of city blocks in the two mile radius of the proposed stadium is developed by the planning staff. A SRS or a systematic sample of those blocks is drawn. This is the first stage of the cluster sample. Next, to get a list of the right units, adult citizens, interns are sent to the selected blocks and literally go door to door, writing down the names of all residents eighteen years or older. Hospitals, nursing homes, and institutions like halfway houses are traditionally left out of the enumeration. Residents of such facilities are considered transients, and many would be incapable of responding to the subsequent interviewer or mailed questionnaire.

Using the new list of adults gathered by the interns, a SRS or systematic sample is drawn of the respondents to be sampled for their reactions to the proposed baseball stadium. This second sampling is the second stage of the cluster sample.

As one can probably tell, a cluster sample is more expensive and time consuming to do than the first three kinds of probability samples because a cluster sample involves sending staff or volunteers to areas to develop a list of the right units. Still, it is the only type of probability sample that is possible if no appropriate list of the theoretical population exists.

The sampling error rate can be computed for a cluster sample just as it can be for the other kinds of probability samples. A cluster sample involves higher error, though, because at least two samples are drawn. To compensate for the higher error rate in a cluster sample, one can increase one's sample size by 50 percent.

E. Random Digit Dialing

Telephones are a quicker and cheaper way to gather information than door-to-door interviews. The difficulty is that today so many people have unlisted numbers. For example, up to 60 percent of numbers in Los Angeles are unlisted. As a result, the phone book is a very inaccurate list to use for a sampling frame. So random digit dialing is used. Random digit dialing is a form of a cluster sample.

First, one develops a list of the area codes, only if more than one area code is used in the area one wants to survey. One then develops a list of the central-office codes in each area code. The central-office codes are the first three digits of the seven digit phone number. To choose which phone numbers will be called, one randomly chooses an area code, then randomly chooses a central-office number. Given that both these numbers are three digits, one would read three digits off the random number chart. Then one needs to randomly choose a four digit number to get a full phone number.

When a phone number is dialed, one then randomly chooses which adult in the household will be asked to answer the questionnaire. This is the second sampling stage. The interviewer has to ask whoever answers the phone how many adults there are in the household and then must randomly choose who in that household is asked to answer the questions. Lists based on sex and age can be used to avoid household enumeration. One randomly chooses combinations of sex and age groups to interview from phone number to phone number. For example, on one call the interviewer asks to speak to the oldest adult female in the household, and on the next phone call the interviewer asks to speak to the second oldest adult male in the household. Asking to speak to the adult in the household whom most recently had a birthday can also be used to pick respondents.

A major problem with random digit dialing is the number of phone numbers which are inoperative numbers. Phone companies where possible do assign new numbers in groups of sequential numbers. If this is true in the area being sampled, then once an operating phone number has been found by the above random method, one can randomly select more random numbers around the operating number as long as one stays within +/−100 units. If the working number is 999–2424, for instance, we might also select 999–2456 or 999–2392.

Conducting a telephone survey is complicated. One might want to seriously consider at least hiring a sampling expert or a survey firm to design the sampling procedure. There are also other major issues, such as training the interviewers and supervising their work and establishing a callback procedure for calls in which no one is home or the right respondent is unavailable at that time.

F. A Future Cities' Sampling Design

Innovative sampling designs are rare and are variations of SRS, systematic, and stratified. Even a cluster sample is a multistage SRS or systematic sample. A unique stratified sampling design was developed in the 1970s at the University of California, Irvine (Kraemer et al., 1981). The aim was to draw a sample of cities that would reflect not the current characteristics of cities but the characteristics of possible future cities. Sampling theory presumes one wants to draw a sample to describe the current theoretical population. In other words, data are collected from the sample at one point in time to learn about the theoretical population as of the same point in time. The aim of the future cities' design departs from this typical intention behind sampling.

The design was developed so that the researchers could answer what would happen to cities if they did x, x being a policy relating to computerization. In order to stratify they needed to know about computer policy in each city. No such information existed, so a survey was done of all cities over 50,000 in population, asking extensive questions about computer policies. From this survey six policy variables were chosen on which to stratify the cities. Note that the sampling frame was being divided into strata based on combinations of six variables, not one variable as described above in detail. Each of the six variables was dichotomized, so they each had two categories. The possible strata or combinations of six variables with two categories are 64. Resources limited the study to site visits to only forty cities although there were 64 strata. So 40 strata were randomly chosen, and a city from each of these 40 strata was then randomly chosen.

The result was a stratified sample that represented possible variation on computing policy not current variation. From this sample, the researchers would be able to collect data that could answer what would happen if cities had this computing policy or that one. A more typical sampling method might not have picked cities for the sample that had a certain type of policy because that policy would have been rare or rare in combination with other policies. Hence, a typical sample would not have allowed researchers to discuss the effects of rare policies.

This innovative sampling method was expensive and time consuming because a survey of the theoretical population had to be carried out to find information about each unit in order to be able to stratify. Drawing the actual sample from this information was also time consuming because of the large number of strata and the fact that not all strata were represented by a real city. A further complication was obtaining permission to study the city's operations from various officials in each city chosen for the sample.

G. Sampling the Role and Not the Person

Another innovation in sampling design is sampling the role and not the person. This approach is particularly appropriate for longitudinal data collection in organizations. People leave organi-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

zations or change their positions in organizations. Thus, it is not always possible nor appropriate to ask the same respondent questions at a second point in time. Thus, if one were conducting an evaluation of a training program or a change in policy, one does not necessarily have to sample the same person at each point of time in data collection.

The key is getting data from the person who holds a particular job that has a perception of the effects of the policy. Asking employees about their morale before and after a policy change requires only that the employees still work for the organization and hold the same job responsibilities so that they are affected in the same way. It does not require that they be the same individuals.

Moreover, the size of one's sample would be greatly reduced if one had to sample the same people where there is high job turnover. By sampling the role and not the person, sample size can be maintained at the two points in time.

H. Sampling Without Replacement

One issue in probability sampling that has not been addressed is sampling without replacement. Probability sampling presumes each unit in the sampling frame has an equal chance of being chosen for the sample. If each unit has an equal chance of being chosen, then there are no biases in selecting the sample. However, this assumption is often violated. A number can appear more than once in the random number chart. Thus, when using the random number chart to select a SRS, a unit in the sampling frame could be selected twice. It does not make sense to interview the same individual twice. So in practice, if a number is selected more than once, that repeat number is ignored. One is actually throwing numbers out of consideration for the sample once they have been selected. The result is that the units in the sampling frame do not have equal chances of being chosen for the sample.

To illustrate, if there are 200 numbers in the sampling frame and a sample of twenty is desired, each unit in the sampling frame has a one out of 10 chance of being chosen. This is before the first random number is chosen. Once a random number is chosen and cannot be chosen again, the remaining 199 units in the sampling frame have slightly more than one out of 10 chances of being chosen. Every time a number is selected and then retired, the chances of being selected for the other units goes down.

When the sample is small compared to the size of the sampling frame, there is a negligible error introduced by throwing numbers out of consideration once selected. Systematic sampling avoids this error all together.

I. Reporting Sample Design

Although it is critical to include in one's report the sampling method and all its essential characteristics, it may not always be appropriate to impose this information at the beginning of the report. One should consider one's audience. If the report is prepared for public officials or public dissemination, the sampling information should be put in an appendix. In fact, major publications, like the *Gallup Poll Monthly*, reserve a section under the title ''Design of the Sample'' at the end of each issue. This enables them to present the findings without the burden of a long, technical introduction. However, putting the information in an appendix may increase suspicion about the quality of the data. If this is possible, one may want to explain where the information can be found when not introduced in the beginning of the report.

For academic audiences, it is crucial to describe one's sample design in the beginning to acquire support and recognition from one's research colleagues, who would not consider any of one's claims unless properly informed on the quality of one's data. *The Public Opinion*

Quarterly and the *Public Administration Review* strictly enforce such up front reporting in every published article.

IV. NONPROBABILITY SAMPLING

Probability samples can be expensive and thus beyond the reach of researchers and organizations. There may also not be the need to go to the trouble and expense of doing a probability survey. Nonprobability samples are one's alternative. The key distinction between probability samples and nonprobability samples is that in the first the researchers have no influence over which units in the sampling frame end up in the sample; the opposite is true in the latter. It is also true that only in probability samples is it possible to compute sampling error, i.e., to what extent does collecting data from the sample differ from collecting data from the whole sampling frame.

A. Judgmental or Reputation Sample

A judgmental or reputation sample is a common kind of sample. Many articles in the *Public Administration Review* are based on studies from one state's experiences or a few cities'. The state or cities are chosen based on their reputations for success or, sometimes, very costly failure in policy making. For example, if one hears that a city has really reduced its trash collection costs, one would likely want to talk with that city and find out how. If a city similar to one's own has made a major advance or suffered a severe loss, it might be beneficial to explore why. Thus, based on reputation, one seeks out certain units or samples from which to gather data. This form of data collection makes sense because it fits with people's logical tendencies to learn from example.

There may be other cities, though, who were even more successful or less successful. Using the reputation approach, one may actually be learning from cities whose experiences are totally unique and unable to be guideposts. The lessons learned would therefore be misleading. That is the weakness of a reputation sample. There is no way to know how representative or typical the experiences of the sample units are. The strengths of a reputation sample are that limited resources can be expended and an in-depth understanding of a situation can be undertaken.

B. Convenience Sample

A convenience sample involves choosing units to study that are readily available to the researcher. Many articles in the *Public Administration Review* and much research in the field are based on samples of convenience. Studies done by government employees based on their work experiences are studies of convenience. Academics find it convenient to study cities that are located near their universities.

A sample of convenience is the least likely to reflect the larger theoretical population. While a reputation and a convenience sample may seem similar, they are not. Unlike a convenience sample, the reputation sample is chosen on a criterion independent of the researcher, its reputation. Still, it is true that like a reputation sample a convenience sample is less expensive than a probability sample and allows for in-depth study.

To improve the quality of a convenience sample, one can vary the time and place for selecting the units. This is a good idea if the units of analysis are individuals. Then one can collect data from individuals at different points of time in different cities or departments or

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

classes. For instance, if an instructor wanted to assess students' satisfaction with his or her classes offered by the Parks and Recreation Department, it would be wise to survey students in different classes in different sessions. One could vary the time by both time of day (morning, afternoon, evening, weekend) and by session (spring, summer, winter).

C. Quota Sample

A quota sample is an important kind of nonprobability sample. It is used often in marketing research and election polls. It also bears a similarity to a stratified sample. In a quota sample the researcher sets quotas on key variables that will shape who is chosen for the sample. The quotas are characteristics of the theoretical population. For example, if one knows the sex and age and racial make-up of the population, one then sets the same ratios for the sample. So if ten percent of the population is African-American, one sets the quota of ten percent African-American for the sample.

The advantage of a quota sample over other nonprobability samples is that a quota sample insures that the sample looks like the theoretical population on the variables on which quotas have been set. However, the researcher gets to choose who ends up in the sample within the quota framework. So if one were standing outside a supermarket doing a quota survey, one would be more likely to approach the smiling male over 35 than the male who is moving fast and avoiding eye contact even though he is also over 35 and thus meets the quota. The biases introduced by the researcher in selecting who ends up in the sample within the quotas is why quota sampling is a nonprobability sample.

Like the first two kinds of nonprobability samples, a quota sample is cheaper than a probability sample of the same size. It can also present good data. The quality of a quota sample rests on choosing good variables on which to set quotas. The best variables are ones highly related to the purpose of the study. So if one wants to know residents' reactions to building a skate board park, set quotas on variables that are likely to affect reactions, such as residence's proximity to the park and whether a household has children under sixteen. Information gathered from such a quota sample would be an excellent supplement to public hearings. The quota sample could balance the bias inherent in using public hearings to gauge wider public sentiment. Moreover, the quota sample can provide the additional information for far less cost and in much faster time than a probability sample could.

The quality of a quota sample also rests on how the units are chosen within the quota allowances. One could give the interviewers a list of one digit random numbers. The interviewers choose who is interviewed using the numbers and not whom they personally would pick. So within the quota of a white, female, under thirty, the interviewer can only approach the woman who fits that description and the random number. If the random number is three, the interviewer can only approach the third woman who fits the quota description. The random numbers are used similarly to pick respondents within the other quotas.

D. Volunteer Sample

A volunteer sample is another type of nonprobability sample, but one that is more common in other fields such as medical research. Sample members are self chosen; they nominate themselves for the sample, following some form of public announcement. Like the other types of nonprobability samples, there is no way to know if the sample is at all representative of the theoretical population. Clearly, a sample based on volunteers looks different from the theoretical population because the sample were the only ones interested in participating.

Volunteers are more motivated for a variety of reasons. In medical research, volunteers

may be more seriously ill than the wider theoretical population suffering from the disease. Hence, such volunteers may be less likely to respond to treatment because the disease has progressed beyond the point of help. Accordingly, the treatment being evaluated may look less successful than it would if tested on a more representative sample.

Programs can also look more successful. In 1995 the Army did a study to see if women were capable of doing very heavy military tasks, involving lifting a 100 lb weight. Using female volunteers from various occupations, the Army measured their strength before a training program and then after completing the program. The results showed a dramatic increase in the volunteers' abilities to lift 100 pounds. The impressive success of the program may be distorted though. The volunteers were likely much more motivated to be faithful to the training and give it their best than a group of women who were required to do so. Many of the volunteers had never exercised before. Some had just had children and wanted to get back into shape. The study did show women can do heavy military tasks like loading trucks and marching under the weight of a full backpack. The study does not show that all or many female recruits would respond as well to a weight training program and thus be able to load trucks.

In 1992 the Long Beach Police Department in California wanted citizen input so that a strategic plan could be developed. Surveys were published in two of the city's newspapers and placed in every public library. This is an example of a volunteer sample that tries to get at a wide segment of the population by using multiple ways to gather respondents. In this case a random segment of citizens were also surveyed. The volunteer sample made sense as a complement to the probability sample for political reasons. Specifically, volunteer samples can serve a valuable purpose by giving citizens or consumers or clients an outlet to express their opinions. Individuals not selected in a random sample can feel that they are being ignored. By supplementing random samples with volunteer samples, an important outlet for discontent is provided. Political discontent may thereby be reduced. Such an approach may also increase acceptance of reports based on the data analysis.

Sometimes, a volunteer sample makes sense because the public policy being evaluated is based on volunteer participation. Consider school magnet programs, which are programs that select students who volunteer for the programs. The students and their parents want to participate because the programs promise specialized, advanced schooling in academic areas. To accurately evaluate school magnet programs, one should take a random sample of programs and a random sample of students in the programs. But if one wanted to do in-depth interviewing of students and parents, a random sample might be too costly and might suffer from a low response rate. The option is to ask for volunteers in the program. Using volunteers from a volunteer based theoretical population is less susceptible to the unique error of selfselection inherent in volunteer samples.

V. HOW BIG A SAMPLE

There are a variety of factors that shape how big a sample is needed. Resources both in terms of staff and budget have to be balanced against getting the kind of data that is needed. The kind of data that is needed depends on the required accuracy of the conclusions, the detail of the analysis, and the political needs of who gets the report.

A. Sampling Error

The issue of how much sampling error can be tolerated only applies to probability samples. Remember that the difference between probability and nonprobability samples is that only in

the former can one say how much the sample differs from the theoretical population. This is because only probability samples are chosen randomly. Random selection of samples involves two kinds of error, one is confidence interval and one is confidence level.

1. Confidence Interval

This type of random sampling error is the best known and is now consistently reported in news stories. Confidence interval is expressed as a ± percentage. So if the confidence interval is ± three percent, it would mean that the data from the sample fall within the ± three percent range as compared to what the results would be using the whole theoretical population. More concretely, if President Clinton's approval rating is 61 percent according to the latest Gallup sample, the President's rating is actually between 58 and 64 percent among all adults.

Confidence interval dictates sample size. The less error one wants, the larger the sample. Since larger samples eat scarce resources, it is important to set the interval according to how much error is needed to draw conclusions.

If there is a controversial ordinance proposed which appears to divide the community, a smaller interval should be chosen, such as three percent. This way the results of the survey can be used to drive the decision whether to approve the ordinance or not, reflecting the public will being the goal.

Often, though, surveys are done that are more exploratory in nature. If one want to know what public service areas citizens consider problematic, then ballpark results are just fine. It does not really matter whether 10 or 12 percent of the citizens surveyed think parks are a problem, the point is it is a low percentage. So what if the interval error is ± even five percent. One would still know only a small percentage feels parks are a problem.

Determining the interval is based on how accurate one need the results to be. Seven percent is probably the widest one would want to set the interval. Seven percent is actually a 14 percent range of error, which is getting large enough to cause serious distortion of data results. The Gallup Poll organization will not do surveys with confidence intervals larger than five percent.

Consider that the food packaging error rate for calories is set at 20 percent. What this means is that a serving portion of 100 calories could really be as low as 80 or as high as 120. If one were on a strict diet, this error rate might be too large to allow for weight loss, or at the other end of error it might be too large to allow for faithful dieting without serious feelings of starvation.

2. Confidence Level

Confidence interval cannot be understood without knowing confidence level. The two random sampling errors are interpreted together. Confidence level refers to the percentage of time that the sample differs from the theoretical population within the confidence interval. It is not true that the sample looks like the theoretical population always within the ± percentage set as the interval.

Confidence level is expressed as a percentage. It can range from one percent to 99 percent. Realistically, confidence level should never fall below 90, 95 percent being the most common. So if the confidence level is 95 percent and the interval is three percent, then in 95 samples out of 100 the sample will differ from the theoretical population by ± three percent. There is a five percent chance that the sample differs from the theoretical population by more than ± three percent.

Confidence level dictates sample size along with confidence interval. The higher the confidence level, which means the more accurate, the larger the sample must be. Like one's decision

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

about how big an interval should be, the size of the confidence level should be dictated by how accurate one needs the data.

Once one has chosen a confidence level and interval, one refers to established tables to know how big a sample is required for those error rates. If one's population is less than 10,000, there are tables based on the size of one's population that one references. If one's population is 10,000 or larger, then population size does not affect how big a sample is required. One merely finds where one's confidence level and interval bisect in the table and reads the sample size.

To give one an idea how error rates affect sample size for populations over 10,000, consider these examples. If one's confidence level is 95 percent and the confidence interval is one percent, the sample size is 9604. If one changes the interval to three percent, the sample size changes to 1067. If one changes the interval to seven percent, the sample size goes down to 196, a big saving in time and survey costs. If one raises the confidence level to 99 percent but keeps the interval at seven, the sample size becomes 339. Keeping the 99 percent level but lowering the interval to three results in a sample size of 1843.

Polling organizations tend to sample between 400 and 1200 adults for congressional districts or statewide surveys (Goldhaber, 1984). They use a 95 percent confidence level but vary the interval between three and five percent. Cost considerations are the major factors affecting sample size. In those districts or states that are very homogeneous, such as Wyoming, polling organizations find the smaller sample size quite acceptable because there is less random sampling error within homogeneous populations than within heterogeneous populations.

3. Detailed Analysis

The above explanation of sampling error applies to interpreting data based on the whole sample. What if one is also interested in learning about subgroups in one's sample? For example, one may want to know not only how the whole state feels about a law but also how the Latino population feels. When the Latinos are separated from the sample, then analysis is being done on a smaller sample. Sample size corresponds to confidence level and interval. So if one lowers the sample size, one raises the error. Therefore, to maintain the same confidence level and interval, one needs to increase the initial sample size to allow for analysis of the Latinos within the accuracy level predetermined. Or if one does not increase sample size, then interpretations of subgroups must take into account their higher sampling error.

An interval of three percent in national samples applies to data based on the whole sample. Breaking the sample into regions of the country to find out about how adults in the West or South feel results in error of about $\pm$ five percent. Analyzing different religious groups results in widely varying error because the sizes of the groups are so different. While the error for Protestants will stay within the three percent range, the error for Jews is 18 percent, plus or minus. The Jewish data would be all but worthless. If one wanted to learn anything about the Jewish population, one would have to increase sample size.

Sometimes it gets down to just having units in subgroups of the sample to analyze. Every year the US Justice Department samples 100,000 people. This incredibly large sample is necessary so that the Department can analyze crimes by sex of victim, age, time of day, type of crime, location such as suburb versus central city, etc. To do such analysis requires breaking the sample down into many different groups; just one such group would be female afternoon rape victim over sixty-five, living in central city. The Justice Department needs an incredibly large sample just to increase the chance that it will have such a person in the sample. For statistical analysis reasons one really wants at least five such bodies for each group. So if one wants to analyze one's sample in terms of fifty subgroups, then one would need a 250 person sample just to get

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

five in a group. Of course, one only needs a 250 person sample if one assumes all group memberships are evenly distributed and randomly occurring. The latter two assumptions never apply. So one needs even a bigger sample.

In sum, if one is going to do detailed analysis within subgroups of the sample, one needs to consider a larger sample size. The number of subgroups as well as the size of the subgroups in proportion to the size of the theoretical population affects how much the sample size should be increased.

B. Who Asked for the Data

Sample size does depend on how much error can be tolerated in probability samples. In both probability and nonprobability samples, there is another important consideration. Who asks for the data to be collected has a personal sense what would be an adequate sized sample. This sense may be unconnected to sampling theory or, rather, may reflect a lack of understanding of sampling theory. City councils have been known to reject out of hand that a sample of four hundred could possibly represent their citizenry. As an employee or outside consultant, one can politely lecture the council why 400 would be adequate. If this fails, try the numbers game. Point out to the council what a 5000 person sample would cost and what a 400 person sample would cost.

Do not underestimate the importance of what is an adequate sample size for those who asked for the data in the first place. Even if those who asked for it understand one's sampling error arguments, they may still want a much larger sample for political reasons. They may think that their constituents would not trust a survey's results based on four hundred people when there are 500,000 of them. This is a very strong argument for using a larger sample. Data based on a sample is only as good as people think it is. The best stratified sample with a two percent confidence interval is no good if the users or readers of the data think it is not large enough to be trusted.

C. Resources

The above factors determine the size of the sample that is needed. Once one has the number, the reality of the budget raises its ugly head. Time and money are scarce resources. One may neither have the time nor the money to collect data from as many units as one would want. It is time then to adjust sample size to reflect the resources available.

Time is a prime consideration. Data are collected for a reason. Perhaps data are needed before the next Board of Education meeting or by the June 16th council meeting. Data collection must be completed before that date. Reports are often ignored or paid less attention to if delivered the day of a big meeting. Data drawn from a probability sample are probably important and thus should be circulated in report form two weeks before a meeting that will make decisions related to those data. Data drawn from a nonprobability sample varies in importance, depending on how many resources have been expended to collect it. Important data that can affect big decisions should be circulated two weeks before the appropriate meeting. Less important data can be circulated closer to the meeting date.

Another time consideration also has to do with the length of time to collect the data. Independent of when the data is needed is the actual time spent collecting the data. Data's accuracy is affected by the length of time it takes to collect. Sampling theory presumes all data are collected at the very same time. This means that surveys of individuals presumes that all individuals answered the survey at the very same time. This assumption is impossible to meet. No organization has enough interviewers to make the phone calls or visits at the same time.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

There is also no control over when a respondent fills out a mail survey or even one distributed at work. There are exceptions, such as surveys of police officers that are done during watch meetings. The rule of thumb is that surveys of individuals should be completed within six weeks. After that time too many other factors could influence responses to trust that the responses of people surveyed the first day would be the same as their responses if interviewed the last day.

The rule of thumb if the unit of analysis is an organization is three months. A classic characteristic of an organization is its resistance to change. Therefore, organizations are expected to be less changeable than individuals, and data can be collected over a longer period of time.

Finally, there is the issue of money. Does the budget allow for data to be collected from the predetermined sample size? One needs to consider the production of questionnaires, mailing or interviewing costs, data cleaning and analysis costs, production of reports costs, etc. If one cannot afford the size of sample needed based on sampling error or views of who asked for the data, reconsider the initial needs. Beware. Increasing sampling error to get a smaller sample size may result in unusable data. This is the challenge of sampling, balancing resources against the quality of data needed.

VI. RESPONSE RATE

Sample size normally is not the same as number of units actually studied. Mail surveys have very low initial response rates, five to 20 percent returned. Telephone and interview surveys can also suffer from low response rates. Surveys that are filled out in a room with supervision will have good response rates.

What is a good response rate? A response rate of 85 percent is excellent. Response rates between 70 and 85 percent are considered very good. Over 60 percent is considered acceptable. Response rates between 50 and 68 percent are questionable. Below 50 percent is just not scientifically acceptable. Thus, a low response rate is simply a waste of resources.

For example, the US Agriculture Department regularly surveys citizens about what they eat. The data are used to regulate school lunches, food stamps, food labels, and pesticide exposures. These data are very important for they not only affect the health of millions of Americans but they also affect the spending of millions of dollars. The 1987–1988 food consumption survey was badly flawed due to its low response rate. While the contractor randomly sampled 9000 households, only 34 percent of the households responded. With two-thirds of households not represented in the data, the federal government was left with data that did not represent the consumption patterns of most Americans.

In contrast, the Gallup Poll has a 88 percent response rate to its telephone interviews (personal interview, 1996). This is twice the industry average and, obviously, is a great response rate. Gallup attributes its high response rate to the prestige associated with being a respondent to a Gallup poll. Thus, who is sponsoring or conducting a study contributes to the likelihood of responding and thereby the usefulness of the data.

In sum, response rate is critical to usefulness of data. There are a number of things that can be done to increase the likely response rate: envelope features, good cover letter, quality of questionnaire, postage, incentives, media blurbs, and follow-up.

A. Envelope Features

If mail delivery is necessary, then the first step in getting a high response rate is getting the respondent to open the envelope. If the survey is being distributed to people in an organization, use the type and color of envelope used for important notices. To illustrate, many offices use

a large brown envelope that can be reused over and over again by just crossing out the old name on it. This type of envelope is used for a variety of regular office communications. It does not signal that the contents are important nor that they need to be read promptly. Do not use this type of envelope. Instead, use the envelope that personnel uses to send out important notifications, such as promotion and benefit announcements. Employees recognize this latter type of envelope as containing important information in need of prompt attention. Thus, the respondents will very probably open it and do so quickly.

If the US mails are to be used, the envelope should not have a bulk rate stamp. Many people throw out mail by just looking at the envelope. Bulk rate stamps indicate mass mailings, which are unsolicited and often sales' pitches. Commemorative stamps look nice and do not send up such a red flag.

The envelope should be addressed, whenever possible, to the name of the household or household member. An envelope addressed to ''occupant'' is another ''please throw me away'' indicator. It is important to note that it is just fine to use mailing labels on the envelopes.

A return address can also encourage or discourage the recipient from opening the envelope. But one does need to be honest. Government or university addresses are good. Envelopes with the names of charities or some nonprofit groups can be viewed as solicitations and therefore thrown away without being opened.

It may be worthwhile to put a phrase alerting the recipient about the contents on the bottom of the envelope. Of course, this works only if it ties into a widely felt interest in the topic. ''Important Survey Inside'' does not always do it. ''Benefits Survey'' would work if the respondent were an employee or government aid recipient.

B. Good Cover Letter

Once one gets the addressee to open the envelope, one has to get them interested in filling out the questionnaire. The cover letter is key. Again, many people look at the letter and within reading just the first few lines make the decision to toss or continue reading. Therefore, the cover letter must give a good first impression.

The cover letter should be on letterhead. It should also be a very well spaced and short letter, one page maximum. Moreover, the letter, whenever possible, should use the respondent's name in the salutation. The letter should be signed with a signature and a title. More than one signature can be used if it encourages completion. For example, a survey being conducted by the Maintenance Department should bear the signature of the department head and the mayor or city manager.

In the very first sentence one needs to create the respondent's interest in reading further. By stating why the survey is important and important to them, one taps into their interest. If more motivation is needed, mention the bribe or incentive next. This is especially important in a survey that is being sent out to a wide community audience.

The letter needs to be honest about who is sponsoring the survey. Normally, this is done in the first sentence as one is trying to tap into the respondent's interest. Statements such as ''the city council and mayor want your input into whether or not the city should build a stadium'' get right to the point of both sponsorship and interest in participating.

Now, if the respondent makes it past the first paragraph, one still has to persuade them to fill out the survey. To do so, it is necessary to assuage some qualms they might have. Explain why they were chosen for the study and if the data are confidential. Also point out how little of their time the survey will take by saying a true estimate. An example would be ''this survey will take only ten minutes of your valuable time.''

Always in the last paragraph there should be directions about how to get the survey back

to the office and a stated date of arrival. For example, ''Please return the survey in the enclosed self addressed envelope by September 15.'' In this last paragraph one should also thank the respondent and give a phone number and name of a person they can call if they have any questions. Just by giving the respondent an option to check on the survey increases response rate and does not necessarily subject the organization to a flood of calls. While few people will call, there should still be a trained staff member prepared to receive any inquiries.

Obviously, the cover letter needs to be written at the language level of the respondent. It should also have a ''friendly'' style.

C. Quality of Questionnaire

Questionnaire construction is treated in a later chapter of this handbook. A well-constructed questionnaire not only produces useful data but also affects response rate. It is important that the questionnaire is easy to follow as well as simple to read. Otherwise, respondents may just quit filling out the questionnaire and never return it. Well-spaced questions, clear directions between sections, consistent set up of questions and response alignment all help to increase response rate.

The actual length of the questionnaire is also a factor that can affect response rate. There are no tried and true rules on what length questionnaire produces what response rate. Too many other factors affect response rate, such as the design features and interest of respondent in the survey.

D. Postage

In the envelope feature's section above, commemorative stamps were recommended for the envelope addressed to the respondent. There is also the issue of return postage if the US mail is being used. First, always provide the respondent with a return envelope, which has already been addressed and stamped. Second, a commemorative stamp on the envelope will again increase response rates. It seems that respondents feel a subtle pressure to return surveys when the envelopes have a stamp on them.

The use of a stamp on the return envelope does increase response rates but not dramatically. Therefore, weight the cost differential of using stamps versus business reply. With stamps one has to pay for each one whether or not they are used by respondents to mail the survey back. In contrast, with business reply the post office only charges for questionnaires returned. The post office does charge a little extra for this service, so that the additional cost needs to be factored in one's decision to use stamps or business reply. Do not forget that if one uses stamps, someone has to stick them on the envelopes.

E. Incentives

Incentives are rewards to encourage response. Whether monetary or material, incentives are effective ways to increase response rates. Incentives can be provided with the survey or upon return of the survey. They work best when the incentives are included with the survey. True, including the incentive with the survey is more expensive. The inclusion, though, works as a subtle contractual obligation, i.e., ''we gave you this, you now must do the survey.'' In contrast, when receiving the bribe depends on the return of a completed survey, the respondent does not need to feel remorse when they throw away the demanding piece of mail that requests work for a small reward.

Monetary rewards do not have to be large to increase response rates. In fact, three quarters

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

or a dollar included with the questionnaire so the respondent can buy a cup of coffee to drink while they fill out the survey works well. Monetary rewards are not always appropriate or even legal if the sponsoring agency is public. The legal issue is whether it is acceptable to give a monetary benefit to only the sample population and not the whole population.

Moreover, monetary compensation may not always be the best reward because some people's time may be more valuable than the small monetary incentive. Furthermore, depending upon who is being polled, money may be misjudged or even unnecessary. When people are concerned about an issue and are eager to make their opinions known, there may be no need for extra spending on such incentives. For example, when polling employees on work related issues that could affect their environment, a bribe may not be necessary to increase response rates. Of course, there may be other reasons for offering an incentive, such as showing respect for the employee's time.

Many material rewards are also possible. The key is that the reward must appeal to the entire sample population. If the sponsoring agency is the library, bookmarks can be included with hours or important city phone numbers listed. A water agency might use a hard water tester. A recent all purpose incentive is the refrigerator magnet on which numbers, information or pictures can be printed.

It is also possible to include coupons as the incentive. Parks and recreation might include a coupon of dollars off next class or team sign-up. This reward would work well if the sample were all past students or sport participants. If the city has an annual fair, free admission coupons work well if the survey is conducted close to the event.

Coupons do not have to cost the sponsoring agency anything. Local businesses often are willing to provide the coupons as a good will gesture or form of advertising. But he careful that the business has no relevance to the aim of the study. One does not want the incentive to introduce a bias. Also be careful that the solicitation of coupons from one business is not considered favoritism, that is if there is a rival business in the city, for example.

Finally, offering to supply the respondents with a brief summary of the study's results can be used as a motivation. This form of incentive may be most useful for an elite sampling population, such as city managers, police chiefs, or civic leaders. This reward for responding should not be used if the summary mailing would occur so far in the future that the respondents may not even remember doing the survey or find the results useful.

F. Media Blurbs or Prenotification

Another technique that can be used to increase response rates is prenotification. One to two weeks before the receipt of the survey one can contact the respondent by mail or phone alerting them that they have been selected to participate in the study. For instance, a brightly colored mailing or picture postcard is likely to be read and get the respondent's attention.

A cheaper form of prenotification would be to use media blurbs announcing the upcoming survey. The choice of media outlet depends on the sample population. If the sample population is the general community, newspapers or the city's cable channel can be used. If the sample population is concentrated in an organization or a building, such as employees, bulletin boards, e-mail, or newsletters are useful. If the sample population is members of a group that gets newsletters or periodic mailings, these outlets can provide prenotification.

G. Follow-Ups

A major technique used to increase response rates is doing follow-ups. After the initial questionnaire has been distributed, reminders to return the questionnaire should be done. Reminders can

be in the form of a postcard, letter, telephone call, e-mail, fax, or a complete redistribution. The method used for reminders depends upon budget and access to information, such as phone numbers.

If one has kept track of who has returned the questionnaire, follow-ups need only be done to those who have not responded. By putting an office code on the questionnaire, one can keep track of who has and has not responded. Thus, reminders can be sent only to those who have not responded.

Follow-ups should be done in two week intervals. The first reminder would go out two weeks or ten business days after the questionnaire distribution. The reminder should stress the importance of the study and also thank the respondent if they have already responded. It is also important to provide a number to call if the respondent has misplaced the questionnaire so a replacement can be sent.

If after the first follow-up, the return rate is over 80 percent, one may decide to stop collecting data or do one final follow-up. Return rates lower than 50 percent demand a second if not third follow-up. Remember that the worth of the data depends upon achieving as high a response rate as possible, 60 percent or higher is a must. One can expect to get about half the return rate in the second follow-up as one got in the first. So if one got 40 percent in the first follow-up, the second follow-up should produce about 20 percent return.

The last follow-up, if more than one, differs from earlier follow-ups in that a questionnaire as well as a reminder message may be included. Depending on budget and time passage since initial questionnaire distribution, it may be wise to include a questionnaire with the last reminder. After a month, it is likely that the initial questionnaire is lost.

Telephone reminders are tricky. They not only require access to phone numbers but also require trained staff to do the calls. Each staff member making the calls should operate from a set script. The script should include the purpose of the study, who is sponsoring it, and its importance. Staff should have pleasant, friendly voices.

If a telephone reminder is possible, one might want to consider giving the respondent the option of doing the survey over the phone. Again, there must be a set script to work from so each staff member responds to respondents' questions the same way. Offering this option requires that the staff go through some training about phrasing of the survey questions. A supervisor should also oversee the calls and verify responses by calling back a few respondents. Unfortunately, a recent study indicated that inviting the respondents to complete their questionnaires by phone did not significantly increase the response rate (Dillman et al., 1994).

Finally, a sample chosen by random digit dialing requires special follow-up considerations. One must keep track of each number called and list those where one got no response or when one needs to call at a different time for the proper respondent to be available. One should have different time slots available for call backs. Try to cover a morning, afternoon, early evening, and an evening, i.e., 9 AM to 12 PM, 12 PM to 4:30 PM, 4:30 PM to 6:30 PM, and 6:30 PM to 9 PM.

H. Demographic Analysis as a Check on Response Bias

When response rate is less than a 100 percent, there is the possibility that those who responded are different from those who did not. If possible, one needs to check for the extent of such a bias on the study's main variables. To do this, one needs to know some characteristic of the sampling population, such as percent that are male or female or the percent that are union or non-union members. Then one sees to what extent one's sample reflects that actual percent. If one finds that the sample has a higher or lower percent of men or women for example, one

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

needs to analyze the responses of the two sexes to some key questions. If they are significantly different, then one has an important response bias.

If there is a response bias, then one has two strategies one can pursue in presenting one's data. First, one can weight the sample to correct for the bias in response. In other words, if one has more women in the sample than in the sampling population, the women's answers are weighted less than men's. This is just like weighting in nonproportional sampling, which is discussed earlier in the chapter. Second, one can present all one's findings within their subgroups, in this case male and female.

Be on the look out for response bias as one is collecting data. An early analysis of the first wave of returned questionnaires may signal how critical multiple follow-ups are to the quality of the sample. Special messages can be put in the reminders to elicit more responses from the under represented group too.

VII. DATA COLLECTION IN DIFFERENT COUNTRIES

Even though one may need to collect data from units in countries other than the US, do not rush out to book one's tickets. One may be able to collect one's data in this country or use data already collected.

For example, one might be able to collect the data desired from foreigners right at one's fingertips. Foreign students at nearby US campuses can provide a sufficient population for one's sample. Be aware that the sample is foreign nationals but within the US, which may create a bias and thus be inappropriate but not always.

Recently, a pharmaceutical company interested in flooding Europe with its new cough syrup turned to European born and raised students who had been in the US for less than two years. By contacting foreign language departments and the Office of International Education at local universities, lists of appropriate students were obtained. Using the lists, the company drew a sample and had the selected students try out their European taste buds on the American cough syrup.

When one really needs the data to come from another country, one should always consider the resources that the country may have to offer. Many countries carry out a census of their population every ten years. While the censuses are mostly demographic information, they often also cover other topics.

Another possibility is a research center or a university in the foreign country that also is interested in the same topic. They may have data or be willing to help in the data collection.

If it turns out that no existing data are relevant, one should probably resort to a local agency to conduct the survey. There are too many cultural differences that can affect data collection in different countries. Language nuances, types of incentives, how many people have phones or unique addresses, and even whether voter or household lists are available to nongovernment employees are some of the issues that show why one needs a native agency or professional to guide the survey.

Some final words of wisdom on data from foreign countries should be offered. It is amazing what is considered unacceptable in the US but perfectly normal in other countries. For example, quota sampling has been regarded as unacceptable in the US for forty years. But many countries like Mexico, France, and the United Kingdom routinely use quota samples. Amazingly, in those countries quota sampling has worked well in public opinion polling. Random digit dialing is considered the only acceptable telephone interviewing sampling method in the US. Yet, many countries do not use it. Some countries like Italy use quota and others like Denmark use directories for their sampling frames.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

VIII. CONCLUSION

The theme of this chapter has been sampling and data collection. There are many important issues to address in doing both. In the end the quality of one's research report rests on the quality of one's data collection. One's ability to do certain statistical analyses is also dependent on the data collection, particularly sample size. Once data are collected, though, there continue to be challenges to the quality of one's research. A large, probability sample done overtime sounds like great data. These data must be accurately transformed to a computer file to maintain their integrity. The problems faced in constructing data sets is the topic of the next chapter.

APPENDIX A: READING A RANDOM DIGIT CHART

How Many Digits to Read?

1. If theoretical population has less than 10 units, read one digit.

   (5) (4) 7 9 3 3 0 6 4

2. If theoretical population has between 10 and 99 units, read two digits.

   (5 4) (7 9) 3 3 0 6 4

3. If theoretical population has between 100 and 999 units, read three digits.

   (5 4 7) (9 3 3) 0 6 4

Where to Start to Read Chart?

The answer is anywhere.
If one decides to read rows and are looking for 3 numbers between 1 and 12, then the numbers would be:

54 79 33 (06) 41 99 43 96 95 34
(01) 49 35 20 27 92 63 20 67 (02)

If one decides to read columns, the numbers would be:

54 01
(01) 92
23 16
60 92
31 (07)
88
59
80
79
(03)

APPENDIX B: ILLUSTRATION OF LIST OF THEORETICAL POPULATION FOR SRS VS. STRATIFIED SAMPLE

Theoretical Population: all city owned vehicles as of August 1, 1996.

SRS: numbered list of all city owned vehicles.

1. 1995 Ford Escort
2. 1996 Ford Escort
3. 1996 Ford Escort
4. 1993 Ford Taurus
5. 1994 Ford Taurus

6–105. 1996 Chevrolet Impalas

Stratified: numbered list of all city owned vehicles stratified by city department.

Police	*Mayor's Office*	*Parks & Recreation*	*Refuse*
1–100. 1996 Chevrolet Impalas	1. 1993 Ford Taurus 2. 1994 Ford Taurus	1. 1995 Ford Escort 2. 1996 Ford Escort	1. 1996 Ford Escort

REFERENCES

D.A. Dillman, K.K. West, and J.R. Clark (1994). ''Influence of an Invitation to Answer by Telephone on Response to Census Questionnaires,'' *Public Opinion Quarterly*, *58*:557.

G.M. Goldhaber (1984). ''A Pollsters' Sampler,'' *Public Opinion, 53*: 47–50.

K.L. Kraemer, W.H. Dutton, and A. Northrop (1981). *The Management of Information Systems*, Columbia University Press, New York.

Personal interview with Tom Reiger at Gallup Poll, Irvine California, March 1, 1996.

RAND Corporation (1955). *A Million Random Digits*, Free Press, New York.

H. Taylor (1995). ''Horses for Courses: How Different Countries Measure Public Opinion in Very Different Ways,'' *Public Perspective*, (February/March): 3–7.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

8

Constructing Data Sets and Manipulating Data

Carmine P. F. Scavo
East Carolina University, Greenville, North Carolina

I. INTRODUCTION

The word data is often used in very general ways—get me the data on Jones; I'm collecting data on my ancestors; we need some data to support this grant application. These uses of the word data are far too general to be useful in the context of this chapter. We will thus utilize a much more limited definition of the word. In the phrasing of Clyde Coombs, psychometrician and author of the classic text, *A Theory of Data*, ''Data may be viewed as relations between points in space.''[1] This use of the word data assumes a mathematical framework that makes the gathering of data, construction of data sets, and the manipulation of data much more logical and understandable, and so we will adopt that definition of data in this chapter. This chapter begins by looking at data, how they are collected and prepared for analysis. The chapter then looks at how data sets are typically structured, how data can be manipulated, changed, recalculated, and so on. Later, the chapter looks at the advantages and disadvantages of using precollected, archived data rather than collecting data for oneself. And last, the chapter looks at ways that data can be reformatted, recalculated, and otherwise changed to fit the user's needs.

II. THE NATURE OF DATA

Data can be thought of as observations about events in the real world. A data set can be thought of as a set of observations, one or more for each unit in which we are interested. Units can be individuals, nations, states, communities, neighborhoods, census tracts, or any other unique entity in which we are interested. Typically, a data set is structured as a matrix, each line representing a different unit or case and each column representing a different observation or variable.[2] For example, each unit or case may be a person for whom data have been collected. Each column would represent an observation on this person—gender, for example. One column in the matrix would have a code—a number or letter—which would designate whether the individual was a male or a female. A unit or case might be a large city in a data set of US cities with populations over 100,000. A range of nine columns in the data set might contain each city's 1990 population as gathered by the US Bureau of the Census. A range of nine columns would be required since the largest US cities—New York, Los Angeles, Chicago, Philadelphia—have populations over one million but under 10 million. The smallest sized cities would have only the six farthest

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

right columns occupied with numbers—since their populations would be slightly larger than 100,000—while the largest cities would have numbers in all nine columns.[3] While the numbers designating the population of a large city have their own intrinsic meaning, other numbers in a data set might not. If we had collected data on the gender of individuals surveyed about quality of city services in a large city, we might code female as ''1'' and male as ''2.'' Naturally, there is no intrinsic meaning in these codes; male could just as easily have been ''1'' and female ''2.'' What is important here, however, is telling the computer what code has been utilized to signify male and what code has been utilized to signify female.

III. SCALING

The numbers used in a data set will be part of one of the measurement scales that social scientists use—nominal, ordinal, interval, or ratio. Nominal scales are those in which observations that share the same value are assigned the same code. Ordinal scales are those in which the order of numbers assigned reflects an underlying ordering in the observations. Interval level scales are those in which differences between the codes reflect differences between the observations. Ratio level scales are those in which differences and ratios between the codes reflect differences and ratios between the observations.[4] Consider the following example from a survey on quality of city services:

How satisfied or dissatisfied are you with the job that the city Fire Department is doing?

1. Very satisfied
2. Somewhat satisfied
3. Somewhat dissatisfied
4. Very dissatisfied

The variable constructed to map this question into our data set would have codes ranging from ''1'' to ''4'' and each of these codes would be associated with the statement in the example. The scale is ordinal meaning that Very Satisfied indicates a higher level of satisfaction than Somewhat Satisfied, etc. but we do not know how much higher a level of satisfaction answering Very Satisfied conveys than answering Somewhat Satisfied. The actual numbers assigned to each of the responses to the question above are, in essence, arbitrary. Rather than coding the responses from ''1'' to ''4,'' they could have been coded from ''−2'' to ''+2'' (omitting ''0'') or from ''10'' to ''40,'' if we counted by tens rather than by ones. Or we could code the responses in the opposite direction—''1'' would be Very Dissatisfied; ''2'' Somewhat Dissatisfied; and so on. Measurement theory is the branch of social science that is concerned with scaling. Much of this theory was developed by the psychologist S. S. Stevens.[5] Perhaps the best explanation of Stevens' work is an article by Sarle[6] who explains that the various scales used to code data are defined by the permissible transformations that can be performed on the data without changing the data's underlying meaning. Thus:

> Nominal level scales can undergo any one-to-one or many-to-one transformation. Thus, if we were coding ethnicity and had established codes as ''1'' White; ''2'' African-American; ''3'' Asian; ''4'' Hispanic; ''5'' Other, we could recode this variable by changing the ordering of the codes (''2'' could become Hispanic and ''4'' African-American, etc.) or by ''collapsing'' the scale—''1'' might remain White while ''2,'' ''3,'' and ''4'' would be recoded into a new category of Non-white. It should be apparent that information about the exact nature of an individual's ethnicity is lost when the latter data transformation is undertaken.
>
> Ordinal level scales can undergo any transformation that monotonically increases. Monotonicity is a scaling concept meaning that each category must be greater than or equal

to the category that preceded it (of course, if the scale is decreasing, then each category must be less than or equal to the category that preceded it). Thus, all of the transformations for the data on fire service satisfaction undertaken above are permissible since the scale is ordinal.

Interval level scales can undergo what are known as affine transformations. These are transformations that allow the origin (the zero point) and the unit of measure to change. An example of an interval level scale and an affine transformation to it is the Fahrenheit scale of temperature and the transformation of degrees Fahrenheit into degrees Celsius. This transformation changes both the unit of measure and the origin through the formula $C = 9/5F + 32$.

Ratio level scales are rare in public administration and social science, but more common in the physical sciences. A ratio level scale is actually an interval level scale with a ''true'' zero. This indicates that the unit of measure is arbitrary but the origin is not. Permissible transformations are any which change the unit of measure but preserve the origin. An example is the conversion of length from the English system (feet, yards, etc.) to the Metric system (centimeters, etc.).

The question about fire services above is one in which each individual logically would choose only one response to the question. There are, however, other types of questions that allow for more than one response in surveys; these are commonly known as multiple response variables. Consider the following questions taken from a recent survey of individuals living in communities near a wild refuge designated by the US Fish and Wildlife Service (USFWS) for reintroduction of red wolves.[7]

1. Please consider the following hypothetical situation. Suppose the USFWS was accepting donations to a ''Red Wolf Recovery Trust Fund.'' The money would be used by wildlife managers to pay for the reintroduction of the red wolf into the Alligator River National Wildlife Refuge. Would you and your household be willing to donate $1.00 every year to the trust fund?
2. If you choose *not* to donate, what are your reasons? (Check all that apply.)
    1. Do not feel the program is worthwhile.
    2. Do not support red wolf recovery.
    3. I feel the red wolf poses a threat to livestock.
    4. I feel the red wolf poses a threat to people.
    5. Some other reason. Specify.

The second question poses a problem since there is no unique single response to the question; in fact, respondents are asked to respond with as many answers as they feel are appropriate. In this situation, a single variable for the question is not appropriate but five variables for the five possible responses are. Each variable might contain a code of ''1'' if the respondent checked that response or ''0'' if he/she did not. There would thus be five columns of data—composed of ones or zeroes—in the data set for the five variables into which the data for this question were mapped.

Often, students have a problem in conceptualizing data, especially when the situation is like the one just described. Variables in the data set are *not* simply the questions in the survey; the data are *not* simply the responses to the questions. Variables are ''observable characteristics that can have more than one value.''[8] As such, variables *can be* questions in a survey (since, presumably, responses to these questions would have more than one value) but variables can also be the actual responses themselves, if we choose to code those responses as ''1'' if the respondent answered ''yes'' to the question and ''2'' if the respondent answered ''no.'' Thus, we can conceptualize this latter variable as the presence or absence of agreement with the specific response to the question. If agreement is present, a ''1'' is coded; if agreement is absent, a ''0'' is coded.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

This latter form of a dichotomous variable—one which is conceptualized as the presence or absence of an attribute, where presence is coded as ''1'' and absence is coded as ''0''—is known as a dummy variable. Dummy variables are particularly useful in social science applications for two reasons. First, any variable coded at any level of scaling can be converted into a dichotomous dummy variable. And second, by the definitions established above, dummy variables are coded at the interval level. A good example of this is religion which is typically coded from survey data at the nominal level. A simple coding of a question concerning religious preferences might resemble: ''1'' Protestant, ''2'' Catholic ''3'' Jewish ''4'' Other. There is no underlying ordinality to this scale and so the assumption must be made that the variable is measured at the nominal level. We can, however, change this variable into a four dichotomous dummy variables—Protestant, Catholic, Jewish, and Other. The new variable ''Protestant'' would have a code of ''1'' for every individual who had responded Protestant to the original religious preference question and ''0'' for all other respondents. The new variable ''Catholic'' would have a code of ''1'' for every individual who had responded Catholic to the original religious preference question and ''0'' for all other respondents, etc.

The major reason that dummy variables are useful in research applications goes beyond the scope of this chapter but suffice it to say that certain of those applications—in particular multiple regression—assume that the variables to be analyzed were measured at least at the interval level. By converting variables measured at the nominal level to dummy variables we are meeting the assumptions of the desired statistical application.[9]

IV. READING DATA

Computer programs that analyze data need to know where data are located in a data set, what the allowable codes are for the variable in question, what each of those codes signifies, how to handle a code that might be out of the range of allowable values, and what terminology to attach to each of the numerical codes and to the variable itself. So, in the above example, the computer program that would analyze the data on satisfaction with fire services would need to know that the information on the question of satisfaction with the city's fire department exists in column 67, that it occupies only one column in the data set, and that only codes of ''1'' through ''4'' are allowed in this field. The program might also be told to set all other codes—those not in the range of 1 through 4—to missing. The computer would also need to be told to attach the terminology of Very Satisfied to the code ''1,'' Somewhat Satisfied to the code ''2,'' etc. And last, the computer program would need to know that the variable under which these codes exist is entitled something like ''Satisfaction with City Fire Services.'' All of these instructions on reading data would appear as a series of statements which, depending on what type of computer one would be working with, could either be typed in one at a time, constructed from a series of menus and also run individually, or assembled into a file which could then be run as a whole against the data. Newer, interactive computing systems allow for one or the other of the first two of these ways of reading in instructions for defining data. In some personal computer versions of computerized data analysis packages such as SPSS or SAS, or in the mainframe analogues to these, the computer program allows the user to type in and run individual statements. In Windows versions of these programs, the user can highlight commands in menus and also run these commands individually. While this interactivity has great advantages in analyzing data, it is not particularly an efficient way to operate when reading in and defining data initially. For this latter operation, running a series of statements as a ''batch'' file has great advantages. Running as a batch file simply means to create a series of statements that do all that one would like to define the data and then to run this file against the data all at once. First and foremost among the

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

advantages of using a batch file to define data is the development of a history of what one has done. It is very easy to make a mistake when defining data—construct a scale that runs high to low rather than low to high; define gender as ''0'' and ''1'' when it should have been defined as ''1'' and ''2,'' etc.—and these mistakes are not easily caught if a copy of the various files used to construct the data set was not saved. And those files are far harder to retrieve and read when the user worked interactively rather than in a batch-type mode. The simplest mistakes do not often show up until the data are analyzed at which point sense needs to be made out of some counter-intuitive finding such as men reporting they have been sexually harassed in the workplace more often than women reporting such harassment. The ''explanation'' for this finding might easily be a data error that coded gender oppositely from what was intended. The actual discovery of this error would be nearly impossible without being able to look back at the files that were used to create the data set originally.

The assemblage of all of the instructions to define data is sometimes called a dictionary. In older data analysis programs such as OSIRIS, the dictionary and data file actually existed separately from each other—they were separate files that needed to be used together. One supplied the raw data; the other supplied the instructions to read the data. Each time the data were read into a computer, the instructions would be read in first and then these would tell the computer how to read the data. This is like reading the raw data in each time one wanted analyze the data. More recent data analysis programs only require a separate dictionary and data file to be read in the first time the data are read into the computer. After this, the data are typically saved as a special file that can be easily read by that data analysis package. This special file combines the dictionary and data into one file and formats that file in such a way as to optimize disk space for the computer being used. These latter files have special names—system files, internal files, etc. to designate that they are to be used only with the specific data analysis package in question. Many computer packages automatically assigned special extensions to system files so that the user can recognize them as files to be used only with that data analysis package. The major advantages to using systems or internal files to store data are ease of use, speed in reading data in and out, and minimization of disk space to store the file.

As noted above, raw data files look like large matrices formatted with variables as columns and observations or cases as rows. If there were a very large number of observations on any unit in a raw data file, we might want to have more than one row of data for each case. In the past, when computerized data were stored on cards, data sets with more than one row (or record) of data per case were very common since each record was limited to eighty columns. With modern computers, data are typically stored on CD's, diskettes, hard disks, or tapes, which can handle records with lengths up to tens of thousands of columns. Nevertheless, many analysts like to divide data sets up into shorter records since many computer monitors can only display eighty or slightly more columns at any one time. Another reason for storing data with multiple records per case might be if one had several sets of observations on each individual. For example, data on satisfaction with city services might have been collected from a sample of city residents in two separate surveys, one conducted before a major reorganization, and one after the reorganization had been accomplished. In this case, one might want to store the first set of observations as one record and the second as a separate one. The major point to be made here is that if this were done, the computer program that was to read the data would need to be told that there actually were two records per case. If not, the program would read each record as a separate case, resulting in double the number of cases and half the number of variables than were supposed to be in the data set.

Once data are stored as a systems or internal file, additional data can be added fairly easily. A wide variety of data input devices make this addition of data simple and economical; the choice of which specific hardware or software to use is typically dictated by the type of

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

project being conducted and the personal experience and preferences of the analysts involved in the project.

Scanners are one kind of device that are being used to automate the inputting of data. One kind of scanner is a screen that is programmed to accept a wide variety of data formats. For example, this kind of scanner could be used to read newspaper columns—the column is placed on the scanner, the cover is closed, and a few buttons are pushed—and to develop a data set of the actual words in the column. This technique has been used by researchers to conduct ''content analyses'' of various print journalism sources (newspapers, magazines, and so forth) in which the number of times certain words or phrases are used is counted and analyzed. A scanner might also be programmed to read other sorts of documents, such as blueprints. The blueprint thus becomes data for a data analysis program to examine. For example, a study of county courthouses might begin with the hypothesis that the design of courthouses impacts the fairness of trials—courthouses where the jury might ride the same elevator as the defendant or where the jury could see the defendant entering the courthouse from the county jail might lead the jury to develop negative impressions of the defendant before the trial begins. By analyzing blueprints of country courthouses and comparing those to decisions of juries in those counties, the role of courthouse design issues could be ascertained, and recommendations for retrofitting courthouses could be developed.

Scanners are efficient devices for inputting large amounts of data since they read entire pages of words, pictures, or images at a time. Scanners (and the software that supports the scanning process) do, however, have certain disadvantages. First, even *very* sensitive scanners make mistakes in reading data.[10] A scanner with a 95% rate of reading individual letters correctly will still most likely read five out of each one hundred letters incorrectly. Given that there are some three hundred *words* on a given typewritten page, this would lead to a fairly large number of errors per page. Scanners with 99% correct reading rates are now becoming the norm, but even here the user must expect that a number of errors will occur in the scanned page. Second, the error rate increases dramatically as the print quality in the scanned document declines, so bad photographic copies of documents present large challenges to even sophisticated scanners and scanning software. And scanners also seem to be better at scanning some kinds of print fonts than others. James Fallows, for example, reports that the scanner and accompanying software that he routinely uses to scan newspaper articles is much more successful scanning the *Wall Street Journal* than the *Washington Post*.[11]

A second type of scanner which is more sensitive and thus makes fewer errors is the more familiar bar code scanner often used to read prices in grocery or other stores. These scanners recognize the bars in the field as individual pieces of data and add them to a pre-existing data set. Of course, for such a system to work properly, bar codes must be set up and installed on each ''item'' to be later input as data and each bar code must be originally defined through a data definition statement when the data set is initially set up. While computer programs exist to write bar codes, a person must still begin the process by writing the instructions that originally set up the codes. This labor-intensive process would appear to be cost-effective only if large numbers of individual items to be added to the data set were identical to each other—in an inventory situation, for example. Where individual items are unique, each would have its own bar code and it is thus difficult to see how any cost or labor savings could be obtained.

V. RULES OF THUMB FOR CONSTRUCTING DATA SETS

Several lessons can be drawn from the discussion of reading data above. In collecting data and preparing them for input into a computerized data analysis package, there are several rules of thumb that should be followed:

1. While the computerized data analysis package into which the data are read will not need to know what level of scale has been used to measure the variable, knowledge of scaling is integral in reading and analyzing data. Thus, attention needs to be paid to how the data are scaled, coded, and input into the computer when the data are being prepared for input. This will allow for efficient data analysis after the data are ready for use.
2. All of the permissible data transformations for each of the scales noted above allow for data to be ''collapsed''; that is data can be transformed from more specific measurements to less specific measurements. It is important to note that the opposite is not true—data cannot be transformed from less specific measurements to more specific. It is possible to recode exact income data for families into larger categories, but it is not possible to recover exact income data for families from data originally coded as larger categories. For this reason, data should always be gathered using the most specific measurements possible—given practical constraints. Income is typically a difficult variable to gather data on in surveys since many people do not like to divulge their income (particularly when the questions are asked by governmental entities!). Thus, asking the question ''What is your family income?'' will typically result in a large number of respondents refusing to answer the question and another large number answering that they simply do not know. Many techniques have been developed to ''trap'' individuals into divulging their income on surveys. One asks the respondent an initial question akin to: ''Is your income above or below $20,000?'' and then proceeds to ask further questions (above or below $30,000?) to categorize income. A second technique displays a list of income categories to the respondent and asks ''Into which of these categories would you say your family income falls?'' It is important to note that both of these techniques will most likely result in larger number of respondents answering the questions (a net plus) but will sacrifice the specificity of the income measure (a net minus). Therefore, care must be taken in setting up the categories for questions like this so that the resulting data are useful for analysis. Imagine the problem of taking an income question worded like either of the two techniques above that was used in a city-wide survey and using that question unchanged for an in-depth survey of poverty neighborhoods in the same city.
3. In the design of the data set, care should be taken to make the data compatible with other data sets that have been constructed by the agency for which the data are designed and/or compatible with the demands of other agencies or data archives. While this may seem a minor point in the initial stages of a data gathering project, a small amount of investment in compatibility in early stages will save major problems in reformatting data after the data set has been constructed and analyzed. Often, state and federal agencies have very restrictive requirements on how they want data collected by local governments or private contractors to be formatted when those data are archived with the parent agency.

VI. USES OF DATA

Typically, data will prove most useful in documenting the existence of problems that public administrators are interested in. But data also have other important uses for the public administrator. Typically, policy analysis texts teach a six step method of analyzing public problems—problem identification and validation, development of evaluation criteria, development of alternative solutions, analysis of solutions in the light of evaluation criteria, display of results, and

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

recommendations.[12] Data on the existence of a given problem would be necessary in determining how extensive a problem exists and whether the problem was worsening. Thus data are necessary in validating and identifying a given problem, but data would also be extremely useful in each of the additional five policy analysis steps. The application of this idea can be best pursued through an example.

> As a manager in a large city's public works department, you receive a memo from your supervisor directing you to look into the situation concerning potholes in city streets. Residents of the city are complaining of the existence of a larger-than-normal number of potholes which seem to have appeared after the particularly harsh winter your city experienced. What kinds of data on potholes would you begin looking for? How would you begin gathering these data? What would you do with the data?

First, note that the question of potholes is deceptively simple—deceptive since it more than likely is not the mere existence of potholes that city residents are complaining about. Most problems have both an objective and a subjective component,[13] and the existence of potholes is no exception to this rule. While the city may or may not actually have more potholes this spring than it has experienced before, city residents might be perceiving a larger number of potholes since the winter was so harsh and the local newspaper just ran a series of stories on pothole problems in neighboring states. Thus, the kinds of data to be gathered initially divide into objective data on the actual existence of potholes in city streets and subjective data on the perceptions of city residents about the existence of potholes. In pursuing the first type of data, one would most likely try to establish a sampling frame of city streets—it would probably be impossible to count all the potholes in all the streets in the city—and then to count, as inexpensively as possible, the number of potholes on those streets. In pursuing the second type of data, one would most likely want to conduct a survey of city residents by taking a sample of the population and asking them questions about the existence of potholes this year in comparison to past years and so on. These two kinds of data would provide a valuable base line in looking at this problem. If the department had collected similar data in the past, comparisons between the observations this year and those from past years could be made to identify whether the problem actually was worse, or whether there was more of a perceptual (subjective) problem. In deciding how to address the pothole problem, very different potential solutions would suggest themselves depending on whether the problem was documented as more of an objective or more of a subjective problem. But the need for data to examine this problem has only begun. While data are certainly necessary to document the problem, to determine if it is getting worse, to compare this city with other cities, etc., data would also be necessary in the later stages of the policy analysis process. For example, a survey of experts in the field might be desirable in order to determine what kinds of evaluation criteria they would advocate in analyzing the policy problem. The survey of experts itself represents a data collection task and would have its own analysis—think, for example, of what statistic might be used to summarize the consensus of opinions of a group of experts. Would a Delphi technique, in which a survey was conducted and analyzed and the results then summarized and made available to the experts for comments and modification of their opinions be useful? In determining how actually to analyze the alternatives that were isolated in this policy problem against the evaluation criteria chosen, policy texts often advocate weighting the criteria such that some are more important than others in the final mathematical analyses. How would one determine how to assign these weights? The naive policy analyst might do this through intuition—a method which is virtually scientifically indefensible. A much better way would be, again, to survey experts in the field, and collect data on their opinions on how important the various criteria are. Weights can be easily calculated from the data collected from the policy experts.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

VII. WHERE DATA COME FROM

Data can come from a wide variety of sources. These sources can be divided initially into those where the data are collected specifically for the individual project and those where the data have been collected by somebody else but are being used for the project currently underway. Collecting one's own data to address problems in policy analysis and public administration has great advantages. Often, the types of problems that public administrators are interested in are limited—bounded both geographically and temporally—resulting in needs for data that are specialized for one or a small number of situations.[14]

In the pothole scenario developed above, it would have been difficult for the analyst to proceed in any other way than to gather the data for him or herself. But this method of collecting specialized data has at least two major disadvantages. First, collecting one's own data has costs in terms of money, time, and effort. Mounting a survey of potholes in a large city can take a few weeks to accomplish, would need personnel assigned to conduct the survey, and would require the efforts of somebody to design the sample and conduct at least minimal training sessions with those who would conduct the survey. And then there are the necessary tasks of processing the survey data into computer-readable files, and so forth, all of which will also add to the cost, time, and effort of answering the question for which the data were collected. Second, the mind set that is developed by collecting specialized data to address one unique problem is one that is self-perpetuating. As long as public administrators operate in this fashion, there will not be large data archives from which other public administrators can profit.

This last point is worth pursuing in some detail. To conduct a survey of city services in a large US city, one would want to collect data on such things as whether city residents think they are getting a good deal on the taxes they pay, on which departments they think are doing a good job and which need improvement. Many cities conduct such surveys on a regular basis; some such as Birmingham, Alabama have acquired a reputation as being particularly activist in conducting citizen surveys.[15]

For the purpose of answering the particular question concerning a specific city's needs, collecting particularized data might be exactly what one would want to do. In order to do so, a sample of city residents would need to be isolated through a scientifically defensible sampling scheme, questions that tap the relevant dimensions on which data were to be collected would need to be developed, the interview instrument would need to be tested for such things as length of interview, proper interpretation of questions, etc., and then the survey would need to be conducted. It might be discovered that 55% of those surveyed thought that taxes in this city were too high and that the sanitation and police departments were specifically identified as not performing their functions very effectively. After the report was submitted to the city manager and the city council, questions most likely would arise as to how this city compared to other cities in the state or other cities of similar population in other states. Given the individualized nature of the data project described, this question could not be answered with any great confidence and a second, comparative data gathering project might be necessary to come up with the council's questions on this topic.

Imagine a different scenario for this project. After being assigned the survey project, the analyst calls a data archive and asks the librarian (or better yet, accesses their world wide web page and personally looks for relevant studies) for information on the types of surveys that other cities have recently conducted. After looking at the half-dozen or so studies, the analyst isolates questions using those in the previous studies as models. This eliminates the step of pre-testing the questionnaire for format (although, most likely, the questionnaire would still need to be pretested for time). The cost for accessing the data archive would, most likely, be small (or covered by a subscription that would allow unlimited access for a given length of time) but in return

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

for access to the archive, one would be required to archive the data generated in the current study with the facility. The efficiencies in the latter mode of operation are clear—many hours of developing wording for questions to make them unbiased and clear in meaning become unnecessary. The analyst would know that the questions that he or she is using have been utilized in other studies and so have a degree of support in the academic community or among experts in survey research. This would greatly lower the overhead in conducting surveys for public administration and public policy uses. And data generated in other geographical settings become available so that any localized study can be compared with other localities or (possibly) national samples to validate the findings one might obtain. In addition, since the data were generated using the same question wordings, possible challenges stemming from differing wordings or formats become moot.

Such data archives are now either in existence or in the planning stages at various universities and other facilities around the US and in other nations. Several large survey-oriented archives have been available for perusal for free or for nominal costs for some time and the developers of these archives have recently put them on-line so that they can be accessed through the World Wide Web section of the Internet. Some large, national surveys are archived at their own survey house's website, but virtually all of the large national surveys are archived either at the ICPSR website described below or at one of a number of other websites mainly associated with universities.[16]

The Inter-university Consortium for Political and Social Research (ICPSR) is a membership based organization that archives and disseminates questionnaires and data. While the typical member of ICPSR is a university, local, state, and national government agencies also are members. Membership carries with it free access to the archive, along with technical assistance in accessing and utilizing the data stored there. ICPSR is housed at the University of Michigan; its governing board is composed of researchers from universities in the US and abroad. ICPSR not only houses a great deal of survey data from such diverse sources as the American National Election Study, ABC, CBS, NBC, The *Wall Street Journal*, *The New York Times*, CNN, etc., but also houses data from the Department of Justice, The Department of Health and Human Services, and the Census Bureau.

The National Opinion Research Center (NORC) at the University of Chicago has been conducting the General Social Survey (GSS) for many years and has developed an archive of questions and data from that series of national surveys. The GSS asks questions from a number of disciplines—sociology, economics, psychology, child development, political science—and so can provide a wealth of information for those looking for possible question wordings or for base line data on which to compare their own jurisdictions. The GSS is also archived at ICPSR and the archive there not only allows for retrieval of question wordings and data, but allows users to obtain marginals (frequency distributions) of questions and simple cross-tabulations over a computer link.

A new ICPSR service (GSSDIRS) allows a user to search all the archived GSS studies by keyword. A recent search on the keywords ''local government'' resulted in a large number of ''hits,'' the following being an example:

> If you had some complaint about a local government activity and took that complaint to a member of the local government council, would you expect him or her to pay a lot of attention to what you say, some attention, very little attention, or none at all?

A set of numbers representing the marginal distribution of answers to this question from GSS surveys dating from 1972 to 1994 is then presented, allowing the user to search for trends in the data, etc.

GSSDIRS is a powerful search engine that not only allows the user to do sophisticated

key word searching through a large amount of GSS data, but it also contains links between questions. So, a user who accessed question number 340, above, could then move to related questions by using the buttons programmed on the computer screen.

In the past several years, several researchers have suggested that state and local governments—who conduct large numbers of surveys typically monitoring citizens' opinions of the services governments offer—begin to develop a common standard set of questions and data formats in order to lower the overhead in the conduct of such surveys. With a common set of questions and formats, a state and local government data archive could be more easily (and less costly) established. While such a data archive has been proposed by several different individuals in several different venues, it is still only in the planning stages.

As noted above, survey data are not the only kinds that have been archived. The largest data archive in the US is that maintained by the US Census Bureau. The uses of Census data are numerous. Granting agencies typically require extensive documentation of a problem in order for an applicant to be considered for financial support. Census data are adaptable for such uses, although the actual steps that one must go through to make the data amenable for use in a grant application are not as simple as one might think. Consider the following example:

> A city manager has been receiving a large number of complaints from residents of two neighborhoods that abut the central business district of a city of 200,000. The residents are complaining that their neighborhoods are deteriorating, or at least are not keeping up with other neighborhoods it the city. The manager wants to be able to address the residents' complaints and to seek financial assistance from the Federal Department of Housing and Urban Development (HUD) if the data show that there are substantial, addressable, problems in these neighborhoods.

The manager turns the project over to a senior analyst and asks for the analyst's staff to collect data to address the economic, social, and physical state of these neighborhoods and the change in all three of these over time. The analyst, in receiving the manager's memo, knows that the necessary data are available through the Census Bureau. The analyst also knows that Census data are available from on-line sources directly from the US Census Bureau, from data archives such as ICPSR, etc., or in paper format through the extensive series of volumes that the Census Bureau publishes (and that are housed at the local university library).

The analyst's first steps in gathering the data would most likely be to think of what possible indicators of deterioration might be utilized in this situation. He/she would know that multiple indicators are better than single ones, and would also know that data need to be gathered to determine if economic, social, and/or physical deterioration were occurring. The analyst develops an extensive list of indicators to research and begins to look at Census holdings to see: (1) if data on these indicators are available; (2) if these data were gathered over time; and (3) to see if the data on indicators that were gathered over time were defined in similar enough ways as to make over time comparisons meaningful.[17]

The analyst assigns the best graduate student intern working in the department to gather the data. At this point a problem occurs. The graduate student knows that there will not be data on all of the indicators that the analyst is looking for; the graduate student also knows that on only some of these indicators will there be data from more than one Census. But those are only minor problems; the major problems that hit are, first, that the city's definition of the neighborhoods under question and the Census Bureau's definition of a Census tract are not conterminous. And second, the Census Bureau constantly updates its definition of tracts and blocks which makes over-time work difficult. Consider these two points separately.

First, definitions of neighborhoods are difficult to develop. Residents of a given neighborhood subjectively know what geographic areas they consider to be in their neighborhood but

these subjective definitions are often not shared with a large number of their neighbors, and the subjective definitions also often do not translate well onto city maps of the neighborhood. In addition to this, whatever subjective definition that can be developed for a given neighborhood will most likely *not* be exactly described in the most available US Census data—the data sets that describe US Census tracts. Census tracts are developed by the Census Bureau in conjunction with local government officials and researchers at local universities, and data from Census tracts are the basic building block on which most Census analyses are based but, at best, a given Census tract is only a statistical approximation of a neighborhood. Data on smaller units—Census blocks, for example—are also available through a variety of sources, and these data can be used to develop a finer approximation of a neighborhood, but these data are difficult to work with since one would need to know exactly what Census blocks are contained in a given neighborhood and then to build up the neighborhood data set from its component Census block data. While this is not an inherently difficult task—and the actual process will be described later in this chapter—it is not one that an analyst would typically undertake for a single study. Several cities—Cincinnati is a notable example—have developed statistical approximations of neighborhoods from Census data and make these data available to both city departments and outsiders for research purposes, thus reducing the need to develop individualized data sets for particular projects.

Second, the Census Bureau is not consistent over time in its definition of a given Census tract.[18] Census tracts are adjusted for each census as neighborhoods, cities, and states gain or lose population from one US Census to the next. These adjustments are attempted with a bias towards preserving the status quo, but often the changes that are made are large. Again, it is possible to re-compute the data for a given Census tract from its underlying block structure and thus to make data from more recent censuses conform to those of the past, but the further back in history one wishes to go, the more changes will be necessary and the more time, effort, and money will have to go into developing the data set necessary to answer the initial question. At times this may be absolutely necessary. For example, if a researcher interested in population growth in cities wanted to disentangle ''natural'' growth (growth as a result of childbirths and in-migration) from growth as a result of annexation,[19] the researcher could attempt to match Census tracts from the original city (pre-annexation) with current Census tracts that approximated those and measure population only in these tracts.

At best, then, Census data can only yield an approximation of the data necessary to answer the question about neighborhood decline. Depending on how close the Census tract definitions for the city in question are to the underlying neighborhood structure of that city, the statistical approximation would be either close or distant. Given an unlimited amount of time, resources, and money, the Census data could be recalculated to yield an exact match to the city's neighborhoods, but for most public administration research tasks, time, resources, and money are almost the exact opposite of unlimited on all three dimensions.

The most important point to be drawn from this discussion is that while Census data are potentially an extremely useful source of data for state and local public administrators, the structure of the Census data make them difficult to use for certain purposes. The US Census Bureau is extremely conscious of the confidentiality of its data; they simply refuse to make data available for small enough units that information about individual citizens can be inferred from the larger unit of analysis. If the Census Bureau errs in reporting data, it errs on the side of caution; data drawn from small units that researchers would be interested in are simply not released to respect individual privacy. Census tracts contain thousands or tens of thousands of individuals while Census blocks contain hundreds or thousands of individuals. In such large amalgamations, individual privacy is protected.

Census data are available from a large number of sources. Many private companies design

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

and market computer software that allow easy access to Census data, thus allowing the user maximum flexibility in how those data might be used. However, purchasing a computer program from a private company is not necessary since the Census Bureau maintains its own World Wide Website at http://www.census.gov. This website provides a number of services to the user, including direct access to Census data (numbers) and maps. Data for counties, municipalities, Census tracts, and Census blocks can be easily retrieved and manipulated. Census maps are available through the Bureau's TIGER (Topographically Integrated Geographic Encoding and Referencing) system. TIGER allows the user to specify maps of whatever location he or she desires down to the level of Census blocks. Customized maps can be assembled using geographical features such as roads and streams as boundaries. For those without World Wide Web access, most municipal and university libraries in the US have Census data available on CD-ROM which allows the user to develop statistical profiles and/or maps of whatever areas the user might desire.

VIII. MANIPULATING DATA

Once a data set has been constructed, the data in it, and the very structure of the data set itself, can be manipulated to fit the needs of the analyst. Since the manipulation of individual variables and the manipulation of the data set itself are quite different in concept, these are examined individually below.

IX. SCALE OR INDEX CONSTRUCTION

Individual variables can be recoded, as described above, by reversing scales, collapsing data from many categories into fewer, and so on. These manipulations are fairly simple; the necessity for undertaking such manipulations should be self-evident. More complicated manipulations involve such things as scale or index construction, in which several individual variables are combined into a scale that summarizes the components.

Many scales are constructed as simple additive indices in which the responses to underlying questions are simply summed into an aggregate score. For example, a researcher on economic development might have isolated a number of mechanisms that economic development practitioners could possibly use in their quests to develop their communities. These mechanisms might have been sent out in questionnaire format to local government economic development practitioners throughout the state and the respondents asked to check off which mechanisms they had used to attempt to develop their communities. The resulting questionnaires could then be coded ''0'' where the respondent had not attempted one of the mechanisms, and ''1'' where the respondent had. By simply summing across all the questions, the researcher could construct an additive index of economic development activity in a given community—the higher the number, the greater degree of activity the community would be presumed to have undertaken.[20]

But this scale would be less informative than one which was theory driven. A different approach to constructing the same scale would be to start with some implicit or explicit theory of economic development and use that theory to assist in constructing whatever scales one might need. For example, one might want to divide all of the mechanisms identified in the survey describe above into those that are ''supply'' driven and those that are ''demand'' driven and to construct scales of each of these by only summing the answers to questions under each category.

A more sophisticated scaling technique that has enjoyed some popularity in social science research is to use a data reduction technique such as factor analysis to organize one's data and then to construct scales using the factors identified by the computer program. For example, James Perry and William Berkes used factor analysis to examine strike behavior by local government employees. Their work isolated two factors—local employee status and past strike behavior—from a list of seven variables. The two factors were then used to predict the likelihood of strikes in other localities.[21] Factor analysis programs typically yield two sets of coefficients for each variable—a score of how highly each variable ''loads'' on a given factor and a set of factor scores. The first of these is a measure of how highly correlated each variable is with the latent variable that is being measured by the factor and thus provides useful information about the factor structure in the data. The second coefficient—the factor score—can be used to create a scale in which each of the components are not equivalently weighted. When creating the scale, the scores for each of the individual variables are multiplied by the factor score and then summed, resulting in a scale in which some of the variables contribute more toward the final index value more than do others. Some analysts find this weighted scale to be a more accurate measure of the underlying latent variable than a simply summing of the variables loading highly on a given factor.[22]

A second way of combining individual variables into a composite index or scale is to standardize each of the variables. This can be done in a variety of ways but one common one is converting the data to *z*-scores and then to sum the *z*-scores. This procedure is often used when the individual variables are measured on vastly different metrics or when one wishes to create a summary scale from a series of sub-scales of varying size. For example, in my own work on citizen participation,[23] four sub-scales measuring the types of mechanisms cities used to foster citizen participation—information gathering mechanisms, open government mechanisms, neighborhood empowerment mechanisms, and citizen coproduction mechanisms—were standardized and summed to create a holistic citizen participation scale. The standardization procedure was necessary since each of the four sub-scales varied in how it was measured—citizen coproduction ranged from zero to nine while information gathering ranged from zero to four.[24] Without standardization, the scales with more categories (coproduction) would determine the outcome of the overall scale more heavily than would the scales with fewer categories (neighborhood empowerment).

A second use of this procedure would be where the composite variables are measuring vastly different items. One example given by Meier and Brudney constructs a measure of performance for a city's garbage collection crews on the basis of two variables—tons of trash collected and number of complaints phoned in by residents. As can be seen, it would be impossible to sum tons of trash and numbers of complaints to arrive at a useful scale since the two items are measured on such different scales (metrics). The solution to this problem is to convert each of the two variables to its *z*-score analogue by subtracting the mean from the raw score and dividing by the standard deviation. The resulting standardized scores are by definition measured on a common metric and thus the scores can be summed to arrive at a usable summary index.[25]

Thus it can be seen that scale or index construction is data manipulation since new data are being created from old data. In performing such manipulations, one important point must be stressed—once the scales or indices have been constructed and the analyst is satisfied with how they are performing, the new variable or variables must be saved—added to the data set—before one exits the computer program. Not saving the newly created variables is one of the most common errors in computerized data analysis, and committing this error means that all of the steps undertaken to construct the index or scale must be gone through again. Since some scale or index construction is quite extensive and complicated, it is easy to see why the analyst would not want to repeat the procedure unnecessarily.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

X. CREATING NEW DATA SETS FROM EXISTING ONES

At times, a data analyst will want to re-cast an entire data set by changing the unit of analysis, in essence creating a new data set. For example, in an early work looking at constituency influence on members of Congress, Warren Miller and Donald Stokes[26] aggregated the 1958 American National Election Study by Congressional District. They calculated a statistic (the mean) for responses on policy questions for individuals in Congressional Districts and compared those scores with members of Congress from those same districts. A data set where the individual was the unit of analysis was transposed into one where the Congressional District (and its member of Congress) was the unit of analysis. This technique can be used fruitfully in a variety of situations which are of interest to those involved in public administration. Several of these might be: the aggregation of a survey of housing for a municipality into the municipality's composite neighborhoods in order to compare housing needs across neighborhoods; the aggregation of samples of individual bureaucratic encounters with the public into agency statistics to compare performance across public agencies; or the aggregation of individual electoral support for an education-related bond referendum into neighborhoods to ascertain comparative support for such issues.

In each of these and other situations, aggregating the data up to a higher level of analysis is a fairly easy task—of course, disaggregating data into lower levels of analysis is virtually impossible for many of the same reasons identified under the section on coding variables above.

In order to allow for aggregation, several important points should be stressed when gathering data:

First, in order for any aggregation to occur, the variable that one would want to aggregate on must be present in the data set. While this sounds trivial, it is extremely important. If one wants to aggregate individual level data up to the Congressional District, a variable containing information about Congressional Districts must be present in the data set. Only then can the computer be instructed to aggregate the cases under each of the Congressional Districts coded.

Second, a statistic must be chosen to summarize the individual data at the group or aggregate level. In the Miller and Stokes example above, the mean was used to summarize the individual policy preference data at the Congressional District level. The median could also have been easily used. In fact, if one considers the sum to be a statistic, all of the discussion above on Census tracts can be thought of as creating a new Census tract data set from its constituent Census block structure. It should be apparent that the choice of a statistic is more than individual whim and should be driven by well-supported statistical theory and practices.

Third, it is important to keep in mind the nature of sample data and what happens to the representativeness of samples when one changes the unit of analysis. The American National Election Study used by Miller and Stokes in the example above is a representative national sample of the American electorate, but it is *not* a representative sample of any individual Congressional District in the US. By aggregating individual responses by Congressional District, calculating the mean across those individuals, and using this statistic as an indicator of district opinion on policy issues, Miller and Stokes were probably pushing their data too far. There were *very* few cases in several Congressional Districts, making the samples in those districts far from representative.[27] In statistical terms, the small samples in some Congressional Districts would result in *very large* standard errors, casting doubt on the validity of statistics calculated for those districts.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

All popular computerized data analysis programs such as SAS or SPSS allow data to be aggregated in the ways described above. The actual command structure to accomplish the data transformations is fairly simply but does require a certain amount of experience with the program in order to work the way the user would like. To accomplish such transformations efficiently, it is usually necessary to consult somebody experienced in the data analysis program in which the work is being performed.

XI. CONCLUSION

Data are available to researchers and practitioners in the field of public administration from a wide variety of sources. Whether one wishes to gather data specifically to address a certain problem or to use archived data gathered by somebody else, there is enough data to satisfy even the most ardent seeker. The important questions concerning data are those involving applicability or use, preservation, and sensitivity.

All data gathering exercises must be guided by some theory of the applicability or use to which the data will be put. Simply gathering data on subjects in which the researcher or practitioner is interested is, at best, an inefficient use of resources. All data gathering has a cost—whether that cost be monetary, human, administrative, or whatever. And thus, all data gathering should be driven by at least implicit cost/benefit considerations. Is a new study necessary? Are there already published, relevant data that are available to us at no or moderate cost? Will those data suffice for the purposes of the current project? All of these are questions that should drive the data gathering task. Often, public administrators who are new to the field will propose a study (a data gathering task) to address a problem their agency is facing. While new studies might be necessary to address such problems, it is often forgotten that there is a wealth of data available for a wide variety of sources which might already address the problem.

When quantitative data analysis first began to be taught in the social sciences in the 1940s, some strict practitioners in the field taught their students to gather data for a specific task, calculate the statistics they wanted from the data, check the statistics against the pre-existing hypotheses they had developed, and then destroy the data set so that they would not be tempted to go back and recalculate statistics or do something else that might compromise the practitioner's strict view of the scientific research process. In such an environment, archiving data was not anything that anybody cared about or advocated. In more recent times, considerations about what to do with the data after they have been used for the specific purpose intended carry much more heavy weights. Public agencies throughout the US and in other countries routinely keep files—computerized or paper—of the results of past studies so that any new work in the field can be compared to past results and to reduce the overhead of the research task by keeping researchers from recreating the wheel. Whether these archives are public or not is unimportant; what is important is that they are accessible by people who need to use them—researchers in the agency who are conducting current research, for whom past research would make their jobs easier. Thus, questions about the format of data sets, the nature and format of questions asked in surveys, etc. become important when one looks at data gathering with archiving as a goal.

Last, data are sensitive simply by their nature. Surveys, for example, sometimes ask sensitive questions of respondents, and in order to get the respondents to answer these questions frankly, confidentiality is promised. These promises of confidentiality often lead to different—at times ''better''—data than if confidentiality was not promised. For example, victimization studies—studies where respondents are asked if they have been a victim of a crime, what the circumstances were—often proceed from a random digit dialing (RDD) sampling scheme. When

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

RDD is used, the interviewer does not know to whom he or she is speaking and this is often communicated in the course of the interview. This form of confidentiality leads to *much* higher estimates of the rate of certain sensitive crimes—especially rape and sexual assault—than do analyses of crimes reported to the police. Society's understanding of crime has changed dramatically since victimization studies have been used to augment reported crime as a source of data on crime. But what would happen to these gains in the understanding of crime if it were suspected that the researchers were *not* treating the data as confidential? It would not take much to trace the phone numbers obtained through a RDD sampling scheme and obtain the names, addresses, etc. of the individuals who had been interviewed, and a nefarious survey organization could make good use of this information—perhaps attempting to sell crime deterrents to people who had been victimized. Or what would happen if a respondent found that the results of a study in which she had participated were reported in such detail that anybody with any passing familiarity with her case could identify her? What keeps this from happening to any extent is a common understanding among the research community about the confidentiality of data. This confidentiality is what allows survey data to be taken seriously; if individuals suspected that the data they were supplying governments, survey houses, corporations, etc. were going to be used in ways that were less than confidential, our faith in survey data would be seriously undermined. Thus, any data gathering task that involves gathering data from individuals must guarantee the individuals' confidentiality. Failure to do so is courting disaster, both for the individual study and for the research community as a whole.

NOTES

1. C. H. Coombs (1964). *A Theory of Data*, John Wiley and Sons, p. 1.
2. It should be apparent that what we are discussing here is specifically what behavioral researchers call *R* methodology. *R* methodology involves seeking commonalities among variables across individual units. A different methodology—*Q* methodology—has been identified and advocated by W. Stephenson most clearly in his book *The Study of Behavior: Q-Technique and its Methodology* (Chicago: University of Chicago Press, 1953). *Q* methodology seeks to represent an individual's subjective view of reality; it has very different data needs and quite different data gathering tools than those described here. See S. Brown, *Political Subjectivity: Applications of Q Methodology in Political Science* (New Haven: Yale University Press, 1980) for a good explication of *Q* methodology and how it differs from *R* methodology. Also, see F. Kerlinger, *Foundations of Behavioral Research*, 2nd edition (New York: Holt, Rhinehart, and Winston, 1973), chapter 34, pp. 582–600.
3. In most computer applications commas would not have to be entered into the data. The computer program would do this automatically or could be told where to insert the commas.
4. W. S. Sarle (1995). Measurement theory: Frequently asked questions, unpublished manuscript to appear in *Disseminations of the International Statistical Applications Institute*, 4th edition.
5. S. S. Stevens (1946). On the theory of scales of measurement, *Science, 103*: 677–680.
6. Sarle, 1995.
7. See W. R. Mangun, N. Lucas, J. C. Whitehead, and J. C. Mangun (1996). ''Valuing Red Wolf Recovery Efforts at Alligator River NWR: Measuring Citizen Support,''

in *Wolves of America Conference Proceedings*, Washington, DC: Defenders of Wildlife, pp. 165–171.

8. E. O'Sullivan and G. Rassel (1995). *Research Methods for Public Administrators*, 2nd edition, Longman Publishers, p. 13.
9. There has been a certain amount of controversy about using variables measured at the ordinal level in multiple regression applications. While statistical purists still insist that only interval or ratio level variables should be used in multiple regression applications, most recent texts have adopted the rule that the regression model is robust enough to support analysis of ordinal level scales provided that there are a fairly large number of categories in the scale (typically, five or more) and that the analyst is willing to assume that the differences between the scale categories are approximately equal. L. Giventer, for example, in *Statistical Analysis for Public Administration* (Belmont, CA: Wadsworth, 1996, p. 404) states flatly, ''A variable having an ordinal level of measurement with an implicit underlying scale can be treated mathematically as if it has an interval level of measurement.''
10. *The Atlantic Monthly* has been gathering humorous incidences of scanning errors in the preparation of their own magazine. Their Word Wide Web page contains a list of scanning errors in recent editions of the magazine.
11. J. Fallows (1994). Recognizable Characters, *The Atlantic Monthly*, February, pp. 110–115.
12. See C. Patton and D. Sawicki, *Basic Methods of Policy Analysis and Planning*, 2nd edition (Englewood Cliffs: Prentice-Hall, 1996).
13. See C. Scavo, Racial integration of local government leadership in southern small cities: consequences for equity relevance and political relevance, *Social Science Quarterly, 71*(2): 362–372.
14. W. Dunn, *Public Policy Analysis: An Introduction*, 2nd edition (Englewood Cliffs, Prentice-Hall, 1994).
15. See C. Scavo, The use of participative mechanisms by large US cities, *Journal of Urban Affairs, 15*(1): 93–110.
16. The Institute for Policy Research at the University of Cincinnati, for example, provides a data archive consisting mainly of data sets obtained from ICPSR, NORC, the International Survey Library Association (Roper studies), and the US Census Bureau. Other universities—Yale, the University of Wisconsin, several of the University of California campuses, and others—maintain data archives focusing on various aspects of public administration, public policy, or other social science-related areas.
17. The Census Bureau did not, for example, publish data on female-headed families until 1990 when interest in these families among researchers and others warranted such publication. Attempting to develop tracking data in this case is thus futile, since the data literally do not exist before 1990.
18. To be fair to the Census Bureau, they *do* work extensively with state and local governments and researchers from area universities in making adjustments to Census tracts so that the changes from one Census to the next are more-or-less incremental in nature. There are, however, instances in which incremental change is impossible and only large-scale change in the Census tract structure for a geographical area can accommodate the population changes that have occurred.
19. Since annexation rates vary tremendously by state and region, comparisons of city population growth rates that seek to make assumptions about natural population in-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

creases would be required to make the types of statistical approximations described here.

20. See P. Eisenger *The Rise of the Entrepreneurial State: State and Local Economic Development in the United States* (Madison, WI: University of Wisconsin Press, 1988), R. Hanson, and M. Berkman, A meteorology of state economic climates, *Economic Development Quarterly*, 1991, 5 #3: pp. 213–228, and L. Reese, Categories of local economic development techniques: an empirical analysis *Policy Studies Journal*, 1993, 21, #3: pp. 492–506 on the use of additive indices in economic development research.
21. See J. Perry and W. Berkes, Predicting strike behavior, *Western Political Quarterly*, Spring, 1979, pp. 501–517.
22. See E. O'Sullivan and G. Rassel, 1995, pp. 279–280 for examples of this approach.
23. See C. Scavo, The use of participative mechanisms by large US cities.
24. The four sub-scales described were first detailed in E. Sharp, *Urban Politics and Administration: From Service Delivery to Economic Development*. (New York: Longman, 1990).
25. K. Meier and J. Brudney, *Applied Statistics for Public Administration*, 3rd edition (Belmont, CA: Wadsworth, 1993), p. 113–115.
26. See W. Miller and D. Stokes, Constituency influence in Congress, in *Elections and the Political Order*, A. Campbell, W. Miller, P. Converse, and D. Stokes (New York: John Wiley and Sons, 1966), pp. 351–372.
27. Miller and Stokes recognize this problem and attempt to address it. See p. 353, note #3.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

9
Threats to Validity of Research Designs

Nicholas A. Giannatasio
University of North Dakota, Grand Forks, North Dakota

I. INTRODUCTION

In the framework of everyday conversation there seems to be little distinction made between the terms *reliability* and *validity*. When we discuss reliability we are describing a quality of something or someone that is ''dependable'' or ''trustworthy.'' Validity has some same connotations as reliability. When one tries to conceptualize something as valid, we often conform this term with similar sounding synonyms as those used for reliability and possibly include: ''sound,'' ''telling,'' or ''cogent.'' Yet, most people would not make the distinction between a scale that measures weight as being reliable or valid. While we would accept either reliable or valid in this context, validity implies much more than reliability. Validity implies logic and well-grounded principles of evidence; and, if one were to place reliability and validity on a continuum, they would occupy opposite poles. In research, such is the case. Researchers want their measurements to be reliable, but often, as in some case studies, reliability cannot be assured to the degree the researcher feels is warranted. On the other hand validity, must be assured. This chapter attempts to clarify the distinction between reliability and validity in research design. If one understands validity and is able to conceptualize its distinction from reliability, the research design will be stronger, more sound, and ultimately more convincing.

The title of this chapter may appear intimidating. A neophyte to data analysis and research design may look at the title of this chapter and probably put as much distance between themselves and this topic as one could. It has a confusing yet important sounding ring to it that can illicit responses like: ''is this something I have to know about?'' The answer to that question is simply—yes. However, it is hoped that this chapter will take the mystery out of this title so that it may be known as ''Our Experiments, Things that can go Wrong with our Experiments, and How to Avoid Them on all Levels of Research.''

This topic also presents logistical considerations of ''which comes first, the chicken or the egg?'' and, from what context, framework, or paradigm does one look at the chicken and the egg? Does one come up with an experimental design and then look for what would threaten the validity of the design? Or, should one be aware of threats from internal and external issues before the research design is developed? In both cases the answer is simply—yes. Therefore, whether we start with explaining threats to validity or the components of a research design, both topics—validity and design—are prominent in their importance to quantitative and qualitative methods and data analysis. Notwithstanding, the equal footing of validity and design, this chapter

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

will discuss validity as a prologue to a discussion of research design, and place both validity and research design in the framework of positivism.

II. POSITIVISM

There is much debate in the social sciences about positivism. Auguste Comte, the French philosopher and the founder of the positivist school, adapted his theory as one that excluded speculation and was based on positive facts and phenomena. Positivism is a valuable reference point for all schools of thought because it is the most dominant framework of rational, comprehensive, and empirical experimental designs that are the closest the social sciences come to the ''hard'' sciences. Threats to validity of research designs, the topic of this chapter, communicates positivism. Positivism looks to the past for causality in order to advise the decision maker on future considerations. Simply put, if a city manager needed to make a decision about whether a new fire station needed to be placed in a growing area of town, the manager would most likely look at historical facts, such as: the number of fire alarms answered by existing fire stations in the proposed district, response time to those fires, multiple alarm fires in those districts that may have been caused by a slow response time allowing the fire to escalate, the cost of present fire stations and their predicted impact on the tax burden of new fire stations. These are positivistic facts that often influence the decision process. The debate begins to rage when detractors of positivism affirm that positivists only consider the descriptive side—the facts of the issue and ignore the value side—the normative issues that may raise questions of whether an additional fire station may save someone's life. Indeed, scholars such as Durning (1993), Denhardt (1993), Bobrow and Dryzek (1987), and Kaplan (1963) feel that positivism provides little help in determining public policy and most likely is the root of the problem in acquiring the knowledge necessary for decision and policy making. Furthermore, positivism implies an all or nothing type of approach to the future of policy actions, i.e. $X_{1...n}$ causes Y. Therefore, the decision must include all factors of X. The problem with this aspect of positivism is that it may contain internal contradictions that can paralyze the practical realization of the analysis (Bobrow and Dryzek, 1987). These contradictions include *self-negation*, described by Kaplan (1993) as self-defeating, in that general laws as prescribed by positivists will at some time in the future, be negated by other laws. An example of this self-negation is how Darwinian evolution negates religious fundamental beliefs in creation. A further contradiction is that the positivistic world view is one of cause and effect and this determinism is too insulated.

Decision makers and policy scientists realized that the parochial approach of positivism had to be adjusted. However, there was hardly a total, realistic intention to ''throw the baby out with the bath water.''[1] Rather, positivism became a tool, one of many others, to be used as appropriate. Popper in the 1930s realized that some aspects of the positivistic approach were necessary in what he termed ''piecemeal social engineering,'' where an all or nothing approach is not needed but rather a piecemeal, moderate, cautious intervention (Bobrow and Dryzek, 1987). Lindblom and Cohen (1979) described a path to what they described as ''usable knowledge'' that included scientific (positivism) *and* ordinary knowledge (common sense, causal intuitiveness, etc.) Hermeneutics, Forensic Policy Analysis, and Pragmatism, use positivistic approaches to weave their narrative case. Fischer (1995) describes a discursive method of policy analysis where a consensus must be reached on each level. In Fischer's model, the first level includes positivistic verification before proceeding with a higher level of discourse.

The point of the newer approaches to analysis is not that positivism is dead, nor is it the ultimate tool in the social sciences, but it remains a prominent, viable tool, part of a total ''tool box'' of analytical tools where verification of programs need an empirical interpretation as part of the argument of analysis.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

III. DEFINITION OF TERMS

The following three basic definitions are the beginning of the discussion not the end; nevertheless, they are the point of reference for this chapter's discussion:

Validity—simply put, are we comparing ''apples to apples?'' A measure is valid if it really measures what it is supposed to measure. For example:

A valid measure of reading scores would be one where those with high reading scores scored high and those with low reading scores scored low.

Threats to Validity—would be those internal and external factors that may prevent one from measuring what one wants to measure or obscure the relationship between the dependent and the independent variables.

For example: The Hawthone effect, or Testing effect (Campbell and Stanley, 1963), if not controlled, would affect the results of scores.

Furthermore, respondents, realizing they are being tested, may give responses based on what they may feel the researcher is looking for.

Experimental Design—The experimental design is a research design where one manipulates the independent variable to see if this manipulation causes changes in the dependent variable. The purpose of the experimental design is to eliminate all competing hypotheses so that the only hypothesis left is the experimental hypothesis. A subgroup of experimental designs are *Pre-Experimental Designs* (Campbell and Stanley, 1963). These experiments are ones that involve a one-time study or a single pretest, or a pretest/posttest study, and are a subgroup of the Experimental Design.

Quasi-Experimental Design—It may be impossible to eliminate all competing hypotheses from the experimental hypothesis, manipulate the independent variable, or randomly assign conditions to the dependent variable. Therefore, one can only come relatively close to an experimental design; or, the researcher is only able to achieve a *quasi*-experimental design.

In the social sciences, experimental designs are difficult to achieve. Experimental designs are found in laboratory settings where it is easier to manipulate an independent variable. An example of an experimental design would be a chemical experiment where the effects of a reagent or catalyst—the independent variable—is manipulated to see the result of this manipulation on the compound—the dependent variable—what the reagent is intended to affect.

Social science quantitatively operates in the quasi-experimental arena. Independent variables usually cannot be manipulated and it is difficult, if not impossible to eliminate all contending hypotheses.

With a conceptual picture of two types of experiments: experimental, where one can manipulate the independent variable and eliminate all competing hypotheses and quasi-experimental, where one cannot manipulate the independent variable, eliminate all the competing hypothesis, or randomly assign subjects to conditions—both experimental and quasi-experimental research designs must measure what we want them to measure in order for them to meet the test of validity.

IV. MEASUREMENT VALIDITY

The following illustration places validity in a framework of types of validity and threats to this framework.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

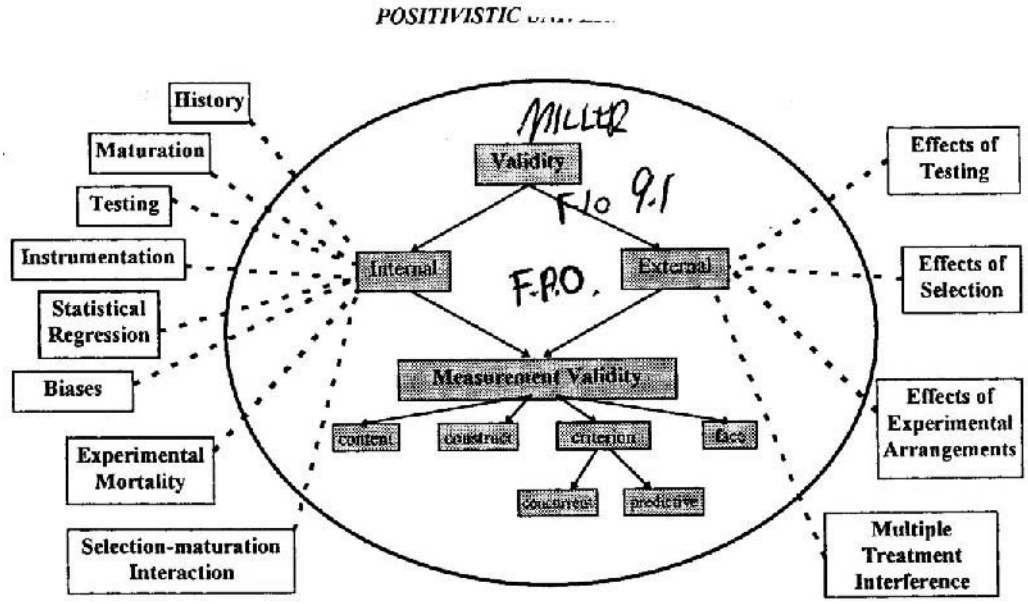


FIGURE 1 The validity framework and concomitant threats.

The above diagram is a representation of how validity exists in a positivistic universe consisting of internal and external validity, where validity is segmented into questions of accuracy based on: *content, face value, criterion,* and *construct*. The Universe of validity is threatened from extraneous factors that affect internal validity, the left side of the picture, and external validity on the right side. Campbell and Stanley (1963) presented the 8 factors that threaten internal and 4 factors that threaten external validity of experiments based on Campbell's earlier work ''Factors Affecting the Validity Of Experiments'' (*Psychology Bulletin*, 1957). All threats to internal and external validity remain applicable forty years later, and will be presented with examples appropriate to public administration.

However, a distinction should be made at the outset of any discussion of validity; validity is *not* reliability. Notwithstanding various tests for reliability of empirical methods: the retest method, alternative form method, the split halves method, and the internal consistency method (Carmines and Zeller, 1979),[2] a measurement can be *reliable* but not valid. A measurement tool can give reliable, consistent measurements but not measure exactly what one wants it to measure and therefore fail the test for validity. For example:

The state highway patrol was monitoring car speed on the interstate and unknowingly, the radar gun they were using was defective and only measured car speeds up to 68 miles per hour, i.e., if a car was going through the speed zone at 55 miles per hour the radar gun would read 55 miles per hour. However, if a car past through the zone at 75 miles per hour, the radar gun would only read 68 miles per hour. The state police measured speed for 24 hours and consistently measured speeds that were reliable and consistent, but they were not valid.

Our measurement tool may also give us consistent, reliable measurements, but validity could be compromised. For a measurement tool to be valid the tool must measure what it is supposed to measure. The police radar speed gun must be able to measure all speeds, not just speeds up to 68 miles per hour.

Reliability in observational settings can be measured by applying a formula for percent reliability. The formula measures how one observer's results may differ from another.

$$\text{Reliability} = \frac{\text{number of agreements}}{(\text{number of agreements}) \times (\text{number of disagreements})} \times 100\%$$

The Percent Reliability is considered reliable if the equation equals 90% or greater. (D.E. Pierson, 1995: p. 96). If we applied this same formula to testing the reliability of a study that was previously done, we can examine the reliability of the test measurement by applying a variation of this same formula and substitute the ''number of 'like or similar' results'' for the number of agreements, substituting the ''number of 'unlike or dissimilar' results'' for the number of disagreements. In the social sciences, more so than the ''hard'' sciences, we cannot qualify the number of ''like'' or similar results by saying the ''number of exact results.'' This distinction is made due to the inability to manipulate the independent variable in quasi-experimental designs and the threat of *randomness* to experimental and quasi-experimental designs.

To conclude the discussion of reliability, it is important to note that reliability is secondary to validity. If the measurement tool is not valid its reliability cannot be considered.

Campbell and Stanley (1963) describe two types of validity: *internal and external validity*. At the beginning of Campbell and Stanley's discussion of validity it is clear that *''internal validity is the sine qua non''* (1963: p. 5)—the essential validity—the essence of the experiment. Internal validity is examined when the researcher asks the question: ''did the independent variable cause the expected corresponding change in the dependent variable?'' An illustration of internal validity using fire stations would be the answer to the question: Did an increase in fire stations cause a decrease in multiple alarm fires in the new district? Or, did an increase in police on beat patrol cause a concomitant decrease in crime?

In contrast to internal validity, which is specific to the experiment, *external validity* asks the question of generalizability; or, to what extent can the findings of an experiment be applied to different groups, settings, subjects, and under what conditions can this experiment be generalized. Campbell and Stanley (1963) explain external validity by comparing it to inductive reference, in that it is never completely answerable (p. 5). In the example of reading scores, an experimental finding may be:

Students in New York City public high schools with an enrollment in excess of 5000, have lower reading scores than public high schools in the same city with less than 5000 enrollment. However, this experiment may not be able to be duplicated in Newark, Chicago, or Los Angeles. In short, while high enrollment in New York City Public Schools may cause lower reading scores it might not have the same effect in another area of the country.

Furthermore, external validity does not rule out the possibility that while more police on beat patrol may reduce crime under one set of circumstances; less crime may reduce the amount of police on beat patrol in another set of circumstances; a case where the independent variable in experiment *A* becomes the dependent variable in experiment *B*.

The important question that persists when one examines experimental validity is: does the measurement tool measure what it is supposed to measure? This question is predicated on matters of precision and accuracy. The accuracy of the measurement tool involves several types of validity questions:

Face validity is the simplest type of validity. It answers the question: Does the measurement tool appear to measure what we want it to measure. For example:

If we wanted to measure Customer Service Department effectiveness at the Internal Revenue Service, we would not measure the eating habits, secretarial skills, or the amount of graduates from accredited graduate schools of accounting in the Customer Service Department because ''on the face of it'' these items tell us little, if anything at all, about customer reaction to customer service.

Face validity, being a simple measure of validity, is also the most innocuous measure of validity. Face validity alone is not sufficient to meet accuracy tests of validity.

Content validity asks the question: is the measurement that is being taken a subset of a larger group of measurements that represent the focus of the study? While similar to face validity, it is a more sophisticated test for validity. An example of content validity can be shown in our study of Internal Revenue Customer Service's Department.

In this study we want to determine if the customer services representative was accommodating to the taxpayer. If a survey instrument was to be used to determine customer satisfaction, the survey could ask one question: ''Were you satisfied with your contact with the Internal Revenue Customer Service Department?'' While this question may be adequate in some cases, most likely the question might attract many negative responses because the customers' needs might not be totally satisfied; or for that matter, affirmative answers might not give you the information you will need to make changes in customer service. A better approach would be to measure responses to questions that come from a subset of customer satisfaction. For example the IRS might inquire if the customer service representative:

- picked up the phone in a certain length of time following your connection to the department?
- did they identify themselves to you?
- did they inquire about your problem?
- did they give you a satisfactory answer?
- if they didn't know the answer, did they say that they would get back to you?
- did they return with an answer in a timely manner, etc?

These are typical questions that would meet the question of content validity for customer service.

In the same example, a question that would *not* meet the criteria of content validity would be: Did you submit your income tax return in a timely fashion? Not only would this question not meet the content validity criteria of customer service, but if used in a survey of the IRS's Customer Service Department, it may illicit negative responses to the relevant questions.

There are two types of *Criterion validity*—concurrent and predictive. Concurrent validity is used to question the validity of a subset of questions already verified by content validity. This subset may be created to save time during the actual questioning during a survey. For example:

A survey given to motorists at a bridge toll booth. The motorist can bring the survey home and return it on the next pass over the bridge. However, the decision makers would like a faster more immediate response to the survey instrument. They decide that they will have the bridge police set up a safe area past the toll booths and before the entrance to the Interstate. Police will assist surveyors in detaining random cars so that motorists can be asked the survey questions. Any anxiety caused by the police detaining the motorist is immediately relieved when the motorist finds that he is only being detained to answer a few questions. In order to ultimately get their cooperation they are told that their names will be entered in a raffle for a free dinner-for-two at a local restaurant. Before this plan can be initiated the survey planners realize that the motorists can't be detained to answer the current questionnaire. This would delay traffic, and slow down the process limiting the amount of motorists that can be questioned, and possibly incur the wrath of the detained motorist. The survey planners decide to create a significantly shorter survey instrument from the original questionnaire that will meet face and content validity questions and give them the information they need to meet the criteria of the survey.

Predictive validity asks the question: does the test that is being administered have some predictive relationship on some future event that can be related back to the test administered?

In the Fire Station experiment of determining alarm response time to newly developed areas of a township, we can determine:

Fire stations within a certain radius of housing developments decrease response time to alarms, while fire stations outside this radius increases response time. In this instance the Fire Station Experiment has predictive validity if we use the results of this experiment as a predictor of future fire station placement in the community. The future placement of fire stations relates the result of the experiment back to the test; and, the test can be related to the placement of the fire stations.

Construct validity relates back to general theories being tested; aptitude tests should relate to general theories of aptitude, intelligence tests should relate to general theories of intelligence, etc. For example: in the bridge repair experiment:

The county engineers realize that certain heavy equipment must be utilized by mechanics hired by the county. They want to give aptitude tests for potential hires to reduce their liability during construction. The assumption is made that the engineers or those creating the aptitude test for using heavy equipment, understand what constitutes aptitude for using heavy equipment during bridge construction. The test to measure aptitude—the construct validity—must relate back to general theories of aptitude, not theories of heavy equipment.

V. THREATS TO RESEARCH DESIGN VALIDITY

Threats to Internal and External validity are variables—different from the independent variable—that affect the dependent variable. When one explains the methodology of their research design, they must address how they confront the threats from these variables, or how these threats are controlled. These threats, or extraneous variables, which need to be controlled in the experimental design, (Campbell and Stanley, 1963) will be presented as an introduction to each threat to validity, as listed in Campbell and Stanley's *Experimental and Quasi Experimental Designs for Research* (1963).

A. Threats to Internal Validity

> *1. History—the specific events occurring between the first and second measurement in addition to the experimental variable* (Campbell and Stanley, 1963: 5).

When events occur that fall outside the boundaries of the experiment that could affect the dependent variable, internal validity has been threatened by *history*. History is a potential problem when studies are conducted in natural settings (O'Sullivan and Rassel, 1995). History is impossible for the experimenter to control for; rather, threats to the experiment's validity due to history need to be explained when discussing causality. History threatens validity when we can ascertain that an event, other than the independent variable may be associated with the dependent variable. The following example illustrates threats to validity from history:

In a study of the adequacy of existing fire stations done during the course of one year, we may find that the existing fire stations were not adequate as evidenced by the number of multiple alarms fires (requiring more than one fire station to extinguish). In this case the relationship we are looking for is that the number of multiple alarm fires is negatively related to the number of fire stations in a district. However, during the course of the year in which the data came from, the summer was exceptionally hot, and there was a drought. The drought lasted for approximately two weeks; nevertheless, it was also a period when the temperature was higher than normal. Since the area encompassed large expanses of rural and undeveloped areas, numerous brush fires occurred. Due to the stage of drying that the brush was in, the fires spread rapidly and soon required a second or third fire station to respond.

In the above example the results were affected by the extraneous variable *history*. It is impossible to control the extraneous variable—weather; and, the affect that the weather had on drying and the spread of fires. The study's validity is threatened, but not totally invalid. In this case, if one explains the affect of history and that the threat to validity is actually a contingency that districts should be prepared for, the study still has merit.

> *2. Maturation—the processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like* (Campbell and Stanley, 1963: 5).

When changes occur—naturally and ineffectively—over a period of time, in the group being studied, the threat to validity is called maturation. Commonly, studying children, or any group that may go through rapid physical and social changes, affects the validity of the experiment. Typically studies of education of a cohort group may occur over a period of years. For example:

A study of reading skills of children in the primary grades is undertaken. Students will be tested over a period of six years from grades kindergarten through Grade 5. Students will be tested five months into the kindergarten school year and then at the end of kindergarten. Subsequently, reading skills will be tested every year at the end of the school year.

In this example maturation is expected to occur. The question that maturation compels us to ask is: without the educators teaching reading skills—the independent variable—would we receive similar changes in the dependent variable—improved reading scores—without the affect of the independent variable? Children grow rapidly both socially and physically; and, this rapid growth, the maturation in both physical and social contexts, may have an affect on the experiment.

> *3. Testing—the effects of taking a test upon the scores of a second testing* (Campbell and Stanley, 1963: 5).

In an experiment where a group is given a pretest before the introduction of the independent variable, the pretest sensitizes a group to experimentation and their response to the independent variable may be attributed to the pretest and not the independent variable. The administration of the posttest, which shows the affect of the independent variable, must be reviewed in the context: did the pretest effect changes in the dependent variable? In short, could a pretest group associate questions from the pretest to the experiment and affect the results by consciously or unconsciously taking that experiment to that end; or, as a result of the pretest and what they remember from it, i.e., the experimental or control group are ''good test takers,'' the group does better on the posttest because their test taking abilities affect causality and not the effect of the independent variable.

A more interesting way of illustrating the effects of testing is what has become commonly known as the Hawthorne Effect. An experiment that was begun to test workplace efficiency at the Hawthorne Electrical Plant soon became the basis of the Organizational Development theories of administration. The essence of Hawthorne was that the employees that were being tested performed better—were more efficient despite workplace conditions being more and then less favorable—because they knew that they were being tested. Researchers who have tried to duplicate this experiment have been unsuccessful and have repudiated the validity of Hawthorne. To the extent that other researchers have disavowed the Hawthorne experiments based on validity there is merit; however, to the extent that they reject Hawthorne as a lesson for Organizational Development, they are mistaken.

Notwithstanding Hawthorne, the following public administration example shows the effect of testing threats to validity in terms of pretest and posttest knowledge. A city may be looking for ways to streamline trash collection and at the same time reduce personnel costs. Other cities,

such as New York found that the introduction of a ''Two-Man Truck'' (as opposed to a three man truck) reduced costs and was an effective means of collecting trash. At a city council meeting the mayor proposes the New York model as one that might work in their city. The city council, at their public meeting, decides to do a study in one of the city's sectors. However, there was concern that the increased costs and maintenance required on the new trucks may not offset the savings in personnel costs. They decided that they would do efficiency and cost measurements under the current system, while awaiting an order for two Two-Man Trucks. The local newspaper reporter, covering the council meeting, reports the results of the council meeting in the next day's edition.

Within two weeks, efficiency experts are dispatched with the sector's two trash teams. Aware that they are being tested, and conscious of the purpose of the study, the men outdo themselves collecting the trash. When the new trucks arrive and a post-test is administered, production and efficiency did not improve, which was anticipated by the council, and the savings in personnel costs of one less man on the Two-Man Trucks, did not off-set the cost of the new trucks and the anticipated maintenance on the vehicles.

Obviously, the fact that subjects became aware that they were to be studied and the concomitant realization that their livelihood may be threatened, affected the results of the experiment. In this case the pretest, as well as information that the groups would be tested, threatened the validity of the experiment and skewed the results.

> *4. Instrumentation—in which changes in the calibration of measuring instrument or changes in the observers or scores used may produce changes in the obtained measurements* (Campbell and Stanley, 1963: 5).

When changes occur in the interpretation of the dependent or independent variable, or the methodology changes during the interval from the pretest to the posttest, these changes are interpreted as threats to validity from instrumentation. It is not unusual that during the course of a social science experiment threats to the validity from instrumentation occur. For example:

During a meeting of the local school district, a principal was concerned that shortly after lunch it seemed that students participated less in activities designed to foster participation. The principal's theory was that the lunch provided by the district was not healthy and the amount of fats and empty calories used in the diet were the major factor for this lack of participation. To illustrate his point, the principal brought with him a nutritionist who attested to the fact that the essence of the school lunch program was ''junk food.'' The board decided that a study should be commissioned to determine if there was a relationship between school lunches and the level of participation in school activities after lunch. The study was to encompass the school term from September to June. Contacts were made with the school district in the next county, which had more nutritionally sound meals, to act as a control group. After the study was in effect for three months the same nutritionist presented her case in front of the state senate. Shortly after her presentation, a bill was introduced, passed by the state legislature, appropriate vendors found, and state-wide, nutritionally sound meals were mandated in all the school districts. However, the commissioned study continued in the school district in question and the final result of the study was that there was little correlation between the lunch meal and the level of participation.

Between the beginning of this study and the end, a change—the state's mandate that nutritionally sound meals be served—occurred that may have affected the validity of the experiment. Instrumentation threats to validity are common when studies examine issues that can be affected extraneously by changes in laws or the court's interpretation of existing laws.

> *5. Statistical Regression—operating where groups have been selected on the basis of their extreme scores* (Campbell and Stanley, 1963: 5).

Statistical regression threatens validity when the study chooses to include in the experiment an outlier score—a score higher or lower than expected—at the time of the pretest. The expectation of such a score is that if the subject is evaluated again the score on the next test will be substantially lower, or higher than the previous test; i.e., their scores will regress toward the mean. However, if choices of subjects for the study were based on pretest outlier scores, and one does not consider statistical regression, validity is threatened. Notwithstanding the essence of validity, i.e., the test measures what we want it to measure, where those with high abilities score high, and those with low abilities score low; it would not be unusual to make errors in experimentation by not considering statistical regression. In this example:

The commissioner of Human Services wanted a breakdown of the department's community mental health partial treatment centers so that a decision could be reached on closing some of the least utilized facilities and privatizing the rest. For the last quarter, due to incidental reasons, the Fairfax Community Mental Health Center showed a decrease in admissions, substantially lower than their previous trends. The Fairfax Center had been operating for approximately ten years, and always maintained a high new-patient census. However, this decrease in new admissions was assumed to be the result of population shifts and better utilization of new treatment modalities. Based on the low admission rate for new patients and the recent increase in new drug utilization, a decision to close Fairfax was reached. Fairfax community leaders were not in any hurry to temper the Department of Human Service's decision as the community mental health center was a continual cause of discontent within the community. Shortly after Fairfax closed, the community witnessed an increase in the homeless population, crime, and the suicide rate.

The above is a typical example of not considering statistical regression as a threat to validity. Fairfax Community Mental Health Center was experiencing some type of ''blip'' in their admission rate. The low admission rate represented an outlier score that was too low. Had the Fairfax admission rate been viewed for the ensuing three months, the rate would most likely revert or regress to Fairfax's historical mean admission rate.

6. Biases—resulting in differential selection of respondents for the comparison groups (Campbell and Stanley, 1963: 5).

Bias or Selection is a threat to internal validity when the subjects, cases, scores, etc., are not chosen randomly. On the face of it, biases are something that we inherently avoid so as not to appear prejudiced. However, the threats to validity from bias and all other threats to internal validity can occur with or without the researcher being aware of these threats. Biases occur when we choose comparison groups that are uniformly different from each other. Our results then become affected by our biases so that the results obtained would not have been obtained if the differences between the experimental and control group were less extreme. The following example of an invalid study is one where the researcher purposely biased the experiment:

Baby formula companies have often lobbied for state Infant Nutrition Programs to bolster product sales. In one such state, pressure from the American Academy of Pediatrics Council on Nutrition lobbied state legislators saying that such programs limit the use of breast milk as the primary choice of nutrition for babies. Seeing that this pressure from the pediatric community might limit the power base of the agency by eliminating the program, there was a need to show program success. Analysts in the Department of Health in favor of the continuation of the Infant Nutrition Program, decided to conduct a study investigating the relationship between infant morbidity and participants in the program and using an experimental control group who were infants who were not participants in the Infant Nutrition Program and who were identified by the Health department as those who were born of crack-addicted and HIV positive mothers.

In the above case, the stark difference between the experimental and control group is so systematic that this difference or selection process has replaced the independent variable—

participation in the Infant Nutrition Program—with the threat to validity—bias, which would be the factor that had the ultimate effect on the dependent variable—infant morbidity.

> *7. Experimental Mortality—or differential loss of respondents from the comparison groups* (Campbell and Stanley, 1963: 5).

The threat to internal validity that makes the researcher more concerned with those experimental subjects who leave or drop-out of the study rather than remain in the study until completion, is experimental mortality. Further, experimental mortality includes those subjects who are misplaced in the control group, i.e., those subjects who at one time before the experiment or during the course of the experiment, were exposed to part of the experimental treatment—a stage or effect of the independent variable—and then incorrectly assigned to the control group. Whether a drop-out or a misplaced, exposed member of the control group, the experimenter must ask if the experiment would have been any different if those who dropped out, remained, or if those that were incorrectly assigned to the control group were assigned to the experimental group. Regarding the drop-outs, the researcher must not only inquire how his results would have been different, but is there an effect of the independent variable treatment that caused the subject to drop out. There are obvious examples of both drop-outs and incorrect assignment that can be applied to any pharmaceutical test on a new drug. Drop-outs can be described by a pharmaceutical experiment where the effects of the drug during the course of the experiment caused the subject to leave. In this case, the researcher must determine if an unfavorable treatment reaction affected the drop-out; if that subject had stayed to the end of the experiment, how would it affect the experiment's results; and, could the person have been exposed to some earlier derivative of the drug, it its natural or chemical form that would have sensitized the subject to the drug?

> *8. Selection Maturation Interaction—which in certain of the multi-group quasi experimental designs...is confounded with...the effect of the experimental variable* (Campbell and Stanley, 1963: 5).

Selection maturation interaction is what can be described as ''design contamination'' (O'Sullivan and Rassel, 1995), ''diffusion or imitation of treatments'' (Jones, 1985), and other sobriquets. At the least, it is contamination of either the control or the experimental group that negates the effect of the experiment unless one is doing research on the effects of contamination. Benignly, selection maturation—contamination—is related to the threat *testing*. This occurs when the experimental groups guess the purpose of the experiment and gravitate toward that end. Malignantly, contamination occurs when one group tells another group what they are experiencing or what they believe the experiment is about, and this cross-contamination places the experiment in the validity danger zone. For example:

A long time problem in education is the use of a testing model to evaluate teaching performance through a testing instrument given to their students. Recently education researchers had developed a testing instrument that would eliminate 65% of the variance. School districts throughout the country are excited about this development. Politicians who feel that teachers are overpaid and not productive enough are eager to see the results of the experiment. The teachers union feel that a test of this type is just another ploy to adversely affect contract negations to their constituency. Before the test is administered a thorough description of the examination is picked up by the press and considering the hot political issue the test has developed into, publish the story and various follow-up pieces. Teachers, unions, and families discussing the test with students, constituents, and children, have sensitized the students to the issue. In many schools, teachers who believe they have an idea of the sum and substance of the test discuss the test in the classroom. Students impressions are influenced and the scores on this test show that teachers are performing well.

These threats to internal validity have been augmented and altered to include other threats that are merely variations on a theme. *Compensatory Rivalry and Compensatory Equalization* (Jones, 1985) are variations of contamination. Similarly, selection maturation interaction (Campbell and Stanley, 1963), where one group finds out that another group is being treated better than they are being treated, are some of the variations. Essentially the basis of internal validity are these eight, all others are derivatives with different ''spins.''

B. Threats to External Validity:

As mentioned previously, external validity focuses on how well the experiment can be generalized or applied to different groups; or, how we can refer the results of one experiment to the hypothesis of another similar experiment? Again, as in discussions of internal validity, Campbell and Stanley's (1963) descriptions of these threats are presented as benchmarks for interpretation. Sometimes the differences between the threats to internal validity and external validity are subtle and it is important to direct one's focus on the nuances of their differences.

> *9. The Reactive or Interactive Effect of Testing—in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results...unrepresentative of the effects of the experimental variable for the unpretested universe...* (Campbell and Stanley, 1963: 6).

Testing, which is also an internal threat differs from testing as an external threat because, internally, testing affects the subjects as a result of a pretest that changes behavior. External validity is threatened by testing when the successful outcome of the program can only be replicated by the administration of a pretest. Without the pretest, the experiment cannot be generalized. For example:

Participants in a county maternity and child health clinic were studied to determine if they knew when to bring their child in for care. The pretest tested their knowledge of child health and disease. The mothers were then given child health classes tailored to meet their requirements based on the pretest. The result of the classes reduced infant morbidity and mortality in the county. Other counties wanted to initiate this program of education to achieve the same result. The educational programs were initiated without the administration of a pretest and the education classes were lectures that were not tailored to educational needs. Child morbidity and mortality were not reduced in the other counties.

In the above, the experiment was not able to be generalized because the interpretation of the effect of the treatment—the independent variable—was education. By not giving the pretest in the other counties and merely initiating a program of education without evaluating needs of the target population, external validity defeats the pertinence of the general application of the experiment elsewhere.

> *10. The Interaction effects of the selection biases and the experimental variable* (Campbell and Stanley, 1963: 6).

When an experiment produces results that show a relationship between the dependent variable and the treatment, despite the fact that biases were used in determining participation, the experiment may still be internally valid, in the context of the original experiment, in that there is some commonality between the two groups. However, when the experiment is repeated elsewhere, the use of a significantly different control or experimental group, more or less applicable, produces disconcerting results. For example:

In New York City, an experiment was conducted to examine the relationship between firemen's heath in fire stations where there are many more alarms than in less busy fire stations. It was determined on the basis of this study that there was little correlation between the firemen's

health in busy fire stations than the less busy fire houses. When the study was duplicated in less populated urban areas, the opposite results were obtained.

In this experiment selection biases are a result of the assignment of firemen to busy or less busy fire stations. In New York, where there are many fire stations in both busy and less busy areas, younger firemen are assigned to the busier fire stations and firemen who are older and have been through assignments in busy fire stations, have been placed in the less busy stations. As a result of this bias, older firemen who may have developed poorer health over the years are being compared to younger firemen who are generally in good health. In other cities where there are fewer fire stations, there is little choice as to assignment and firemen may stay in one station for their entire tenure as firemen.

> *11. Reactive Effects of Experimental Arrangements—which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental settings* (Campbell and Stanley, 1963: 6).

When the experimental arrangements—the situation in which the experiment is being conducted—is so clinical or controlled that the duplication of the experiment in nonclinical surroundings is virtually impossible, then the experiment is threatened by the threat to external validity of the arrangements themselves. This threat also applies to the situations where testing validity is threatened in that the subjects know that they are being tested and alter their behavior accordingly as in Hawthorne or in the following:

Residents of a community are told that they are to pilot a program of community policing initiatives to lower crime in the area. The residents of the community are individually visited by the members of the township's police department and the program is explained to them. The residents are to report any suspicious cars or people that they see in their community to a special police telephone number. The residents are enthused about this experiment and the new relationship with a formally aloof police department that the experiment is effecting. The residents perform exceptionally well and community crime is reduced. Other townships decide on the basis of this experiment to implement their own test of community policing. However, the lack of partnership between police in the other townships and the community—the lack of special arrangements—shows that community policing programs make little difference in reducing crime.

> *12. Multiple Treatment Interference—likely to occur whenever multiple treatments are applied to the same respondents because the effects of prior treatments are not usually erasable.*

When experimental treatments are applied to a group of subjects they affect the participants in the study and cannot be undone. When this occurs and other independent variables are applied to the same group, the effect of the previous independent variable affects the reaction to the new independent variable. When attempts to duplicate the experiment is attempted without the previous experimental treatment given to the original group, the experiment can't be duplicated. In the case of the community policing, if the original test community was also the community used to test volunteerism in reporting the location of trash and refuse along the community streets in order to develop a cleaner, more attractive community, the effect of this treatment may have influenced their participation in community policing.

This discussion of validity is one that should raise the level of consciousness of the researcher that there are threats to all experimentation that have to be considered in the research design. These threats must be considered early in the design process. Constantly throughout the experiment attempts to control and limit these threats are what makes the experiment more valid and applicable in settings other that the experimental and observational environment.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

V. EXPERIMENTAL DESIGNS

Experimental designs offer researchers the format for inferring a relationship between a theory and the application of that theory. The relationship between the dependent variable can be an association or a correlation but more often than not the relationship cannot show true causality—the cause leads to the effect—in the strict sense of the word. Even in total, clinical, experimental research, randomness threatens causality (Blalock, 1961). Furthermore, if all attempts to control for random selection are observed, total randomness is always questioned. This phenomenon is increasingly demonstrated more in quasi-experimental designs than in experimental designs; nevertheless, it appears to be more accurate to refer to the relationship between variables as an association, a relationship, or a correlation, especially in the social sciences, rather than refer to the relationship between the dependent and independent variables as one of cause and effect.

Often designs are expressed using symbols of:

R = randomly chosen subjects
O = the observation or the measurement of the effect on the dependent variable
X = the independent variable

While looking at combinations of these symbols in describing experiments presents some confusion on the part of the student, there is little alternative to presenting a diagram of the research designs in this way. However when a description of the experiment is plainly given, the diagram of the experiment eventually presents a visual representation of which design is desired.

Furthermore, it is important to note that experimental designs and variations and derivatives of those designs are numerous and often the researcher uses different combinations of designs to investigate a phenomenon or prove a theory. It would be unduly cumulative to present all of the combinations that could be created; and, at the same time some combination would invariably be omitted.

The following designs are listed and described using the typical R, X, O format. The threats to validity and the internal and external controls—the strength of the experimental design over the threats to validity are also identified.

A. Pre-Experimental Design

One-Shot Case Study

$X \quad O$

Threatened Internally by:	Threatened Externally by:	Controls Internally:
History	Selection Interaction	none
Maturation		
Selection Bias		
Mortality		Controls Externally:
		none

As the design states a one-shot case study does nothing more than observe the effect of the independent variable. It reports nothing more than some contrast or difference in a group attributed to some treatment. There is little scientific value to the one-shot case study; it is at risk from the most relevant threats to validity; and, it does not control for any threats. Other than a bearing from where to begin a discussion of experimental design, one-shot case studies offer little utility in the social sciences other than single ''snapshots'' of a group at one point in time.

One-Group Pretest-Posttest Design

O X O

Threatened Internally by:	Threatened Externally by:	Controls Internally:
History	Testing Interaction	Selection Bias
Maturation	Selection Interaction	Mortality
Testing		
Instrumentation		Controls Externally:
Selection Maturation Interaction		none
Mortality		

With its numerous threats to validity, the One-Group Pretest-Posttest Design is just slightly better than the One-Shot study or as Campbell and Stanley state: ''...enough better than Design 1 [One-Shot Case Study] to be worth doing when nothing better can be done'' (1963; p. 7).

An example of the One-Group Pretest-Posttest Design are studies of reading skills development where a group is tested and then after some period of time the same group is tested again.

Static-Group Comparison

$$\frac{X \quad O}{O}$$

Threatened Internally by:	Threatened Externally by:	Controls Internally:
Selection Bias	Selection Interaction	History
Maturation		Testing
Selection Maturation Interaction		Instrumentation
		Regression
		Controls Externally:
		none

Static-Group Comparison studies are done where two groups are observed, one receiving the effect of the independent variable and the other group not experiencing the treatment. Static-Group Comparisons are useful when comparing program participants—children that have participated in operation ''Head Start''—with the reading level of those who did not participate in the Head Start program. The single accomplishment of the Static-Group Comparison is that it establishes the effect of the independent variable.

1. The Classical Experimental Design

Notwithstanding the Pre-Experimental designs where there are drawbacks that often preclude the use of these designs in order to protect research from threats to validity, the experimental designs offer more insulation from internal and external threats and are more appropriate as a research design. For this discussion the focus is on those experimental designs that:

- reflect a random selection of subjects and where there is no significant difference between an experimental and control group
- a pretest that measure the dependent variable is given to the experimental and control groups

- Both experimental and control groups will experience equal conditions except for the treatment of the independent variable
- the researcher controls the amount of treatment to the experimental group
- a posttest is given after the exposure to the treatment to both the experimental and control group
- changes due to the dependent variable and the differences between the dependent variable effect on the experimental and control group, evidenced by the posttest, is attributed to the independent variable (adapted from O'Sullivan and Rassel, 1995).

Pretest-Posttest Control Group Design

R O X O
R O O

Threatened Internally by:	Threatened Externally by:	Controls Internally:
none	Selection Interaction	History
		Maturation
		Testing
		Instrumentation
		Regression
		Selection Bias
		Mortality
		Selection Interaction
		Controls Externally:

The Pretest-Posttest Control Group Design is also referred to as the classical experimental design. This design enables the researcher to choose experimental and control groups of randomly assigned subjects. One group receives the experimental treatment while the control group does not. After the introduction of the independent variable, both groups are observed again. Differences between the experimental and control groups are attributed to the effect of the independent variable.

From the beginning of this research design—the assignment of random subjects to experimental and control groups—threats to validity are being controlled. If the selection is truly random, then biases, regression toward the mean, and all other internal threats are initially controlled in the experiment. However, as the experiment progresses over time it is practically impossible to control for maturation and maturation interaction. The following example is a description of a Pretest-Posttest Control Group Design.

A random selection of mothers at the local community health station were chosen to test if there were differences in satisfaction levels between the random group of mothers that were in the experimental group or those in the control. The study wanted to determine if they could eliminate, as a cost containment technique, nurses at the intake level and run the clinic with ancillary health professionals, nurse practitioners, and doctors. The experimental group was to be interviewed by a nurse to take a history of the current complaint, whether this was a ''well-baby care'' visit, or a ''sick-baby'' visit and answer any questions the mother may have about her child. After the initial visit by the nurse, the nurse practitioner or the doctor would come into the room to examine or treat the child. The control group would not receive the nurses visit. Both groups would receive a pretest one month after being enrolled as patients. Then the independent variable would be introduced, and a posttest on both groups for customer satisfaction.

Solomon Four-Group Design

$$\begin{array}{llll} R & O & X & O \\ R & O & & O \\ R & & X & O \\ R & & & O \end{array}$$

Threatened Internally by:	Threatened Externally by:	Controls Internally:
none	none	History
		Maturation
		Testing
		Instrumentation
		Regression
		Selection Bias
		Mortality
		Selection Interaction
		Controls Externally:
		Testing Interaction

The Solomon Four-Group Design is the first experimental design presented that controls, to some extent, threats to generalizability or duplication. The design is set up where the component of the Pretest-Posttest Control Group Design make up the first two randomized groups. In addition to this experimental and control group, a third group is added that is not given a pretest but is exposed to the independent variable; and, a fourth group that is given neither the pretest nor exposure to the independent variable. In the example of eliminating nurses at a clinic, there would be a group of clients who were not given the pretest for customer satisfaction but received the pre-visit by the nurse, and a fourth random group that received neither the pretest nor the experimental treatment of the nursing previsit.

Posttest Only Control-Group Design

$$\begin{array}{lll} R & X & O \\ R & & O \end{array}$$

Threatened Internally by:	Threatened Externally by:	Controls Internally:
none	none	History
		Maturation
		Testing
		Instrumentation
		Regression
		Selection Bias
		Mortality
		Selection Interaction
		Controls Externally:
		Testing Interaction

The Posttest Only Control-Group Design is also known as the Randomized Posttest Design (O'Sullivan and Rassel, 1995). This design also protects the experiment from the same threats to internal and external validity as the Solomon Four-Group Design. The Posttest Only Control-Group Design also presents the same opportunity for generalization as the Solomon design. However, there are times when it may not be practical, feasible, or possible to administer a pretest. The option of a pretest is removed when we are studying large groups of subjects, there are no pertinent questions to be asked in a pretest, or there is not adequate funding to administer

a pretest to the experiment's participants. Furthermore the application of a pretest takes enormous time and may not be of value. Consider the following:

The Federal government was considering changing the style of uniforms for the Air Force. Since the end of World War II, there was much discontent among members of the Air Force that the uniforms were drab and generally lacking in the type of military style that may be found in the other branches of the military. While the discontent over the uniforms ebbed and flowed over the years, recently recruitment quotas were consistently below expected levels and it was thought that changing the uniforms would enhance recruitment. To see if the new uniform would increase recruitment, new recruits, from random cities on the West Coast were given the new uniforms, while new recruits from random cities on the East Coast were issued the old uniforms. The result of the experiment was that recruitment quotas were met on the West Coast but remained at a continuous low level on the East Coast.

This posttest design experiment is one that illustrates the point that it would be difficult to administer a pretest to every adult eligible to join the Air Force; nevertheless, the posttest was able to show that there was an association between the independent and dependent variables.

VI. QUASI-EXPERIMENTAL DESIGNS

The experimental design is predicated on the ability for the researcher to be able to manipulate the independent variable, the ability to randomly assign subjects and the experimental treatments, and to eliminate competing hypothesis in their experimental research so that only the working hypothesis remains to be proved or disproved by the experiment. Once the research leaves the controlled environment of the laboratory or other controlled environment, the amount of control that the researcher normally has in experimental settings is virtually unrealizable. When the researcher is unable to randomly assign, manipulate the treatment, or eliminate competing hypothesis, the experimental design that remains is *quasi-experimental*.

The quasi-experimental design is one where the research is left to design the best possible alternative to the experimental design including as many components as possible from experimental designs. The creation of a quasi-experiment and the inability to assign random sampling, open the experiment to threats to external validity. The inability of generalizing results brings fields like public administration into the argument of whether public administration and other related social science disciplines are truly a science. The use of the quasi-experiment also leads to inertia in disciplines as the findings are difficult to duplicate; or, when others attempt to duplicate the experiment, their findings are different or explain less of the variance than the original experiment. Nevertheless, the social science tool for doing research largely involves a research design that is quasi-experimental.

Interrupted Time Series

O O O O X O O O O

Threatened Internally by:	Threatened Externally by:	Controls Internally:
History	Testing Interaction	Maturation
		Testing
		Regression
		Selection Bias
		Mortality
		Selection Interaction
		Controls Externally:
		none

The Interrupted Time Series is a design that enables the researcher to make multiple observations of a group before and after the introduction of the independent variable. The independent variable is not usually introduced by the public administration researcher; rather, this design is one that observes changes in groups that cannot be attributed to time when an independent variable was introduced by some agency, law, or action, which in the researcher's view would have caused a change to occur in the group observed. For example:

If a researcher had the hypothesis that stricter drug laws are associated with increased state prison populations: The researcher defines a period before the enactment of stricter drug laws to observe yearly prison populations. The independent variable—the stricter drug laws—is introduced, and the prison populations are observed for a period of years after the introduction of the laws.

Notwithstanding other variables that would need to be controlled, the above example illustrates the utility of an interrupted time series. Finally, in public administration, and in some other social sciences, the Interrupted Time Series takes a snap shot of some past time. The use of this technique in the present would have to entail a dependent variable that would be affected in a very short period of time or the researcher must be committed to studies that will encompass an expanse of past, present, and future time.

Equivalent Time Samples Design

$X_1\ O \quad X_0\ O \quad X_1\ O \quad X_0\ O$

Threatened Internally by:	Threatened Externally by:	Controls Internally:
none	Testing Interaction	History
	Reactive Effects of Arrangements	Maturation
	Multiple Interference	Testing
		Instrumentation
		Regression
		Selection Bias
		Mortality
		Selection Interaction
		Controls Externally:
		none

An Interrupted Time Series design does not offer the researcher the option of testing the effect of the independent variable on the test population over a period of time more than once. In contrast, the Equivalent Time Sample Design allows the researcher to do a time series experiment with repeated introductions of the independent variable. However, as the diagram illustrates, the treatment is introduced (X_1) and the observation taken; then, after a lapse of time, the observation is taken without the effect of the treatment (X_0). In this manner the researcher can observe the effect with and without the independent variable on the same population, varying the amount of observations and length of time of the experiment. The benefits of this quasi-experiment is that the effect of the independent variable may be transient or reversible (Campbell and Stanley, 1963).

The Equivalent Time Sample is useful in the social sciences in education, the workplace, or in any environment where the effect of the experimental treatment can be exposed and withdrawn.

Quasi-experimental designs continue to evolve with variations on the presented examples. For example the Equivalent Materials Design takes the model of the Equivalent Time Sample

Design and augments it with a materials aspect as an independent variable. At each point where the independent variable (X_1) is introduced, materials become the independent variable and these materials can be varied from points X_1 and X_0. The Equivalent Materials Design does control for threats from interactive arrangements and the design would be diagrammed as (Campbell and Stanley, 1963):

$$M_a X_1 \; O \quad M_b \; X_0 \; O \quad M_c \; X_1 \; O \quad M_d \; X_0 \; O$$

As shown above, the materials ($M_{a...d}$) are not the same materials, rather they change at different time points in the experiment.

The Non-Equivalent Control Group Design is similar to the experimental Pretest-Posttest Control Group Design. The difference is simply that randomness is not required; and, as previously mentioned, the inability to assign random subjects qualifies the design as quasi-experimental. The Non-Equivalent Control Group Design is diagrammed as:

$$\begin{array}{ccc} O & X & O \\ \hline O & & O \end{array}$$

The random sampling in Pretest-Posttest is replaced by an experimental and control group that are not determined by similar characteristics, i.e., the experimental and control group are similar in all characteristics except for the exposure to the independent variable. However, not so similar that the pretest can be dispensed with (Campbell and Stanley, 1963). Understandably, the Non-Equivalent Control Group Design cannot control internally for Selection Interaction; nevertheless, this type of design is useful to compare similar, defined groups that the researcher identifies as the two groups that must be tested. For example, an education study of third grade students might compare the two classes in one school without examining the subjects for like characteristics. In this way all third graders are examined, i.e., they are similar enough—not identical—and appropriate for the Non-Equivalent Control Group Design.

VII. SUMMARY

What would the state of air travel be if pilots flew wherever they wanted to without filing a flight plan? Would you even attempt, as a passenger, to fly aboard an airline where there was no plan for the flight, where the plane is going, how it should get there, and the myriad of information considered by the pilot and air-traffic controllers? The absurdity of this notion is synonymous to social science research without a plan—the research design. For the reason that a pilot would not proceed in this manner, neither should the researcher. This chapter offered the basic information on experimental design and threats to validity—the flight plan one needs to conceptualize a design for research. Having all the knowledge about experimental models offers little help if that knowledge remains on the pages in this handbook. It is hoped that in the manner that threats and designs were presented that one will be able to conceptualize threats and designs. At first, this may seem difficult, but don't give up at that point, or think it's impossible...you are just not used to doing it. Over time that will change. Nevertheless, if one is serious about research, one must conceptualize these models and be aware of the threats to those models. When these concepts are discernible, and an idea, problem, or question for research presents itself, devising a research design comes naturally. Finally, being fluent in experimental design will give the researcher the confidence needed to defend their research in all environments.

NOTES

1. There have been approaches to policy analysis that have disdained any type of positivistic approach, Post-Modernism being the most striking example. However, post-modernism in its detraction of positivism, is even more contradictory, i.e., using logic to denigrate logic, etc.
2. Reliability testing methods are discussed in various texts. In Carmines and Zeller's *Reliability and Validity Assessment*, Sage, 1979, the authors present models of reliability. The re-test method is where the same test is given to the same group of people after a period of time. The Alternative Form Method is similar to the retest but an alternative test is given after a period of time. The Split Halves method is where the test instrument is divided into two halves. The scores of each half are correlated to test reliability. The Internal Consistency method test reliability at the same time and without splitting or alternating tests. It uses Cronbach's alpha formula: $\alpha = N/(N-1)\,[1 - \Sigma\sigma^2\,(Y_i)/\sigma_x^2]$.
3. All titles for Pre-Experimental, Experimental, and Quasi-Experimental Designs are from Campbell, D.T., and Stanley, J.C. 1963. *Experimental and Quasi-Experimental Designs for Research*, Houghton Mifflin Co., Boston., unless otherwise noted.

BIBLIOGRAPHY AND REFERENCES

Blalock, H.M. 1964. *Causal Inferences in Non-Experimental Research*. New York: W. W. Norton.

Bobrow, D.B. and J.S. Dryzek (1987). *Policy Analysis by Design*, Pittsburgh: The University of Pittsburgh Press.

Carmines, E.G. and R.A. Zeller (1979). *Reliability and Research Assessment*, Thousand Oaks: Sage.

Campbell, D.T. and J.C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research*, Boston: Houghton Mifflin.

Cook, D.T. and D.T. Campbell (1979). *Quasi-Experimentation: Design Analysis Issues for Field Settings*, Boston: Houghton Mifflin.

Denardt, R.B. (1993). *Theories of Public Organizations*, Belmont: Wadsworth.

Dickenson, M. and G. Watson (1976). *Political and Social Inquiry*, New York: Wiley.

Durning, D. (1993). ''Participatory Policy Analysis in a Georgia State Agency,'' *Journal of Policy Analysis and Management, Vol. 12, No. 2.*

Fischer, F. (1995). *Evaluating Public Policy*, Chicago: Nelson Hall.

Howard, G.S. (1985). *Basic Research Methods in Social Science*, Glenview: Scott Foresman.

Jones, R.A. (1985). *Research Methods in the Behavioral and Social Sciences*, Sunderland: Sineaur Associates.

Kaplan, A. (1963). *American Ethics and Public Policy*, New York: Oxford University Press.

Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook*, Englewood Cliffs: Prentis Hall.

Lindblom, C.E. and D.K. Cohen (1979). *Usable Knowledge*, New Haven: Yale University Press.

Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods*, Thousand Oaks: Sage.

Rutman, L. (1984). *Evaluation Research Methods*, Thousand Oaks: Sage.

Spector, P.E. (1981). *Research Designs*, Thousand Oaks: Sage.

O'Sullivan, E. and Rassel, G.R. (1995). *Research Methods for Public Administrators*, White Plains: Longman.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

10
Qualitative Research Methods: An Overview

Vatche Gabrielian
Rutgers University, Newark, New Jersey

I. INTRODUCTION

There has been an increased use of qualitative research methods in social sciences in recent years. Once marginalized in theory, qualitative research methods are claiming their place in the arsenal of social science inquiry, with more and more traditional research method textbooks devoting chapters to it, and a growing number of researchers and disciplines engaging in qualitative research. The phrase ''qualitative research'' started as an umbrella term for a variety of research methods and techniques that could not be ''quantified'' for various reasons (inability to clearly formulate fuzzy concepts; small number of observations; study of unique events; loosing essence in coding the situation, etc.), but increasingly began to gravitate towards an umbrella term uniting various research methods with nonpositivist epistemology.

There is no unanimity in scientific (and practicing) community on what exactly qualitative research methods are, what are their inherent characteristics, what is their underlying epistemology (if there is any), how compatible they are with quantitative methods, to what fields of human (scientific) inquiry do they relate, what questions do they answer. Qualitative research methods are often used to mean three concepts: (1) underlying research epistemology (i.e., methods based on postmodern, constructivist or naturalistic paradigm of knowledge); (2) specific research strategy (e.g., research design that aims more to interpret and reveal meanings that actors attach to their actions rather than generalize causal relationships to the larger universe of events); and (3) specific techniques that are not operating with numbers (e.g., interviewing). The domains specified by these definitions often overlap, but are not identical—qualitative methods can be applied in a research based on a positivistic paradigm (e.g., ethnography applied to structuralist anthropology); and research based on constructivist or naturalistic paradigm can employ simple quantitative techniques such as tabulations or frequency counts (e.g., content analysis); or rigorous quantitative techniques can contain explicit articulation of researchers subjective preferences (e.g., Q-methodology). It is also important to note the qualitative-quantitative dichotomy in research methods is not very accurate: what is not quantitative is not necessarily qualitative and vice versa. For example, although renown economists such as R. Coase and A. O. Hirschman did not employ statistical or mathematical techniques and did not operate with empirical data in their classic studies, one still cannot call the logic or the method they employed as qualitative in any of the senses identified above. Very often this type of knowledge is called ordinary knowledge (Cohen and Lindblom, 1979). It is also often correctly argued that any type of human

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

argument contains interpretive elements, and quantitative research designs is not void of it, either. For example, Herbert M. Kritzer (1996) identifies three levels at which interpretive process operates in quantitative research. Qualitative research is being applied not only in social sciences, but also in such ''technical'' fields as information science or management information systems (Myers, 1996).

In recent decades the appeal of non positivistic qualitative methods have increased partly because of the following developments: (1) with the advent of information age, bombardment with images and texts created layers of illusory, virtual reality, that undermined the commonsense positivistic notions of reality, objectivity and causality; and (2) the doctrine of analytical positivism (neopositivism) came under fierce attacks from critical theorists, post-structuralist and post-modern theorists, who began to ask value-laden questions and question underlying assumptions of ''neutral'' scientific assumptions (e.g., Rosenau, 1992). The rise of qualitative methods thus was stemming from the failures of conventional social science to answer certain questions (as many felt, partly because the right questions were never asked); and the increasing attacks on philosophical underpinnings of traditional science (i.e., positivism). This led the researchers to look elsewhere for answers. As a result, qualitative methods can be seen as cross-disciplinary: they often combine knowledge from different fields and apply to an increasing number of fields and topics. As opposed to quantitative methods, which were borrowed by social science from natural sciences such as physics and chemistry, qualitative methods came to social sciences from two different sources—arts and humanities and clinical research (Chenail, 1993), where the emphasis was more on interpretation of human cognition and action (even if it is one particular person) rather than on objective and veritable generalization of confidently established causal relationships from examined group to a wider population. Another important aspect of this new approach was the emphasis on practice—on conducting naturalistic (i.e. unobtrusive research in natural settings, without manipulation) research (very often trying to change the object of the study), as opposed to ''objective'' positivistic research which was criticized as detached, ivory-tower-type of enterprise aimed more at proving existing dogmas than solving actual problems.

It is not the intention of this essay to provide a comprehensive coverage and full classification of qualitative research methods. It is a far more daunting task that will require more erudition and experience than I am able to provide. Perhaps, the best volume to refer to for this purpose is the *Handbook of Qualitative Research*, edited by Norman Denzin and Yvonna Lincoln (1994). Neither it is the aim of this essay to lay down a practical guide of conducting qualitative research. One article, I am afraid, will not suffice for it. Readers interested in more down-to-earth, clearly written practical guides can turn to growing number of volumes from Sage Publications (e.g., Strauss and Corbin, 1990; Patton, 1987, 1990; Marshall and Rossman, 1995; Miles and Huberman, 1984, 1994), as well as examine classical studies in qualitative research that clearly describe employed procedures (e.g., Glaser and Strauss, 1967). This essay will rather try to sketch a brief roadmap of different paradigms of qualitative research, discuss various strategies and tools of qualitative inquiry and their possible combination with quantitative methods. As everything else connected with qualitative research, nothing in this article can be claimed to be exhaustive and/or objectively true.

II. DEFINITIONS

Qualitative research defies easy classification. It is a loose assortment of complex and interconnected concepts, terms and assumptions that crosscut disciplines, fields and subjects matter, and which assume different meanings in different historical contexts (Denzin and Lincoln, 1994:

1–2). Qualitative research is often described by listing its diverse methods and the fields that they are applied to. Among often mentioned categories are ethnography; participant observation; ethnology; textual, hermeneutic, semiotic, and narrative analysis; analysis of archival and material culture; discourse and communication analysis; analysis through symbolic interactionism; ethnomethodology; psychoanalysis; feminist inquiry; phenomenology; phenomenography; survey research; deconstruction; action research and participatory action research. Qualitative research is not confined to certain discipline, and is employed in wide range of disciplines such as anthropology; education; sociology; literary and art studies; cultural studies; history; archaeology; biography; program evaluation; clinical studies; medicine; psychiatry; nursing; family therapy; and cognitive and ecological psychology (e.g., Denzin and Lincoln, 1994; Marshall and Rossman, 1995).

Perhaps, the most significant conclusion that one can draw from this diversity of methods and fields is that the most important tool of qualitative research is the researcher him/herself, who employs multiple methodologies and very often has multi-focus tasks. As Lincoln and Guba (1985) argue, human beings possess unique qualities as an instrument of research—they have the capacity to respond to a wide range of hints, to make often unpredictable mental associations and references, to see the phenomena from a holistic perspective, while detecting atypical features, to process data on spot, and test out the new knowledge immediately. Many qualitative researchers speak about the importance of what Barney Glaser (1978) labeled as ''theoretical sensitivity.'' Anselm Strauss and Juliet Corbin define it as ''the attribute of having insight, the ability to give meaning to data, the capacity to understand, and capability to separate the pertinent from that which isn't'' (Strauss and Corbin 1990: 42). Theoretical sensitivity can stem from mastery of literature, as well as professional and personal experience. Qualitative research often implies multiple methodologies. For example, in grounded theory approach, multitude of theories can be verified against existing data, when a new perspective is being tested based on one's conclusion of centrality of emergent categories. The diversity of methodologies is often called bricolage, and the researcher a *bricoleur*—a person that renown anthropologist Claude Levi-Strauss (1966: 17) defined as a ''Jack of all trades or a kind of professional do-it-yourself person.'' Qualitative researchers by and large espouse a tolerant view of the field and see variety of methods as equally important and able to provide important insights. One important characteristic of qualitative research is the tendency toward *triangulation*—the act of bringing more than one data source or more than one perspective to bear on a single point. Initially started as triangulation of data—use of variety of sources for the research, the concept of triangulation moved to include investigator triangulation (use of multiple researchers); theory triangulation (use of multiple perspectives on a single set of data); methodological triangulation (use of multiple methodologies for a single problem); interdisciplinary triangulation (looking at the same problem from different vantage points) (Denzin, 1978; Janesick, 1994). Although triangulation is highly desirable, it is also quite costly. It is important to note that the same tools of qualitative research can vary in their meaning and relevance across different fields, depending on research design, field of study and scientific paradigm. For example, use of ethnography by in cultural studies will yield different results than if it was used in classic structuralist sociology. The first would focus on the establishment of multiple meanings for various persons or subgroups within the studied population, while the second will try to explore latent but real structures that define the behavior of the community.

Based on this, many researchers distinguish between research techniques (tools) and methods (strategies of inquiry). In this view, research method is qualitative if its intent and focus is one interpretation and understanding rather than explaining and predicting (e.g. Erickson, 1986). Understanding is seen as more contextual and specific, while explaining is seen more like laying down lawlike patterns of phenomena under investigation that will apply in the future and similar

situations as well.[1] For example, one can understand the politics of budgeting in the field of water resources in the 1950s, but cannot explain the budgetary politics of water resources in the 1980s based on that understanding: the context—the structure of the Congress, the clientele, the agency leaders and the personnel, media awareness, mass communications, etc.—has fundamentally changed for any outcomes to be predicted accurately according to earlier models. In essence, this line of reasoning is an argument for defining qualitative research as a paradigm of research with certain assumptions about ontology (reality), epistemology (knowledge) and methodology (tools). This argumentation rejects the definition of qualitative research as addiction to methods (that many would dismiss as ''soft'' science) and tries to picture qualitative research as an expression of non-positivist scientific paradigm. This brings us to examination of competing paradigms of scientific inquiry.

III. UNDERLYING PARADIGMS OF QUALITATIVE INQUIRY: A HISTORICAL PERSPECTIVE

Since the classic work of Thomas Kuhn (1962), the notion of scientific paradigms has been in the center of the social science debates. There have been so many announcements of collapses of old paradigms and emergence of the new ones, that as Paul Diesing (1991: 55) notes with deft irony, if all of these were true, social sciences would have been experiencing birth of a new paradigm every six months. Egon Guba and Yvonna Lincoln (1994: 106) define paradigm as ''a set of *basic beliefs* (or metaphysics) that deals with ultimates or first principles. It represents a *worldview* that defines, for its holder, the nature of the ''world,'' the individual's place in it, and the range of possible relationships to that world and its parts, as for example, cosmologies and theologies do. The beliefs are basic in the sense that they must be accepted simply on faith (however well argued); there is no way to establish their ultimate truthfulness.'' Guba and Lincoln distinguish three main attributes along which paradigms differ: ontology; epistemology; and methodology. Before discussing particular paradigms of inquiry, it is important to mention one important characteristic of paradigms. In Kuhnian model, old paradigms collapse and new ones take hold, and mature science is united by a single paradigm. Others have contested this dynamic, saying that history of science ''has been and should be a history of competing research programs (or, if you wish, paradigms)'' (Lakatos, 1970: 155). It has also been argued that this Kuhnian dynamic of paradigms, even if true for natural or experimental sciences, is not necessarily true for social sciences, where coexistence of paradigms does not mean undeveloped, ''pre-paradigm'' state of affairs (Diesing, 1991: 56). Still another view can be Hegelian dialectic interaction between paradigms—when new paradigms are a result of competing paradigms. Some will argue that the ''punctuated equilibrium'' model or Schumpeterian ''classical situations,'' when there is more or less consensus about the body of the scientific theory, is more characteristic for social sciences (Heilbroner and Milberg, 1995). One should be cautious when discussing paradigms in social sciences—the picture should not be of extreme diversity with scores of paradigms (e.g., including feminism or ethnic studies as paradigms when in reality they are fields of inquiry with definite standpoint and diverse methodologies and epistemological assumptions), neither it should be a mortal battle between only two paradigms (e.g., positivism and postmodernism). As shown above, qualitative research is seen by many as anti-positivistic in essence, as a method of inquiry geared towards understanding rather than explaining (e.g., Erickson, 1986), though the opposite viewpoint also has a following (e.g., Miles and Huberman, 1994, 1984). Such vision entails certain assumptions about ontology and epistemology, or subscription to certain paradigm. There are several insightful classifications of paradigms one can draw upon for further elaboration (e.g., Burrel and Morgan, 1979; Morgan and Smirchich, 1980),

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

but perhaps for our purposes the most appropriate is the one provided by Guba and Lincoln (1994) (Table 1).

From the discussed four paradigms positivism is the reigning queen in social sciences. In positivist paradigm there is an objective truth that can be uncovered through structured and rigorous quantitative study. The results are applicable to a larger part of the society than the study examined. For example, if one tries to introduce a new social program, they may test the program as the natural sciences do (e.g., testing a drug).[2] The researcher will randomly choose two groups to participate in the experiment, of which the first will receive the benefits of the program, while the other—control group, will not. Both will be monitored, and by the end of the experiment, the researcher will quantify the change in both groups (e.g., how many in each group have changed their behavior, occupational status, become more active in policy participation, etc.) and apply statistical techniques to see whether there is significant difference between the two groups and can it be generalized to the whole population. The researcher will control for many variables (e.g., the gender and age of the recipients of the benefits of the program) to make the comparison meaningful. The results will be judged on the basis of validity (internal and external), reliability and objectivity. Here the assumptions are that given similar structures and incentives people behave similarly (one objective truth); that there is clear separation between the researcher and the participants of the experiment and the researcher does not influence their behavior otherwise and can observe their behavior (the object and the subject are separate and the truth is knowable); that by having a control group and controlling for age and gender and other characteristics the researcher can correctly test the hypothesis about the benefits of the program influencing the behavior of the recipients (testing of hypothesis through manipulation); and that the application of the findings to the society at large will solve the problem the program is addressing (controlling the problem).

In postpositivistic (in sense Guba and Lincoln use the term) perspective a similar design will be applied. The only difference, perhaps will be greater tolerance for error—the findings will be probable, rather than established and verified laws; but they will be considered true until falsified. Qualitative research may be employed to augment what is essentially positivistic design. For example, some (or all) aspects of participant behavior that cannot be quantified easily, will be given to independent experts, who, based on unstructured interviews, will grade them according to certain scale, which later will be statistically analyzed. Here the process of interpretation is explicitly incorporated into the research design.

The researcher in critical perspective first of all will attack the premise that there is no link between the researcher and the participants (the subject and the object). Just the fact that participants know they participate in experiment, the proponents of critical approach will argue, will change their behavior and will not make it authentic (the so-called Hawthorne effect). If the change in experiment is not ''authentic,'' they would argue, why not to combine the process of learning (the experiment) with process of desired change (the practice). The critical perspective is not value-free, or more correctly, is explicit about the values. The critical perspective actively advocates emancipatory, empowerment ethic. On the other hand, the critical researchers will argue that the aim of research should not be finding whether certain incentives influence behavior, but rather, understand what causes that behavior—what are the existing structures that shape undesirable behavior (historical realism) and correct those. The change can be achieved through a dialogue between the investigator and the participants, which will help to educate and emancipate the participants and transform the unjust structures through more informed consciousness (overtly participatory research leading to critique and transformation). For example, whereas in positivistic approach manipulation of welfare benefits (e.g., changing the mix of benefits, duration and eligibility) can be seen as a way of studying the problem of teen pregnancy, in critical approach the research is more exploratory (the variables are not

TABLE 1 Basic Beliefs (Metaphysics) of Alternative Inquiry Paradigms

Item	Positivism	Postpositivism	Critical theory et al.	Constructivism (naturalism)
Ontology	naive realism—reality is ''real'' and apprehendable.	critical realism—reality is ''real,'' but only imperfectly and probabilistically apprehendable: it should be approximated as much as possible, but cannot be fully captured.	historical realism—reality consists of crystallized (reified) structures (that are ''real'' for all practical purposes) that over time were shaped by social, political, cultural, economic and ethnic, and gender factors.	relativism—there are multiple realities that are constructed; experiential; local; specific and dependent for their form and content on the individuals or groups holding the constructions.
Epistemology	dualist (meaning clear separation between the knowing subject and examined object of the study with no influence in either direction) and objectivist; findings are true.	modified dualist (meaning clear separation between the knowing subject and examined object of the study with no influence in either direction) and objectivist; critical tradition; findings probably true.	transactional/subjectivist (meaning interactive link and mutual influence between the investigator and the investigated object); findings are value-mediated.	transactional/subjective (meaning interactive link and mutual influence between the investigator and the investigated object); findings are literally created as the investigation proceeds.
Methodology	experimental/manipulative; (propositions are stated in form of hypotheses and are subject to empirical or logical verification; confounding conditions are controlled); methods employed are chiefly quantitative.	modified experimental/manipulative; critical multiplism (hypotheses are not verified but tested against possible falsification); discovery is reintroduced as an element in inquiry; there is a drive to include qualitative methods and do research in more natural settings.	dialogic/dialectical; the dialog between the investigator and the subjects of the inquiry must be dialectical to transform ignorance and misapprehensions (accepting unjust status quo as immutable) into more informed consciousness (seeing and showing how the structures can be changed).	hermeneutic/dialectical; social constructions can be elicited and refined only through interaction between and among investigator and respondents: constructions are interpreted through hermeneutic techniques, with an aim of creating more informed and more sophisticated new consensus.
Inquiry aim	explanation: prediction and control.		critique and transformation: restitution and emancipation.	understanding and reconstruction.
Nature of knowledge	verified hypotheses established as facts or laws.	non-falsified hypotheses that are probable facts or laws.	structural/historical insights.	individual reconstruction coalescing around consensus.
Goodness or quality criteria	internal validity (isomorphism of findings with reality); external validity (generalizability): reliability (stability and replicability) and objectivity (does not depend on observer).		historical situatedness; erosion of ignorance and misapprehensions; action stimulus.	trustworthiness criteria (credibility; transferability: dependability and confirmability); and authencity criteria (fairness, enrichment, education, stimulation to action and empowerment).

Source: Adapted from Guba and Lincoln, 1994: 109–117. Reprinted by permission of Sage Publications, Inc.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

clearly identified and are thought to be in more complex than causal relationship) and directly addresses the problem. The research design may be working with teenage girls to understand and transform their consciousness about their possibilities and prospects, as well as to ameliorate the structures that induce the undesirable behavior (e.g., alleviation of poverty, providing better education). This emancipatory action stimulus will be part of criteria that the research will be evaluated. This type of research here will be with more local than global ambitions (i.e., striving for local relevance).[3] More often the format of critical research will be narrative rather than quantitative.

With the constructivist approach the research design would be different, non-experimental. In study examining causes of progress (success) in higher education a constructivist may argue that one cannot compare experiences (and thus, their understanding of the world and their logic of action) of minority students in an urban state university and students from preppie schools in Ivy League colleges—because education and grades (and a host of other variables) have different meanings for these groups of students. They mean different access to jobs (they may be seen as mattering or inconsequential), they have different cultural relevance (certain values may seem imposed while others overemphasized), they have different social meaning in the peer circle (sports may be more important), etc. Thus, from a constructivist perspective, a single theory cannot possibly cover the asked question, because there are multiple realities that are constructed in each particular environment. Also, in constructivist paradigm one can ask broader questions than in positivist paradigm. In positivist paradigm, a set of specific factors (independent variables) should be linked to the studied phenomenon (the dependent variable). For example, the question may be formulated like this: Do the characteristics of the professor (age, gender, number of publications, tenure, the ability to entertain, etc.) have impact on the grades? While in constructivist/qualitative research the question may be broader, say: Why only a small fraction of professors succeed in teaching English composition? There may be no substantial theory explaining the problem, and tested variables may seem exhausted. Thus, variables are not known beforehand, and are created (emerge) during the process of investigation, through hermeneutic/dialectical interaction between and among investigator and respondents. Through in-depth interviews and discussions, analysis of documents and texts, a particular constructed world is interpreted: certain meaning is attached to particular actions (phenomena) and the relationships between these new categories is examined. Thus, the link between the investigator and the investigated is interactive—the categories are created and examined in their hermeneutic/dialectic interaction. The research in this approach does not only test theories, but also generates new theories, which are usually aimed at understanding and reconstructing of ''local'' knowledge, rather than explaining a generalizable behavior. In fact, some argue that qualitative research in this perspective should be judged in terms of the range of its variations rather than generalizability (Strauss and Corbin, 1990). The criteria for evaluation in constructivist paradigm are also different—a new cluster of authenticity criteria is introduced. Thus, what perhaps can be labeled as moral sentiments by positivist researchers—issues of fairness, enrichment, empowerment—are explicitly articulated.

The classification above is neither exhaustive nor final. While positivism is fairly accurately represented, the definitions of other schools of thought are still a subject of controversy in social science literature. Even positivism is not an obvious creed that many researchers subscribe. For example, A. Michael Huberman and Matthew B. Miles—perhaps, the best-known ''positivistic'' qualitative researchers, see themselves as ''realists'' or ''transcendental realists,'' rather than positivists (Huberman and Miles, 1994). With other schools of thought, the consensus is even less. Some may see, for example, the constructivist paradigm (which is also known as naturalistic paradigm—by its drive of conducting research in natural environment, without manipulation or experiment) as a version of postmodernism (e.g., in Rosenau's (1992) terms

they will be ''affirmative postmodernists''). It can also be argued that postpositivism is less ''positivistic'' than it is portrayed here and is rather pragmatic (in philosophical sense) and multimethodological, with inclination to employ different methods when necessary to better understand the situation and complement the research conducted in a different mode.

Qualitative research methods were employed much earlier than social sciences acquired any coherent set of principles that could be dubbed as paradigm. It started with Europeans' desire to study other, often exotic cultures since the middle ages (Vidich and Lyman, 1994) and philosophically can be traced back to Kant's revival of Aristotelian idea of distinguishing between theoretical and practical knowledge (Hamilton, 1994). According to Kant, practical knowledge is a field ''governed by autonomous principles which man prescribes to himself'' (as quoted in Hamilton 1994: 63), and thus, knowing how the world works is different from how one makes decisions as what to do about it. While positivism in qualitative research is never dead (e.g., Miles and Huberman, 1994, 1984), qualitative research is increasingly come to be seen as more interpretive, geared more towards understanding than explaining. Denzin and Lincoln (1994) examine the development of qualitative research methods from the beginning of this century and document gradual retreat from positivism to more interpretivist and multi-alternative state of affairs. Qualitative research methods started as tools of inquiry within the positivistic paradigm. In fields were quantitative methods were inappropriate qualitative methods were employed to explain reality. The starting traditional period, for example, has been called by R. Rosaldo (1989) as the period of Lone Ethnographer—a larger-than-life figure that went into distant lands and brought stories about exotic people, and who operated on four terms and commitments: a commitment to objectivism; a complicity with imperialism; belief in monumentalism (creating museum-like pictures of cultures); and belief in timelessness (Denzin and Lincoln, 1994). Soon after there were attempts to make qualitative methods as rigorous as possible, including use of simple statistics. By the mid-70s, with more and more serious defeats of positivism and its various brands, and increasing popularity of newer trends (e.g., phenomenology, hermeneutics, semiotics, poststructuralism), given the ability of qualitative methods to work with much richer and more holistic data that positivism was failing to explain, qualitative methods began to be more explicit in their interpretivist leanings. Based on an argument by Clifford Geertz (1973) stipulating that by social scientists turning to humanities for models and theories (e.g., semiotics, narrative analysis) boundaries between the social sciences and humanities have become blurred, Denzin and Lincoln (1994) call the period from 1970 to 1986 as the period of blurred genres. Since the mid-eighties, postmodernism has deconstructed and questioned every major assumption inherent in research (gender bias, ethnic bias, colonialist bias, political bias, historical bias, etc.) as well as the ability of qualitative researchers to capture lived experience and, respectively, represent it. This period is labeled by Denzin and Lincoln (1994) as ''crisis of representation.'' One can argue that whereas quantitative methods and positivistic inquiry have by and large ignored the challenges of postmodernism, qualitative research methods on the other hand, have gone to another extreme reflecting every major contradiction of postmodernism, which, if taken to extreme, can undermine every enterprise of scientific inquiry. For the skeptical school of postmodernism, for example, there is no truth, there is a demise of subject, death of the author, and all that is left is play, the play of words and meaning (Rosenau, 1992). And finally, Lincoln and Denzin (1994: 576) define the present as ''the fifth moment,'' where six fundamental issues continue to torment the ''interdisciplinary, transdisciplinary, and sometimes counterdisciplinary'' field of qualitative research: (1) critique of positivism; (2) crisis of representation; (3) crisis of legitimation; (4) ''the continued emergence of a cacophony of voices speaking with varying agendas;'' (5) shifting scientific, moral, sacred and religious discourses that shape qualitative research; and (6) influence of technology.

Qualitative methods are increasingly becoming more and more interpretivist, relativist and

constructivist, and the term increasingly means attitude and substantive focus rather than specific, non-quantitative techniques. Still, one cannot claim in the field of qualitative research methods postmodernism or constructivism reigns. For example, Miles and Huberman's (1984, 1994) ''realist'' sourcebook of qualitative methods is very popular. Huberman and Miles (1994) give the best rationale, that I am aware of, for ''cohabitation'' of realist (i.e., positivist) and constructivist approaches to qualitative research, and identification of qualitative research as a field not determined by epistemology. They argue that ''there do appear . . . to be some *procedural commonalities*, in the sequential process of analyzing, concluding, and confirming findings in a field study format. . . . the researcher shifts between cycles of inductive data collection and analysis to deductive cycles of testing and verification'' (Huberman and Miles, 1994: 438).

IV. QUALITATIVE-QUANTITATIVE DICHOTOMY: COMPLEMENTARY OR CONTRADICTORY?

A recent attempt by three prominent political scientists to lay down principles of qualitative social inquiry is rather positivistic (or postpositivistic in Guba and Lincoln's terminology), since they emphasize such concepts as causality and sample size, and insist on situating every research inquiry in the framework of a broader theory in order to test for generalizability—i.e. bring the issues of internal and external validity to qualitative research (King et al., 1994). Qualitative research is seen as research based on *in-kind* rather than *in-degrees* differences (Caporoso, 1995: 457). Thus, qualitative variation (differences across categories such as types of government) is not a variation of magnitude as quantitative variation is (differences across the quantities of the same variable, such as income). King and his collaborators (1994) argue that quantitative and qualitative research have the same underlying logic of inference. While valuing the role of interpretation in clarifying and defining concepts and idea and hypothesis generation. King et al. (1994: 39) argue that for evaluation of these hypotheses ''the logic of scientific inference is unsurpassed.'' Thus, the argument follows, after establishing new concepts and ideas, one should be able to test their validity and answer to such technical questions as: How many observations are enough for valid inference? How do you measure and improve data when you have already established the concepts? How valid is generalization? Perhaps, the question can be formulated as: How do you build an empirically sound theory on the basis of small number of often unique observations? Such an argument brings one to an inevitable question of compatibility of qualitative and quantitative research methods.

First, there is the purist, or epistemological position. Because qualitative research is in the domain of the constructivist (or any other non-positivistic) paradigm, it will be absurd to mix it with positivist quantification. What is the point of testing external validity of a constructed concept that exists only for a small group of people? Say, what is the meaning to generalize the social experience of people suffering from some rare disease in Amish community in Pennsylvania to the population of US? Or, how can you compare meanings that are attached to public space by adolescent Navajo Indians with that of teenagers from New York's Upper East Side? Or, are there any lessons to be learned from, say, Manhattan project if it was one of a kind and will never recur? This posture does not deny right of existence to quantitative methods—rather, it points that they are answering different questions by employing different logic. According to this view, naturalistic research is more preferable, because it gives more holistic picture through ''thick descriptions'' of local situation, whereas as quantitative-positivistic methods as if dissect reality and through quantification decrease the complexity contained in data.

Another argument against combining of two research methodologies is practical. The proponents of this view do not argue that quantitative and qualitative methodologies are epistemo-

logically incompatible. Rather, they argue that qualitative research produces complete and practical knowledge by itself, and while its results later can be used in quantitative study (say, to test generalizability), there is absolutely no need to make qualitative researchers to engage in both types of research for the same program (Morse, 1996). Very few researchers are trained to employ both methods, and as a result, they are either going to do poor job or hire someone to do the missing part.

Researchers arguing for complementary nature of the research are coming from different perspectives. Some point to the fact that quantitative research is not ''purely'' empirical, rather, they argue, even the most rigorous quantitative research, along with quantification, uses interpretation as well. Herbert Kritzer (1996), for example, identifies three levels of interpretation in quantitative research. First level is the interpretation of statistical data, when statistical measures are explained (e.g., what is R^2? Is there a lowest threshold for explained variance? How good is satisfactory?). This is often achieved through ''some type of analogy to a machine-like process based on Newton's action-reaction third law of mechanics. One such analogy is to a rigid lever: as one variable changes the other variable changes in the same general way that one end of the lever moves as the other end is moved (albeit in opposite directions)'' (Kritzer 1996: 5). Kritzer (1996: 6) argues that whereas ''the experienced data analyst backs off from the simple analogy to recognize the ambiguities of causation, and to introduce the stochastic component,'' without such an analogy ''most first order interpretation would be extremely difficult.'' Second level is the use of statistical results to identify ''problems'' in the data and analysis (e.g., what is indicated by regression coefficients that are large in absolute terms, but have the ''wrong'' sign and fail to achieve statistical significance? Does that indicate no link, or opposite relationship? Is it a result of collinearity?) Another focus of second-order interpretation is ''that of recognizing how specific features of the data can influence statistical results in ways that are not closely tied to the substantial theory'' (Kritzer, 1996: 8). For instance, regression results may be significantly altered by small number of extreme outliers. While tools like regression diagnostics can be useful for detecting such influence, the range of diagnostic procedures is too broad, and researchers selectively choose specific procedures based on their interpretation of initial data. Another type of second-order interpretation ''arises from recurring patterns in data that have roots in substantial theory,'' where ''knowing what kinds of patterns to look for involves learning how to interpret data within a given substantive context'' (Kritzer, 1996: 9). For example, in aggregate state-level analysis controlling for region—an ''intervening variable that has recurring influence across a range of substantive questions,'' will significantly enhance the results of analysis. Third level is connecting the statistical results to broader theoretical patterns, which is closely tied to contextual elements such as substantive theory, data collection/generation, and side information available to analyst. Based on complimentary arguments that interpretation is always political, regardless of the object of interpretation, and that interpretation is a problem of language and communication, even if the language is mathematical in form, Kritzer concludes that the lines between quantitative and qualitative social science are less clear than often presumed.

Another view is that it is not the methods that matter, but the paradigms behind the methods (Olson, 1995). Though, as opposed to majority of qualitative researchers, this argument sees qualitative methods more as different techniques, rather than strategies of inquiry or epistemology and does not see any way of resolving differences between competing epistemologies (it is a matter of belief). For example, one can use interviews (a qualitative technique) in a controlled experiment (positivistic design). Whether interviews will be seen as representing particular, contextual beliefs of interviewees or will they be seen as representing the objective truth that can be generalized is a matter of one's subscription of certain epistemology and does not depend on the value-neutral tool of interview. The researchers should be aware of the choices they

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

are making and understand their standing on epistemological and methodological issues. Thus, argument is for more reflexive research, and it is urged to constantly question epistemology and ontology of the research and the biases of the researcher rather than rigidly follow to certain pre-established principles.

On the other hand, there are King et al. (1994) whose argument can be paraphrased that qualitative research produces hypotheses, but for science you also have to test them for internal and external validity, so that it will be possible to infer general conclusions. A single case study, for example, can contribute to theory greatly if disproves (falsifies) the predominant theory. As such, this single observation is becoming an observation in a larger data set that the theory is drawn from. For example, Arendt Lijphart's *Politics of Accommodation* (1975)—a study of single country—falsified what was called pluralist theory by David Truman and others (King et al., 1995). By showing that the Netherlands had deep class and religious cleavages, relatively few of which were cross-cutting and at the same time, it was an especially stable and democratic nation, Lijphart (1975) falsified the pluralist theory stating that cross-cutting cleavages increase the level of social peace and stable democratic government. In this case, King and his collaborators argue, the single case was useful, because it was a part of research program, and it was compared against other observations (perhaps gathered by other researchers). Arguing from positivist (postpositivist) paradigm, they maintain the ''positivist'' criteria for goodness of research (internal and external validity, reliability and objectivity) and do not mention alternative criteria proposed for qualitative research (e.g. credibility; transferability; dependability and confirmability, and empathy proposed for constructivist paradigm). Although they leave room for qualitative research (hypotheses generation and concept definition and clarification), their attempt still may be seen as emphasizing ''the third part of scientific inquiry, the rigorous testing of hypotheses, almost to the exclusion of the first two—the elaboration of precise models and the deduction of their (ideally, many) logical implications—and thus point us to a pure, but needlessly inefficient, path of social scientific inquiry'' (Rogowski, 1995: 467).

A modified version of this argument is basically the approach espoused by most of applied researchers. This view sees qualitative and quantitative methods fully compatible and answering to different questions. Qualitative research is not limited to academe only or only to education, psychology, nursing, anthropology and related sciences. Its appeal is much wider—for example, it is used in such a ''practical'' discipline as marketing. Marketing research represented, for instance, through a specific medium—Internet, is often approached from hands-on, how-to-do perspective rather than from epistemological standpoint. A World Wide Web page on marketing research, for example, simply poses the question ''why?'' for qualitative research and ''how much?'' for quantitative research, adding that not only methods are different, but also the answers (Urban Wallace and Associates, 1995). Another site on marketing research (Qualitative Research, 1996) lists the following reasons for qualitative research:

1. preliminary exploration (getting a feel for market), where there is insufficient information for quantitative research and the results are too confusing;
2. sorting and screening ideas;
3. exploring complex behavior (probing into concealed and unconscious motives and attitudes):
4. explanatory models of behavior (not only correlation, but causality);
5. experiencing the world as consumers see it;
6. using consumers to develop innovations.

Others bring examples how quantitative analysis has enriched the results of qualitative research in social science (e.g., for the topic of ''evil eye'' in immigrant Greek community in the US) (Nau, 1995); or how qualitative analysis have helped to ''salvage'' quantitative analysis

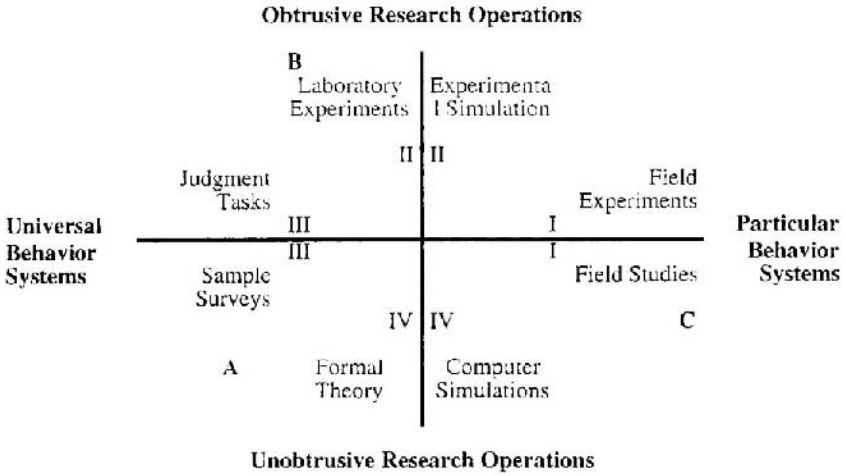
(topics were selected from teaching) (Weinholtz et al., 1995). This trend, although not always articulated, rests on philosophical tradition of pragmatism, which has more practical and tolerant view of diversity in scientific inquiry.

My own inclination will be to treat qualitative and quantitative research traditions as complementing and enriching each other. The task of the researcher should be open-minded consideration of all research alternatives suitable for particular problem at hand and reflexive process of analysis whereby the analyst constantly questions his or her personal bias and the ontology and epistemology of the research inquiry (a term preferred by many qualitative researchers over term ''design'') along the examination of data and theory testing. Interpretation exists in every form of human thought and researchers ought to be aware of this fact—whether conducting positivistic, naturalistic, quantitative or qualitative research. Qualitative research is an umbrella cross- and inter-disciplinary term, unifying very diverse methods with often contradicting assumptions, which defies simple definition. Although qualitative research tends to be explicitly interpretive and more suited for certain tasks (establishing meanings, clarifying concepts and proposing hypotheses) and quantitative research more empirical and suited for other purposes (testing hypotheses), the demarcation lines between qualitative and quantitative domains of research are not very clear, and neither of them has inherent privilege over the other. One should be aware of biases and incoherent definitions of the term ''qualitative research'' and always clarify the meaning of the term in concrete context before deciding on its application—whether the term denotes epistemology, strategy of inquiry or a specific technique. Concrete requirements of situation should decide application of particular research methods and techniques. Following brief descriptions of some widely used methods of qualitative inquiry are intended to sketch their logic for introductory purposes only.

V. STRATEGIES OF QUALITATIVE INQUIRY

Before proceeding with a brief description of methods of qualitative inquiry, it will be useful to give some overview of research methods in social sciences. Perhaps the most comprehensive classification of research strategies in behavioral science has been given by Joseph McGrath. McGrath (1981: p. 179) conceptualizes the research process as ''dilemmatic,'' i.e. as ''*series of interlocking choices*, in which we try *simultaneously to maximize several conflicting desiderata*,'' and draws ''sharp distinction between: (a) strategies or research settings for gaining knowledge: (b) plans or research designs for carrying out studies: and (c) methods of research techniques for measuring, manipulating, controlling and otherwise contending with variables.'' Based on this, McGrath distinguishes eight research strategies along two continua: obtrusive-unobtrusive research operations and universal-particular behavior systems. None of the research methods maximizes more than one of the three conflicting goals of behavioral research: precision, generality, and concreteness or faithfulness to a real situation (Figure 1).

Most of qualitative research fits what this classification calls field studies, with its primary concern being faithfulness to the situation. Despite numerous attempts of classification of qualitative research methods, there is no general consensus on the boundaries and contents of qualitative research methods, with some researchers even arguing against fruitfulness of attempts to bring taxonomy to such a diverse field (Atkinson et al., 1988). Diesing (1971) identified four traditions under the rubric of *case study methods*: (1) participant observation; (2) history; (3) case history; and (4) clinical research. Arguably the best-known (perhaps, also the most-contested) taxonomy of the methods is provided by Evelyn Jacob. Jacob (1987, 1988) identified six major domains of qualitative research in education:



I. Settings in natural systems.
II. Contrived and created settings.
III. Behavior not setting dependent.
IV. No observation of behavior required.

A: Point of maximum concern with generality over actors.
B: Point of maximum concern with precision of measurement of behavior.
C: Point of maximum concern with system character context.

FIGURE 1 McGrath's classification of research strategies (McGrath, 1981:183). (Reprinted by permission of Sage Publications, Inc.)

1. *Human ethology*, which seeks to understand the range of behaviors in which people naturally engage, by observing and quantitatively analyzing the data.
2. *Ethnography of communication*, dealing with ''patterns of social interaction among members of a cultural group or among members of different cultural groups'' (Jacob 1987: 18). Based on their participant observation, as well as audio- and video-recordings, these ethnographers analyze verbal and non-verbal interactions of the members of the group they study.
3. *Ecological psychology*, which emphasizes the interaction of the person and environment in shaping behavior. Ecological psychology ''relies on observational data, supplemented with specimen records,'' with its objective being ''describing these behaviors and analyzing the influence of environment on them'' (Marshall and Rossman 1995: 2).
4. *Holistic ethnography*, which studies ''the culture shared by particular bounded groups of individuals'' (Jacob 1987: 11). The main tool here is participant observation, and the main aim is to reveal and document the perspective of ''Others''—bearers of a different culture.
5. *Cognitive anthropology* studies the ''system for perceiving and organizing the world'' (Jacob 1987: 23) that is unique for each bounded group of individuals. Data are gathered through in-depth interviewing and later are classified into cognitive categories of meaning that are systematically linked to each other.
6. *Symbolic Interactionism* studies how ''interpretations are developed and used by indi-

viduals in specific situations of interactions'' (Jacob 1987: 27), or how individuals ''take and make meaning'' (Marshall and Rossman 1995: 2) in social organizations. Initially developed by H. Bloom (1969), its main tenets are also present in *interpretive interactionism* (Denzin 1989).

The taxonomy proposed by Jacob was criticized by British researchers as ethno-centric and wrongly grounded in ''Kuhnian models'' (Atkinson et al., 1988). They identify seven approaches to qualitative research: (1) symbolic interactionism; (2) anthropology; (3) sociolinguistics; (4) democratic evaluation; (5) neo-Marxist or critical ethnography; (6) ethnomethodology; and (7) feminist research. Marshall and Rossman (1995) add action research or participatory action research to this list. Morse (1994) discusses five main types of qualitative strategies: (1) phenomenology; (2) ethnography; (3) grounded theory; (4) ethnomethodology; (5) qualitative ethology. Janesick (1994) mentions eighteen possible research strategies (with a reminder that the list is not all-inclusive), among which are mentioned such categories as oral history; microethnography; literary criticism, etc. The comprehensive *Handbook of Qualitative Research* (1994) has chapters devoted to: (1) case studies; (2) ethnography and participant observation: (3) phenomenology, ethnomethodology and interpretive practice; (4) grounded theory methodology; (5) biographical method; (6) historical social science; (7) participative inquiry; (8) clinical research. Some authors see methods of inquiry being independent of data collection methods, others claim that qualitative research strategies often define particular data collection strategies (Marshall and Rossman 1995: 40). There is also a little confusion about methods—some authors see often mentioned research strategies as topics of study rather than methods. For example. Robert E. Stake (1994: 236) sees cases studies not as ''methodological choice, but a choice of object to be studied,'' adding that case studies can be both quantitative and qualitative. The same logic obviously applies to biographical method—it is more of an object of study rather than method for studying a phenomenon, though some biographies may give more insight to particular phenomena than scores of rigorous studies. Feminist studies similarly are bound not by methodology but by their focus and sensitivity to women's issues—they can fit in cultural studies, ethnography, ethnomethodology, etc. Some of the above-mentioned research strategies may be expressed in terms that make them too discipline specific (e.g., democratic evaluation seems education-specific, while clinical research seems better fitted for health-related sciences and social work). In this diversity of approaches, I will limit my brief discussion to four main types of qualitative research, that in my view, present the most interest for generic public administrators (assuming there is such a phenomenon). I will discuss general process of qualitative research, the nature of case studies and give brief descriptions of (1) ethnography and participant observation; (2) phenomenology; (3) grounded theory; and (4) action research; emphasizing the grounded theory approach as the most generic and systematic theory-generation strategy.

Before proceeding with description of different qualitative research methods strategies it is important to identify the steps of qualitative research inquiry or research design. Janice Morse (1994a, 1994b) provides us with useful insights. Generally, Morse (1994b) argues, qualitative research consists of four processes:

1. *Comprehending*—i.e., ''learning everything about a setting or the experiences of participants.'' It is over when ''the researcher has enough data to be able to write complete, coherent, detailed, and rich description'' (1994b: 27). When overlayed on concrete research design, this process parallels data gathering (e.g. conversation and dialogues in phenomenological analysis).
2. *Synthesizing*, which is ''the merging of several stories, experiences, or cases to describe a typical, or composite pattern of behavior or response'' (1994b: 30). The

equivalent of this process in research design will be the actual method employed (e.g., content analysis and saturation of categories in ethnography).

3. *Theorizing*, which is ''the process of constructing alternative explanations and of holding these against the data until a best fit that explains the data most simply is obtained'' (1994b: 33). In research design this will be the phase of laying down the end result of the research—i.e. connecting specific phenomena in the study (e.g., for ethnoscience it will be developing linkages between categories and putting them in a taxonomy).
4. *Recontextualization* is the development of emerging theory so that the theory is applicable to other settings and to other populations to whom the research may be applied'' (1994b: 34). This process will be adequate for generalization of the results of particular research in abstract terms (e.g., development of substantial and formal theory in grounded theory approach). After identifying these cognitive processes, Morse argues that the ''way each process is applied, targeted, sequenced, weighed, or used distinguishes one qualitative method from another and gives each method its unique perspective'' (1994b: 34).

Morse (1994a) is more specific in her presentation of funded qualitative research design. She identifies the following phases of developing qualitative research design:

1. The stage of reflection:
   A. identification of the topic
   B. identifying paradigmatic perspectives
2. The stage of planning:
   A. selecting a site
   B. selecting a strategy
   C. methodological triangulation
   D. investigator preparation
   E. creating and refining the research question
   F. writing the proposal
3. The stage of entry:
   A. Sampling
   B. interview techniques
4. The stage of productive data collection/analysis:
   A. data management techniques
   B. ensuring rigor of the data
5. The stage of withdrawal
6. The stage of writing

Marshall and Rossman (1995) address eight major topics in qualitative research design: (1) the overall approach and rationale; (2) site and sample selection; (3) the researcher's role; (4) data collection techniques; (5) data management; (6) data analysis strategy; (7) trustworthiness features; and (8) management plan on time line.

I will touch briefly upon main stages of research, paying more attention to research method selection, data collection and management, and criteria for evaluation. First of all, there is the issue of selection of the topic. Some authors classify sources of such a decision—e.g., personal experience, discrepancy in literature and research findings, assignment, etc. (Strauss and Corbin, 1990, Morse, 1994a). As Morse (1994a: 221) mentions; the important issue here is not where the topic comes from, but the researcher's awareness of his or her motives that may result in a bias when studying particular phenomenon. Reflection upon researcher's epistemological paradigm, as well as on his or her ''posture'' in qualitative research are also important stages

of research process. Wolcott (1992) identifies three ''postures'' in qualitative research: (1) theory-driven (i.e., based on certain broad theory, such as cultural theory in ethnography); (2) concept-driven (i.e., based on certain concept within a theory, such as the concept of care in clinical ethnography); and (3) reform-focused (i.e., political project with predetermined goals, such as feminist research) (Morse 1994a: 221).

The next important topic the researcher has to deal with is site selection. It is important to discuss in this stage the concept of case study. As Robert E. Stake (1994) convincingly argues, case study is not a methodological tool, but a choice of object to be studied. Cases can be studied from multiple perspectives, employing different methodologies (e.g., phenomenology, clinical research). Stake (1994) identifies three main types of cases:

1. *Intrinsic case study*. Particular case is studied because researcher's inherent interest in the case. For example, a study of a tragedy like the Challenger disaster may be undertaken because the researcher is interested to learn not only why organizations malfunction, but also why this particular disaster happened.
2. *Instrumental case study*. Here a particular case is examined to ''provide insight into an issue or refinement of theory.'' For example, the already mentioned study of social cleavages in the Netherlands by Lijphart (1975) is a case that refuted or refined the prevailing ''pluralistic theory,'' which claimed that cross-cutting cleavages are essential for social peace and democratic institutions.
3. *Collective case study*. Here the researchers study a ''number of cases jointly in order to inquire into the phenomenon, population, or general condition.'' This is basically, the instrumental case study extended to several cases. For example, Henry Mintzberg's classic *The Nature of Managerial Work* (1973) is such a study. Realizing that real-life managers hardly engage in what they are supposed to engage according to the literature (Henry Fayol's four functions—planning, organizing, coordination and control; or POSDCORB), Mintzberg ''shadowed'' five executives in five different types of organizations (public, private, and nonprofit), recording their every activity. Based on research, Mintzberg identified 10 activities that managers engage in.

Noting that these three types are ''heuristic more than functional,'' Stake identifies two more, differential types of cases: (1) *teaching case*, which results from instrumental case study; and (2) *biography*. There are always features that make a particular case unique. Stake (1994: p. 238) identifies the following features on which the researchers should gather information: (1) the nature of the case; (2) its historical background; (3) the physical setting; (4) other contexts, including economic, political, legal and aesthetic; (5) other cases through which this case is recognized; (6) those informants through whom the case can be known. Case studies can be quantitative and qualitative, positivistic, and non-positivistic. Site (case) selection is especially important if one holds to positivistic paradigm, because for one best explanation it is instrumental to have the case that provides the richest (or essential) information.

Perhaps the most important issue in research design is choosing a particular strategy for pursuing the research topic. Several authors provide heuristic guidelines for choosing research strategies. In selecting research strategy, it is not sufficient to distinguish the mode of control in research and the desire for local or global (universal) knowledge, as proposed by McGrath (1981). Yin (1984) proposes three questions for selecting the soundest research strategy. First, what is the nature of the research question? Second, does the research require control over behavior, or it should be naturalistic? Third, is the phenomenon contemporary or historical? The key issue here is for researcher to identify what is the research question. Researchers should be clear in what aspect of the phenomenon are they interested. Yin (1984) identifies three types of research questions: exploratory, descriptive and explanatory. Marshall and Rossman (1995)

add predictive questions as another type. In this framework explanatory and predictive questions are discriminated by the object of study. The explanatory question is seeking explain the phenomenon under study, while the predictive question tries to study the consequences of the phenomenon. I would add action-oriented and critical questions as other types. Based on three questions (type of question, control and historical nature), Yin (1984) identifies five distinct strategies of research: (1) experiments; (2) surveys; (3) archival analysis; (4) histories; and (5) case studies. Marshall and Rossman (1995) add field studies, ethnographies, and in-depth interview studies. Marshall and Rossman also (1995: 40) find that particular qualitative methods often define data collection methods. Based on this, they propose a heuristic guide for selecting research strategy and data collection methods for specific research questions (Marshall and Rossman, 1995: 41). According to the guide, experimental and quasi-experimental research design is best suited for predictive questions, with qualitative research being more appropriate for other questions, most notably, for exploratory questions. While one can argue about using case studies or ''multiple case studies'' as specific research method as Marshall and Rossman (1995) do, the taxonomy of questions section of the guide is quite useful. The section, adding action-oriented and critical questions, is reproduced below (Table 2).

TABLE 2 Types of Research Questions

Purpose of the study	Research question
Exploratory	
To investigate little-understood phenomena;	What is happening here?
To identify/discover important variables;	What are the salient themes, patterns, and categories in participant's meaning structures?
To generate hypotheses for further research.	How are these patterns linked with one another?
Explanatory	
To explain the forces causing the phenomenon in question;	What events, beliefs, attitudes, and policies are shaping this phenomenon?
To identify plausible causal networks shaping the phenomenon.	How do these forces interact to result in the phenomenon?
Critical	
To uncover implicit assumptions and biases (and structures) on which the predominant argument (narrative) rests.	What are the assumptions about human nature, society, reality, and type of knowledge that define the existing views on the phenomenon? Are they right? Are they fair?
Descriptive	
To document the phenomenon of interest.	What are the salient behaviors, events, beliefs, attitudes, structures, and processes occurring in this phenomenon?
Action-oriented	
To change the phenomenon by educating and mobilizing people involved in it and affected by it.	What events, beliefs, attitudes, and policies are shaping this phenomenon? How the target group (people needing help) see the phenomenon? How can they change it?
Predictive	
To predict the outcomes of the phenomenon;	What will occur as a result of this phenomenon?
To forecast the events and behaviors resulting from phenomenon.	Who will be affected? In what ways?

Source: Adapted and amended from Marshall and Rossman, 1995: 41. Reprinted by permission of Sage Publications, Inc.

Pamela Brink and Marilynn Wood (1989) distinguish three levels of research design, each having two subcategories. According to them, level III research consists of *experimental and quasi-experimental design*, which are used to test theory. The greater the control in the experiment, the more reliable are the results. Level II research yields statistical analysis of the relationships between and among variables. The first type of design in this level, *comparative design* is based on prior research findings in the literature, and tests theory without manipulation of the independent variable. *Correlational design*, on the other hand, examines the relationship between two or more variables when no previous research findings support a prediction of cause and effect. As in level III, these designs are mostly quantitative. Level I research is exploratory-descriptive. *Descriptive design* is used to ''describe a single variable or population completely, accurately, and thoroughly (Brink and Wood, 1989: 21). Descriptive research can use both qualitative and quantitative studies. And finally, the central purpose of *exploratory design* is ''to develop valid definitions of a concept, describe a process, or yield beginning theories that explain the phenomenon under study. Data are collected in depth and over time in order to increase the validity of the concept being developed. Consequently, samples are usually quite small (from one to twenty). Data are most frequently collected by means of qualitative field study subjected to inductive analysis'' (Brink and Wood, 1989: 21).

Assuming for a moment for simplicity that qualitative research is confined to exploratory, critical and descriptive questions (or designs), there is still one question looming. What strategy (or method or tradition) of qualitative research to use if one wants to explore a particular topic? Phenomenology? Ethnography? Grounded theory? Can they be triangulated? Because of possible multiple combinations of research questions, strategies and techniques, a type of heuristic guide offered by Morse (1994a), I believe, bears a more fruitful approach here. She lists several major types of qualitative research strategies, describing for each strategy typical research questions, the underlying paradigm, methods of data collection and data sources, and finally, showing how each research strategy would have handled a hypothetical example. For the purposes of this article, I will limit my discussion only to four qualitative research strategies, and bring a hypothetical example that is more related to public administration. Below I will draw from Morse's (1994a) descriptions of essential attributes of phenomenology, ethnography and grounded theory, and add similar characteristics for action research (Table 3). Focusing on procedural aspects of research will help us to circumstep discussing epistemology all the time, although, always having the issue of epistemology in perspective.

As Morse (1994a: 223) advises, it is often useful to *imagine* what one wants to find out—''by projecting the research outcome, the researcher may begin to conceptualize the question, the sample size, the feasibility of the study, the data sources, and so on.'' Following her example, let's sketch what our four research strategies would yield in a mock project entitled ''Managing a nonprofit organization'' (Table 4).

For purposes of methodological triangulation, more than one qualitative methods can be used in a research project, provided that the analysis is kept separate and methods are not muddled (Morse, 1994a: 224). Different methods can coexist in other ways, too. For example, critical questions or postmodern deconstruction of existing theories and realities are not only candidates for the primary focus of research. They are a part of every research enterprise that tries to legitimate a new direction in research, and thus, has to deligitimate and pinpoint the shortcomings of the prevailing approach. It is important to realize how the transition from critique to theory-building is going to proceed, and what is the relationship between normative (the result of critical deconstruction) and empirical (the actual collection and interpretation of data) aspects of the argument. One has to show not only that data in general can be interpreted from a new normative vantage point, but that the actual data in research project support such an interpretation. It is also important to remember that the four methods identified above are not all-agreed-

TABLE 3 Comparison of Four Major Types of Qualitative Research Strategies

Strategy	Type of question	Primary method	Other data sources	References
Phenomenology	Questions about meaning—eliciting the essence of experiences.	audiotaped ''conversations,'' written anecdotes of personal experiences; in-depth interviews.	phenomenological literature; philosophical reflections; art; poetry.	van Maanen (1990); Hummel (1994a).
Ethnography	Descriptive questions—of values, beliefs, practices of cultural group.	participant observation; unstructured interviews; field notes.	documents; records; pictures; social network diagrams; maps; genealogies.	Atkinson (1992); Jorgensen (1989).
Grounded theory	Process- and action-oriented questions—experience, interaction over time or change.	interviews (tape-recorded); participant observation; coding	memos; diary	Glaser and Strauss (1967); Strauss and Corbin (1990).
Action research	Action-oriented questions—reflection upon perceptions; implicit models of thought and action.	interviews (audio-and video-recorded); conversations.	observation.	Argyris et al. (1985); Torbert (1991).

Source: Adapted and modified from Morse, 1994a: 224. Reprinted by permission of Sage Publications, Inc.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 4 Comparison of Four Major Types of Qualitative Research Strategies for the Hypothetical Project ''Managing a Nonprofit Organization''

Strategy	Type of question	Participants/informants/ sources	Sample size	Data collection methods	Type of results
Phenomenology	What is the meaning of managing a nonprofit organization?	managers of nonprofit organizations; autobiographies and biographies.	about 5–6 participants or until saturation.	in-depth conversations.	reflective description of the experience of ''what it feels like to manage.''
Ethnography	What are the activities the manager engages in? How does he or she relate to others?	managers, assistants, secretaries, other subordinates, clients, funders, various documents.	about 30–50 interviews or until saturation.	participant observation; unstructured interviews; ''shadowing,'' examination of calendars, organization charts, etc.	description of manager's daily activities, routines, relationships with subordinates and clients.
Grounded theory	What is the essence of specific actions (e.g. is the observed act of writing planning, information processing, or communicating?) In what context did it occur?	managers; secretaries; assistants; clients; diaries; calendars; organization charts; plans; phone bills.	about 30–50 or until saturation.	interviews (tape-recorded); participant observation; coding (open and axial); writing and analyzing memos; drawing schemata.	description of socio-psychological aspects of managing—what action/interaction is more likely under different stimuli and in different situations.
Action research	What are the perceptions of the manager and his/her subordinates about his/her work? How different are they from his/her activities? How the practice can be changed to make it more efficient and just?	the manager and his or her subordinates in one organization; clients.	about 15–25 interviews or until saturation.	interviews (audio-and video-recorded); conversations; observation.	arriving to enlightenment (reflection) for managers so they can detect their shortcomings and take a corrective action.

MARCEL DEKKER, INC. 270 Madison Avenue, New York, New York 10016

upon procedures that are ''carved in stone''—each of them is rather a family of similar methods than a concise methodology.

Phenomenology. Phenomenology is an attempt to reveal the essential meaning of human actions. Originated by Husserl (1970), developed by Heidegger (1972), introduced to social sciences by Schutz (1967), phenomenology has been successfully applied to the study of bureaucracy and public administration (Hummel 1994a). At least two schools of phenomenology can be identified: (1) eidetic (descriptive) phenomenology, based on Husserl's ''transcendental subjectivity''; and (2) hermeneutic (interpretive) phenomenology, based on Heideggerian ontology (Ray, 1994; Cohen and Omery, 1994). Others sometimes distinguish phenomenography as a different branch of phenomenology (Marton, 1994). Very often, phenomenology is grouped under larger group of hermeneutic-interpretive research methods (Diesing, 1991; Holstein and Gubrium, 1994). Still, the very brief description below gives some common features of their methodology.

In phenomenology, comprehension is achieved first of all, by reflecting upon one's own experiences. Then, in-depth interviews and conversations are carried out with subjects, aiming to bring forth experiential descriptions of phenomenon. These conversations are taped, transcripted and thoroughly examined. Descriptive words and phrases are highlighted and studied. Data from other relevant sources (the sources should describe experience and not charts and graphs) can also be used. The principal means for combining data is the process of conducting thematic analyses by identifying common structures of the particular experience (Morse, 1994b: 36). Van Maanen (1990: 101) proposes four ''existential'' guidelines for phenomenological reflection: (1) lived space; (2) lived body; (3) lived time; and (4) lived human relations. The result of phenomenological research is an abstract reflective statement purified through several iterations of writing. Ray (1994: 130) argues that ''affirmation and credibility of phenomenological research can be best understood by Heidegger's (1972) concept of truth as unconcealment and Ricooeur's idea that truth of the text may be regarded as the world it unfolds.'' Perhaps, the other idea she proposes, is more helpful—a researcher ''can recognize that his or her description or interpretation is correct because the reflective process awakens an inner moral impulse'' (Ray, 1994: 130).

Ethnography. Historically originating in the field of cultural anthropology (Vidich and Lyman, 1994), ethnographic approaches to social research have been applied in numerous fields: social and cultural anthropology, sociology, human geography, organization studies, educational research, and cultural studies (Atkinson and Hammersly, 1994: 257). Not easily subdued by any single definition, ethnography and participant observation, perhaps can be understood as the description of some group's culture from the group's perspective. As phenomenology, ethnography is not an agreed-upon precise body of methodology. For example, Boyle (1994), following Werner and Schopfle (1987), discusses four types of ethnographies (classical or holistic; particularistic; cross-sectional; ethnohistorical); as well as ethno-science. Muecke (1994) discusses classical, systematic, interpretive and critical directions in ethnography. Some authors consider ethnomethodology (Garfinkel, 1967) being part of this tradition, while others see ethnomethodology as more hermeneutic practice (Holstein and Gubrium, 1994). There is more agreement on the term participant observation, which is, essentially, the method/technique of ethnography. The terms are very often used synonymously, though are not exactly the same.

One can identify different levels of involvement in participant observation: (1) complete observer; (2) observer as participant; (3) participant as observer; and (4) complete participant (Atkinson and Hammersly, 1994: 248). The following brief statement by Danny Jorgensen (1989: 23) fairly accurately summarizes the essence of participant observation:

> [Participant observation] focuses on human interaction and meaning viewed from the insiders' viewpoint in everyday life situations and settings. It aims to generate practical and theoretical

truths formulated as interpretive theories. The methodology of participant observation involves a flexible, open-ended, opportunistic process and logic of inquiry through which what is studied constantly is subject to redefinition based on field experience and observation. Participant observation generally is practiced as a form of case study that concentrates on in-depth description and analysis of some phenomenon and phenomena. Participation is a strategy for gaining access to otherwise inaccessible dimensions of human life and experience. Direct observation and experience are primary forms and methods of data collection, but the researcher also may conduct interviews, collect documents, and use other methods of gathering information.

Grounded Theory Approach. Because grounded theory is the most recent and systematic approach of theory generation, I will describe it in more detail than the other three research strategies. It was first articulated by Barney Glaser and Anselm Strauss in *The Discovery of Grounded Theory: Strategies for Qualitative Research* (1967). Grounded theory approach shares many features with other types of qualitative research (e.g., sources of data, data gathering, and analyzing techniques, as well as the possible use of quantitative techniques), but the characteristic that sets it apart from others is its explicit emphasis on theory generation. Grounded theory approach is based on constant comparative method, where the evolving theory (i.e., propositions about the nature of relationships between phenomena that are examined) is being iteratively validated against the data (i.e. being grounded in the data) until a substantive theory emerges that relates the concepts, their properties and dimensions in a systematic manner. As the previously described methods, grounded theory also is not an all-agreed-upon research strategy, though disagreements in this approach are of much lesser magnitude than in others. Stern (1994) distinguishes Glaserian and Straussian approaches in grounded theory methodology, first of which, following Glaser (1992), she labels as *grounded theory*, while the second—Straussian method—as *conceptual description*. The most accessible introductory book to grounded theory, I believe, is Strauss and Corbin's *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (1990), from which I will draw the following brief description of the method.

The research question in a grounded theory study is a statement that identifies the phenomenon to be studied. It can come from literature (both technical (i.e., scholarly) and nontechnical), from personal experience, and is oriented toward action and process. The researcher should rely on his or her ''theoretical sensitivity''—the ability to recognize what is important in data and to give it meaning. There are specific techniques that help to enhance researcher's theoretical sensitivity, such as the flip-flop technique (imagining the opposite condition, or turning an observation on its head—e.g., imagining a monopolistic market as perfectly competitive, in order to project how the relationships would change and thus gain insight into the phenomenon) or the far-out comparisons (comparing the examined phenomenon with something totally dissimilar, e.g. comparing violinists and body-builders, in order to elicit insights that otherwise would skip one's mind). The key analytic process in grounded theory is coding. There are two types of coding—*open coding* and *axial coding*. Open coding is ''the process of breaking down, examining, comparing, conceptualizing and categorizing data.'' Drawing from sources like observation and interviews, the researcher first of all conceptualizes the data—i.e., labels the studied actions or attitudes under some conceptual label. This act is not simple description. For example, when a manager is speaking on the phone, it is not simply recorded as speaking, but can be labeled as information passing. Next, concepts are grouped in categories. For example, the concepts of passing and receiving information can be grouped under a category called communicating. After naming the categories are being described in terms of their properties (characteristics) and their dimensions (measurement of properties along some continua). For example, if a manager is engaged in communicating, he or she receives information in the process of that activity. This is a property of communicating. How often does one receive information? In what amount?

From whom (subordinates, outside sources, etc.)? All of these are dimensions that describe the property of receiving information. Also, depending on the question studied, one can ask what part of communicating is information receiving and what part is information passing. Is there any imbalance? The next step in grounded theory approach is axial coding, whereby data are put back together in new ways after coding, by making connections between categories. This is done in the paradigm model, simplified form of which looks like this:

(A) causal conditions → (B) phenomenon → (C) context →
(D) intervening conditions → (E) action/interaction strategies → (F) consequences.

Causal conditions refers to the ''events or incidents that lead to the occurrence or the development of the phenomenon.'' For example, abrupt change of weather may cause a road emergency. Each of these categories should be described along dimensions of their properties. For example, the abrupt change of weather may be characterized by wind, change in temperature, precipitation, fog. Each of these properties may be measured in their dimensions—e.g. the speed and the chilling factor of the wind. Road emergency may have the properties of actual damage to the road because of flooding, low visibility, slippery road, the exact place of the damage. Context refers to the ''specific set of properties that pertain to a phenomenon. It is also the particular set of conditions within which the action/interaction strategies are taken to manage, handle, carry out, and respond to a specific phenomenon.'' In our example, e.g., the context may be managing the road emergency that is a result of (1) heavy rain; (2) flood; it is a (3) damage to a federal highway; (4) damage on an important stretch of the road close to a big metropolitan center; (5) with large stockpiling of cars on the road; (6) with no casualties. Intervening conditions are the ''broad and general conditions bearing upon action/interaction strategies,'' and include: ''time, space, culture, economic status, technological status, career, history, and individual biography.'' Action/interaction has two features: (1) it is processual, evolving in nature; (2) it is purposeful, goal-oriented. Studying failed action/interaction strategies is as important for the grounded theory approach. Action and interaction have certain outcomes or consequences. It is important to remember that the failure to take action/interaction strategies also has consequences. Axial coding is complex analytic process, where four distinct analytic steps are performed simultaneously:

1. ''the hypothetical relating of subcategories to a category by means of statements denoting the nature of relationship between them and the phenomenon''—through the above-mentioned paradigm model;
2. ''the verification of those hypotheses against actual data'';
3. ''the continued search for the properties of categories and subcategories and the dimensional locations of data indicative of them'';
4. ''the beginning exploration of variation in phenomena, by comparing each category and its subcategories for different patterns discovered by comparing dimensions locations of instances of data.''

In grounded theory there is a ''constant interplay between proposing and checking.'' This back and forth movement between inductive and deductive thinking makes the emergent theory grounded. The final theory is limited to actual data. The next step in grounded theory approach is selective coding. This is similar to axial coding, with analysis done on more abstract level. Selective coding is ''the process of selecting the core category, systematically relating it to other categories, validating those relationships, and filling in categories that need further refinement and development.'' When the story line—''the conceptualization of the story'' is explicit and the ''data are related not only at the broad conceptual level, but also the property and dimensional

levels for each major category,'' the researcher has formulated the ''rudiments of a theory.'' And finally, ''validating one's theory against the data completes its grounding.''

Grounded theory methodology is performed by using such tools as memos and diagrams, building conditional matrices, following not random but ''theoretical sampling''—i.e., sampling on the basis of concepts that have theoretical relevance to the evolving theory, etc. I will not focus on these topics, and will conclude this brief description of grounded theory with the issue of criteria for judging a grounded theory study. Strauss and Corbin (1990: 252) mention that while judging a research study, judgments are made about three issues: (1) validity, reliability, and credibility of the data; (2) adequacy of the research process; and (3) empirical grounding of research findings. Because the first two are generally covered in qualitative research literature, they concentrate on the third issue, and offer set of criteria against which a grounded theory study can be judged. Among those offered are questions that specifically focus on the theory-generation aspects of the research: (1) Are concepts generated? (2) Are concepts systematically related? (3) Are there many conceptual linkages and are the categories well developed? (4) Is much variation built into the theory?

Action Research. Action research is a research strategy that studies action with triple goals of: (1) making that action more effective and efficient; (2) empowerment and participation; and (3) developing scientific knowledge. Action research is again a family of methods rather than precise research methodology, and is often covered under the title of participative research. Chein et al. (1948) identified four varieties of action research: (1) diagnostic; (2) participant; (3) empirical; and (4) experimental. Peter Reason (1994) identifies, among others, three main approaches to participative inquiry: (1) cooperative inquiry; (2) participatory action research; and (3) action science and action inquiry. Deshler and Ewert (1995) identify five fields of practice ''that have made contributions to participatory action research approaches'': (1) action research in organizations; (2) participatory action research in community development; (3) action research in schools; (4) farmer participatory research and technology generation; and (5) participatory evaluation. As one can conclude from the above, action research is different research strategy in different environments. Students of public administration are more familiar with three varieties of action research: (1) action research as a form of organizational development (e.g., Argyris et al., 1985); (2) participatory evaluation (e.g., Guba and Lincoln, 1989); and (3) participatory action research in community development (e.g., Whyte, 1991). In this variety of approaches, perhaps, a description of action research approach as applied to one context will suffice for introductory purposes. This is the way French and Bell (1995: 7) characterize action research as applied to organization development:

> Action research is essentially a mixture of three ingredients: the highly participative nature of OD, the consultant role of collaborator and co-learner, and the iterative process of diagnosis and action. The action research as applied in OD consists of (1) a preliminary diagnosis, (2) data gathering from the client group, (3) data feedback to the client group, (4) exploration of the data by the client group, (5) action planning by the client group, and (6) action taking by the client group—with an OD practitioner acting as facilitator throughout the process. Widespread participation by client group members ensures better information, better decision making, and action taking, and increased commitment to action programs. . . . Action research yields both change and new knowledge. . . . New knowledge results from examining the results of the actions. The client group learns what works and what doesn't work.

VI. METHODS OF DATA COLLECTION AND ANALYSIS IN QUALITATIVE RESEARCH

Qualitative research employs a host of techniques for collecting and analyzing data. As Punch (1994: 84) observes, three are central—observation, interviewing, and documentary analysis—

that may be employed across a variety of disciplines. Marshall and Rossman (1995: Ch. 4) identify the following types of techniques, as well as evaluate their strengths and weaknesses across a wide range of criteria: (1) participant observation; (2) interviewing; (3) ethnographic interviewing; (4) elite interviewing; (5) focus group interviewing; (6) document review; (7) narratives; (8) life history; (9) historical analysis; (10) film; (11) questionnaire; (12) proxemics; (13) kinesics; (14) psychological techniques; (15) unobtrusive measures. As we can see, they can be broadly grouped under three categories Punch has identified. Others may specify specific methods of data analysis, like narrative and content analysis (Manning and Cullum-Swan, 1994). Methods of open and axial coding, discussed above, can also be viewed as specific research techniques. As everywhere else in this chapter, considerations of space induce me to impose arbitrary lines on what techniques will be discussed. I will only sketch the contours of: (1) interviewing; (2) observational techniques; (3) textual analysis; and briefly discuss data management and use of computers in qualitative research.

Data Management. First of all, there is the issue of data management. Huberman and Miles (1994: p. 428) define data management as ''the operations needed for a systematic, coherent process of data collection, storage and retrieval.'' One should design the data management system long before the actual data collection starts. Data management and analysis can be significantly enhanced by software for qualitative analysis. Following Levine (1985), Huberman and Miles (1994: p. 430) distinguish five general storage and retrieval functions that should be addressed in the data management system:

1. formatting (physical layout of the materials and their structurization into types of files);
2. cross-referral (linkage across different files);
3. indexing (defining codes, organizing them into a structure, and pairing codes with specific parts of database);
4. abstracting (condensed summaries of longer material);
5. pagination (numbers and letters locating specific materials in field notes).

Interviewing. Interviewing is basically the act (or the art) of asking questions and getting answers. Interviews can be distinguished along three dimensions: (1) type of questions (structured or unstructured or semi-structured interviews); (2) number of interviewees questioned simultaneously (individual or group interviews); and (3) selection of interviewees (random or specialized interviews). The most popular form is random, one-on-one, individual interview, very often, using structured or semi-structured questionnaires. Polls, surveys, and censuses are example of such interviews (Fowler, 1984; Babbie, 1990). One-on-one, face-to-face, in-depth unstructured interviews are often called ethnographic interviews (Fontana and Frey, 1994: pp. 365–366).

There is more than one type of group interviews—focus groups, brainstorming, Delphi technique, etc. (Fontana and Frey, 1994), with focus groups being the most common (Asbury, 1995). Focus groups, in essense, are 6–12 individuals who have some knowledge or experience of the topic the researcher is interested in, and whose thinking on the matter is stimulated and enhanced by group dynamics and interaction. The result is a rich and detailed perspective on the topic that the researcher draws from discussions in several groups. Finally, interviewees can be selected randomly (as for surveys), and selectively, as in focus groups, because they are thought to have greater knowledge of the subject. In this case, the researcher may engage in specialized or elite interviewing (Dexter, 1970).

There may be other classifications of interviews—postmodern, gendered, creative, phenomenological, etc. (Fontana and Frey, 1994; Marshall and Rossman, 1995). Usually, different styles of interviews require different techniques. For example, in structured interviews the researcher should be more neutral, while in ethnographic interview he or she should be more ''involved''—trying to engage in conversation, elicit answers, be empathetic, etc. The answers

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

are not treated as simple texts, but are analyzed in conjunction with respondent's body language, use of interpersonal space, tone of voice, flow of speech, etc. In ethnographic interviews it is important to locate the ''key informants,'' establish rapport with respondents, understand their culture and language (including body language and cultural norms), etc. (Yeager, 1989; Fontana and Frey, 1994). And finally, there are ethical considerations involved in interviewing—issues of anonimity, privacy, consent, etc. (Punch, 1994; Fontana and Frey, 1994).

Observation. Observation ''entails the systematic noting and recording of events, behaviors, and artifacts (objects) in the social setting chosen for study'' (Marshall and Rossman, 1995: 79). Sometimes observation is seen as more general activity, where participant observation also fits. Sometimes, it is classified as different from participant observation by the different levels of involvement of the researcher. Subscribers to this view distinguish the following levels of engagement in observation: (1) complete observer; (2) observer as participant; (3) participant as observer; and (4) complete participant, and call observation only the first one. The important point here is not rigid classification, but researcher's clear understanding of his and her position and possible biases because of that position. For example, the researcher may misjudge some actions, because he or she has ''gone native,'' or there may be an ''observer effect'' (or Hawthorne effect) when the examined group behaves differently because they know they are being watched. If to discount steps that are general for other methods of data collection (e.g., interviewing), such as gaining access to the social setting, establishing rapport with the people in the setting, etc., observation proceeds in two of stages: (1) unfocused and descriptive; and (2) focused, when research questions become clearer (Jorgenson, 1989; Adler and Adler, 1994). Observation works best through triangulation—e.g., having multiple observers or verifying observations with document analysis, interviews, etc. As with interviewing, there are ethical considerations with observation that researchers should be aware of.

Adler and Adler (1994: 382) identify five ''observational paradigms,'' or theoretical and/or research traditions that are clearly associated with observational methods. They are: (1) formal sociology, focusing on structures according to which social interactions are patterned (e.g., Buban, 1986); (2) dramaturgial sociology, which is concerned with how people construct their self-presentations and act according to that in front of the others (Goffman, 1971); (3) studies of public realm, which ''address the issues of moral order, interpersonal relations, norms of functioning, and norms of relating to strange individuals and different categories of individuals'' (e.g., Lofland, 1989); (4) auto-observation (e.g., Douglas and Johnson, 1977); and (5) ethnomethodology, with focus on how people construct their everyday lives (Garfinkel, 1967).

Textual and Artifact Analysis. The third source of data gathering is what Hodder (1994) calls ''mute evidence''—written texts and artifacts. Lincoln and Guba (1985) distinguish between records (texts attesting some sort of formal transaction, such as contracts) and documents (texts created largely for personal reasons, such as diaries). Records (e.g., census records, archival materials) are a widely used source of data in public administration. Artifacts are pieces of material culture that characterize the social setting, like dresses.

There are several methods for analyzing texts. The most common, perhaps, is *content analysis*. The essence of content analysis is basically, deriving numerical measures from nonnumerical texts. Content analysis is often performed to study the dominant themes and trends, say, in research journals in a particular field (e.g., White and Adams, 1994), or, violence in TV programming. Content analysis has several steps (Johnson and Joslyn, 1991): (1) sampling of materials (e.g., research journals); (2) definition of categories (e.g., quantitative and qualitative research); (3) choosing the recording unit (e.g., articles and research notes); (4) system of enumeration of for the content being coded (e.g., qualitative research is recognized as such when it is not mixed with quantitative research, and the article employs only qualitative techniques and methods). The most serious criticism against content analysis is that it neglects the context

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

(Manning and Cullum-Swan, 1994). Other types of document analysis include narrative analysis—examining the form of a narrative in conveying meaning (Manning and Cullum Swan, 1994); and poststructuralist and postmodern deconstruction—the practice of ''taking things apart'' and showing that the meaning of a particular text is indeterminate and can be rendered differently in a different semiotic system (sign system) (Adams, 1994). Artifacts are generally analyzed through situating them in a context and analyzing their function in that social setting (e.g. the role of clothing in showing social status).

Data Analysis. Data gathering continues until researchers achieve theoretical saturation (Glaser and Strauss, 1967; Morse, 1995)—i.e. when the generic features of their new findings consistently replicate earlier ones. Data analysis can be conceptualized as three linked subprocesses: (1) data reduction—i.e., choosing the conceptual framework, research questions, cases and instruments, and further condensing the data by coding, summarizing, clustering, writing up; (2) data display—condensed and organized layout of the data that permits conclusion drawing and/or action taking; and (3) conclusion drawing/verification—i.e., interpreting, drawing meaning from data (Miles and Huberman, 1984, 1994). As Huberman and Miles (1994: 429) point out, ''these processes occur *before* data collection, during study design and planning, *during* data collection as interim and early analyses are carried out; and *after* data collection as final products are approached and completed.'' The process of data analysis is not completed in one decisive step, but is iterative, with consecutive inductive and deductive reasoning in each pattern identification-verification cycle.

Qualitative data analysis is not achieved only through endless hours of abstracting field notebooks. Modern technology has made advances in the field of qualitative research as well. There are many qualitative software tools on the market that significantly enhance the research process. Richards and Richards (1994), after discussing qualitative analysis potential of general-purpose software packages (such as wordprocessors and relational database management systems), classify special-purpose qualitative data analysis into the following categories: (1) code-and-retrieve software (e.g., the *Ethnograph*); (2) rule-based theory-building systems (e.g., the *HyperRESEARCH*); (3) logic-based systems (e.g., the *AQUAD*); (4) index-based software (e.g., the *NUD.IST*™); and (5) conceptual network systems (e.g., the *ATLAS/ti*). *NUD.IST*, which is now distributed by Sage Publications, for example, allows one to code (index) and retrieve units of records (e.g., sentences or paragraphs), write memos about records in dialog boxes that can be easily retrieved with the records, systematically orders codes in trees (or hierarchies), searches for text patterns in documents, systematically relates (compares) different codings, etc. This type of software can be very useful, for example, for conducting a grounded theory research. Computer programs for qualitative data analysis are discussed in the volume edited by Kelle (1995), and discussed and evaluated by Weitzman and Miles (1995). Qualitative research has a proper place on Internet as well. In addition to class curricula, articles on occasional home pages and on-line journals (e.g. *Qualitative Report*), there is now a repository of qualitative data—*QUALIDATA*, that just like *ICPSR*, can be accessed electronically on distance (*QUALIDATA*, 1996). Researchers depositing qualitative datasets for public use should be aware of ethical concerns, such as informed consent and confidentiality, etc.

VII. CRITERIA FOR JUDGING QUALITATIVE RESEARCH

There are no universally accepted criteria for judging soundness and goodness of qualitative research. All discussions on the matter draw from criteria of soundness of mainstream (i.e. quantitative) research—internal and external validity, reliability, and objectivity. Positions are ranging from approach asserting these four criteria being incomplete for qualitative research

(they should be amended and modified) to these criteria being completely inadequate (they should be abandoned and new criteria should be formulated for qualitative research), with practical guidelines fitting somewhere in between. The criteria of judgment reflect epistemological paradigms of the researchers—from pragmatical approach of complementing the accepted ''positivistic'' criteria to denying them at all. Generally, criteria proposed specifically for qualitative research are articulated within non-positivistic paradigm.

Huberman and Miles (1994: 439) propose the following criteria for the goodness of research that focus on procedure: (1) sampling decisions made, both within and across cases; (2) instrumentation and data collection operations: (3) database summary and size, the method it was produced; (4) software used, if any; (5) overview of analytic strategies followed; (6) inclusion of key data displays supporting main conclusions. Some researchers propose the idea of carrying out ''audits'' of the study (Schwandt and Halpern, 1988).

As mentioned above, Strauss and Corbin (1990: 252) argue that while judging a research study, judgments are made about three issues: (1) validity, reliability, and credibility of the data; (2) adequacy of the research process; and (3) empirical grounding of research findings. Concentrating on the third issue, they offer set of criteria against which a grounded theory study can be judged. Among those offered are questions that specifically focus on the theory-generation aspects of the research: (1) Are concepts generated? (2) Are concepts systematically related? (3) Are there many conceptual linkages and are the categories well developed? (4) Is much variation built into the theory?

Lincoln and Guba (Lincoln and Guba, 1985; Guba and Lincoln, 1994) offer the most elaborate criteria for qualitative research. They offer two sets of criteria: trustworthiness criteria (credibility, transferability, dependability, and confirmability); and authencity criteria (fairness, enrichment, education, stimulation to action, and empowerment). Trustworthiness criteria parallel those in positivistic-quantitative paradigm. Credibility is the counterpart of internal validity and is concerned with establishing the ''truth value'' of the study—it should be ''credible to the constructors of the original multiple realities'' (Lincoln and Guba, 1985: 296). In order to achieve this, the researcher should carefully identify the setting of the research, the population, and underlying theoretical framework. Transferability (paralleling the criterion of external validity) denotes the applicability of one set of findings to another context. This is usually problematic in qualitative research, and there are basically two strategies to achieve it: explicitly stating theoretical parameters of research (so that other researchers can decide upon generalizing the approach in their settings) and triangulation of research methodologies. Dependability (paralleling reliability) is the criterion through which consistency in the research is shown—i.e., how the researcher accounts for changing conditions in the phenomena and changes in design. Because of constructivist perspective Lincoln and Guba subscribe, the criterion of dependability is different from positivist understanding of replicability—the social world is always constructed, and thus, replicability is a problem. Confirmability (paralleling objectivity) should show neutrality of the research. Here the emphasis is moved from the researcher and placed on data. The criterion is: ''Do the data help to confirm the general findings and lead to the implications?'' (Marshall and Rossman, 1995: 145). The researcher can never eliminate his or her bias, but should build in strategies to balance for balancing bias in interpretation, like playing devil's advocate for research partner, constant search for negative instances, etc. (Marshall and Rossman, 1995: 145–146). The second set of criteria Guba and Lincoln (1989, 1994) present are those of authenticity. These include fairness, ontological authenticity (enlarges personal constructions), educative authenticity (leads to improved understanding of constructions of others), catalytic authenticity (stimulates to action), and tactical authenticity (empowers action).

There are two more aspects the researchers should pay attention to when designing and judging qualitative research—the questions of ethics and ''the art and politics of interpretation.''

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

As Punch (1994: 89–90) argues, three developments have affected the ethical dimension in research. First, ''the womens' movement has brought forth a scholarship that emphasizes identification, trust, empathy, and nonexploitive relationships.'' Second, ''the stream of evolutionist and interventionist work, or ''action'' research, has developed to a phase where ''subjects'' are seen as partners in the research process.'' And finally, with politicization of these issues, ''the concern with harm, consent, confidentiality, and so on has led some government agencies to insist that financing of research be contingent upon an ethical statement in the research proposal and that academic departments set up review and monitoring bodies to oversee the ethical component in funded research.'' The ethical issues that Punch (1994) discusses include ''informed consent,'' deception, privacy, harm, identification, and confidentiality, etc. The researcher should be aware of all of these issues in the context of research project, and make sure that he or she follows the established codes of conduct.

Discussing ''the art and politics of interpretation,'' Denzin (1994) holds that ''the age of putative value-free social science is over.'' Accordingly, he asserts, ''any discussion of this process must become political, personal and experiential.'' Whether subscribing to this view or not, one must be aware of the tendency that Denzin predicts—proliferation of ''race-, ethnicity-, and gender-specific'' interpretive communities and epistemologies, because an important characteristic of research is how its findings are communicated—to scholars, to government, to communities, and individuals. Especially for action-oriented interventionist research it is very important to tell stories that ''subjects'' or partners may be willing to listen.

Marshall (Marshall and Rossman, 1995: 146–148) presents more practical checklist of 20 questions helped to judge the quality of qualitative research. Although not necessarily applicable to all research situations, these guidelines give a good understanding of the criteria employed to judge qualitative research. With some abridgment, they are as follows:

1. The method is explicated in detail so that a judgment can be made about method's adequacy.
2. Assumptions and biases are expressed.
3. The research guides against value judgments in data collection and analysis.
4. There is evidence from raw data demonstrating the connection between the findings and the real world; and it is done in accessible and readable manner.
5. The research questions are stated and answered, and answers generate new questions.
6. The relationship with previous research is explicit, and the phenomena are defined clearly.
7. The study is accessible to other researchers, practitioners and policymakers.
8. Evidence is presented that the researcher was tolerant to ambiguity and strive to balance his or her biases.
9. The report recognizes limitations of generalizability and helps the readers to find transferability.
10. It should be a study of exploration, and not reasserting theories from literature.
11. Observations are made of a full range of activities over a full cycle of activities.
12. Data are preserved and available for reanalysis.
13. Methods are devised for checking data quality (e.g., informants' knowledgeability, ulterior motives) and for guarding against ethnocentric explanation.
14. In-field work analysis is documented.
15. Meaning is elicited from cross-cultural perspective.
16. Ethical standards are followed.
17. People in the research setting benefit some way.
18. Data collection strategies are the most adequate and efficient available. The re-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

searcher is careful to be reflexive and recognize when he or she is ''going native:''

19. The study is tied into ''the big picture.'' The researcher looks holistically at the setting to understand the linkages among systems.
20. The researcher traces the historical context to understand how institutions and roles have evolved (Marshall and Rossman, 1995: 146–148; Reprinted by permission of Sage Publications, Inc.).

VIII. QUALITATIVE RESEARCH METHODS AND PUBLIC ADMINISTRATION

The arsenal of research methods of public administration has been influenced by research methods in political science and economics, both of which are more concerned with generalizations, usually operate with aggregate data and have more or less established beliefs about human nature and motivation (in most of cases that is a rational person).[4] Subsequently, there is an overwhelming dominance of quantitative research methods as tools of inquiry in university curricula, although quite an impressive share of the theory-generation in the field has been achieved through non-quantitative methods—usually a case study or deductive reasoning (often speculation) based on non-structured or incomplete data. As opposed to political science which is more concerned with the role of institutions in the society, public administration has also a micro-focus—the study of organizational life, a focus that it shares with sociology, anthropology, and psychology. Although there is increasing quantification in this direction of research, most insights are still coming from traditional nonquantitative studies. Very often there is an interesting gap between the rhetoric and practice of public administration research. Though in rhetoric it is predominantly quantitative and statistics-oriented, in practice it still relies heavily on qualitative research methods. Yeager (1989) calls basically qualitative research strategies and methods employed in public administration ''classic methods'' and documents their extensive use in the field. Whelan (1989) shows that computer-statistics oriented paradigm of research in public administration is dating only since the 1960s.

The first textbook on research methods in public administration dates back to 1940. John M. Pfiffner's *Research Methods in Public Administration* (1940) is a textbook with positivistic approach, but one that is at ease with rudiments of both quantitative and qualitative research strategies, though a little bit skeptical towards the former. Noting that ''the mental processes which lead to scientific knowledge are briefly of two kinds, observation and inference.'' Pfiffner discusses two types of observation (bare observation and experiment) and two types of inference (induction and deduction). Then he discusses science—''search for rules which govern orderliness in phenomena,'' through analysis and synthesis, formulation and testing of hypothesis. Testing of hypothesis is achieved through ''classification, comparison, and analogy,'' while formulation of hypothesis is a much creative and less structured process, where it is ''legitimate to resort to imagination, supposition, and idealization, even though they may result in barren hypotheses'' (Pfiffner, 1940: 10). In his critique of quantitative methods. Pfiffner does not follow the popular cliché of the time: ''Figures don't lie, but liars figure,'' but holds to a rather balanced view (Pfiffner, 1940: 168):

> Good quantitative work must be based on good qualitative work. This does not mean that no statistical treatment should be attempted until perfection is reached as to the collection of data. That is often impractical and impossible, although it is reasonable to hope for improvement in this respect with each passing year. What is necessary is that the quantitative re-

> searcher realize the limitations of his data and select his hypotheses accordingly. If the data are admittedly crude, one should look for an underlying trend rather than attempt refined treatment.

Some of the techniques Pfiffner describes, such as work flow and charting techniques, personnel classification are not considered (and perhaps, justly so) in the domain of research methods now. But Pfiffner also discusses some topics that are by and large ignored in today's public administration research methods textbooks (e.g., O'Sullivan and Rassel, 1995; Meier and Brudney, 1993)—the human factor in the research process. The chapter devoted to human factor in research ranges from ''handling politicians'' to ''handling the 'hothead''' to ''wangling''—''the use of influence, suggestion, button-holing, politics and expediency to obtain action'' (Pfiffner, 1940: 130). There are also chapters devoted to interviewing, field data studies and biographical method. Pfiffner also pays more attention to the process of research design and planning than modern textbooks. Without doubt, new research methods textbooks are much more sophisticated statistically, and offer better tools for operationalizing research variables, but they miss more practical, people- and organization-oriented research agenda of Pfiffner.

Since the behavioral revolution in social sciences in the 1960s and enormously increased capacity of sophisticated statistical analysis of large amounts of data, positivistic research agenda modeled after natural sciences became the reigning paradigm in social science research in general, and public administration research in particular. Against this force, public administration scholars from time to time tried to reevaluate seemingly perfect procedures of natural sciences and discuss their applicability to public administration. Still in 1940, Pfiffner (1940: 18) wrote: ''The social scientist who feels inferior in the presence of the physicist, chemist or engineer, should remember that a great share of their knowledge is based on accepted practice rather than precise measurement.'' This line of thought in the 1980s and 1990s was pursued with great eloquence by Mary Timney Bailey (1994) and Robert Behn (1992). Discussing experiments as endeavors to control extraneous variables, Bailey (1994: 187) convincingly shows similarities between case studies and experiments: ''The outcome of an experiment, then, is essentially a hypothesis, and each experiment is, in reality, a case study. A set of case studies can be used to challenge dominant theories or for applied research projects in fields (medicine, engineering, etc.) that are derived from ''pure'' disciplines.'' Later she discusses how criteria of scientific rigor can be applied to case studies. Behn (1992: 409) argues that ''nothing better fits our concept of science than physics. Nothing better fits Karl Popper's concept of science than physics. And yet, physics does not fit all that well.'' Discussing how physicists use various empirically non-proven concepts in their theories. Behn urges for the use of adequate metaphors in public administration research. Discussing the ''ultimate physics metaphor''—neutrino, Behn (1992: 111) writes:

> Physicists want the neutrino to exist. It solves a lot of problems. Their research logic, however, is somewhat like observing People dancing in the streets of Boston and concluding that the Red Sox have won the World Series. You might wish the long-elusive Red Sox victory to be the cause of the dancing, but there are always other possible explanations. Physicists both postulate reality and confirm it. Many of the observations that they use to create reality are only indirect. Neutrinos exist—just like the gravity exists—not because they are observed directly, but because something is observed that should happen if neutrinos or gravity exist. The only advantage that neutrinos and gravity have Over other realities—over, say, angels—is that the mathematics that the physicists have invented to go along with these metaphors can be used to make very specific predictions about how other things should behave and that these predictions are confirmed by observations.

Although strikingly resembling some of Milton Friedman's (1953) positivist arguments, Behn achieves something quite different—he manages to convincingly legitimize the explor-

atory, meaning-seeking nature of the research in the field of public administration (Behn, 1992: 418):

> The reality of managerial world is created by those who write about the world. Sometimes these writers are scholars. Sometimes they are practitioners turned scholars. Regardless, those who are most persuasive in their writings—those who use metaphors that others find most evocative of the managerial world they ''know''—define managerial reality. Research in public management—just like research in physics—is a search for meaningful metaphors.[5]

While ''the envy of physics''—the issues of comparison with natural sciences have been raised since the formative years of public administration as science, issues of epistemology in public administration (as in social sciences in general) are being discussed only since the 1970s. Mitroff and Pondy (1973), for example, identify Leibnizian inquiry systems (epitome of deductive, formal reasoning); Lockean inquiry systems (epitome of inductive reasoning); and finally, Kantian inquiry systems, which try to reconcile Leibnizian and Lockean inquiry systems, arguing that scientific observations are not theory-free. Jay White (1994) identifies three approaches to social research: explanatory, interpretive and critical; and discusses their implications for public administration research. Adams (1994), White and Adams (1994), Farmer (1995); Fox and Miller (1995, 1996) discuss public administration from postmodern perspective. A good source for debate over the nature of research in the discipline of public administration in the 1980s is the *Public Administration Review* articles collection edited by Jay D. White and Guy B. Adams, *Research in Public Administration: Reflections on Theory and Practice* (1994), as well as articles in *Administrative Theory and Praxis*, and sometimes, *Administration and Society*. More and more papers are being delivered at conferences dealing with such ''postmodern'' tools as deconstruction, in public administration context (*Proceedings of the Nineteenth National Conference on Teaching Public Administration* 1996). Increasingly, the argument for new, more inclusive criteria to judge the research in public administration are taking hold in the mainstream public administration. As opposed to radical postmodern conception, criteria derived from positivism are not seen as completely wrong, but rather incomplete. As Jay White (1994; 57) argues:

> The growth of knowledge in public administration can be satisfied by interpretive and critical research as well as explanatory research. . . . reflection on each mode of research is called for to discover what norms, rules, and values pertain to each. The norms and rules will constitute the method of each mode of research, while the values will indicate criteria by which to judge the truth of each type of knowledge. . . . Practical reasoning is fundamentally a matter of interpretation and criticism. It is very much a political endeavor requiring the giving of reasons why one rule should be followed rather than another, or why one criterion should be met rather than another. The growth of knowledge in public administration is based on this type of argumentation.

Qualitative research methods are reclaiming their place in research arsenal in public administration and are now being discussed from public administration (Yeager, 1989), as well as related policy analysis and evaluation (Fischer, 1995), and business management (Gummerson, 1991) perspectives. Still, some authors argue that because in public administration education research methodology courses are taught separate from the main body of study, and are often delayed by the students who take it, students lack ''a critical eye'' when examining basic literature of the field (Bailey, 1994). Although the observation is generally true, there are changes in this tendency since the 1980s, when many studies questioning traditional research methodology or employing different epistemology have been used in public administration classrooms. The connection between organization theory and epistemology have been explored by Thayer (1980) and in a volume edited by Lincoln—*Organization Theory and Inquiry: The Paradigm Revolution* (1985). Bureaucracy has been studied from phenomenological (Hummel 1994a),

critical (Denhardt, 1981), postmodern (Fox and Miller, 1995; Farmer, 1995) perspectives. And finally, there is more practical qualitative methods guide designed for organizational researchers, edited by Cassell and Symon (1994).

IX. CONCLUSION

Qualitative research methods are becoming more and more popular in social science. ''Qualitative research'' is a very general term denoting a host of research strategies and techniques that always should be specified in particular research context. Mostly geared to exploratory, descriptive and interpretive tasks of scientific endeavor, they are invaluable tools of research. Simple absence of numbers does not make one's story qualitative research. Different traditions of qualitative research have established body of procedure and criteria of goodness, and qualitative research designs should conform those requirements for particular tradition. In research process in general, and in quantitative research process in particular, the researcher should be aware of three foci of research: (1) epistemology, (2) research design or strategy; and (3) techniques or tools of data collection and analysis. Though clearly interrelated, these three components of research are not squarely determined by each other—the same epistemology can use different research strategies, and research strategies can use variety of data sources and tools of analysis, which in turn can ascribe to different epistemologies. Very often actual research is a combination of different research methods with variety of sources. This does not mean that the researcher may or should use different methods without discrimination. He or she should be aware of possible implications that employed epistemology, research strategy and data collection and analysis methods will hold for each other, and have a sound rationale for employing a particular design with that particular mix of epistemology, strategy, and methods of data collection and analysis. For example, the use of different methods may be justified for the purposes of triangulation—i.e. trying to explain the studied phenomenon from different perspectives. But this should be done very carefully, without jeopardizing the integrity of each strategy, and clearly integrating them at meta-level. Qualitative research also does not mean absence of criteria for evaluating the research. While criteria may not be accepted across all of the domains of social science, for each research design there are criteria of soundness that have been established through systematic practice in particular subfield. This approach is especially useful in public administration research. Being an interdisciplinary field, public administration draws from multiple sources. When crystallizing research question, the researcher may not only follow the heuristic guidelines like the one suggested above, but as well determine what is the ''sister'' social science that examines similar issues. Is it economics? Is it sociology? Is it political science? What are the traditions in that field? If new research methodology is employed, what new insight will it bring? Will it be accepted by practitioners and scholars? Qualitative research also requires constant reflection. The researcher should strive to distinguish and analyze his or her biases—if not to balance them, at least make them as explicit as possible.

NOTES

1. The interpretivist approach in social science—the desire to understand, rather than explain, has intellectual underpinnings in German tradition of hermeneutics and *Verstehen* tradition in sociology, phenomenology of Alfred Schutz (1967), and critiques of scientism and positivism by ordinary language philosophers. Historically, the interpretivists held the view that mental sciences (*Geisteswissenschaften*) or cultural sci-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ences (*Kulturwissenschaften*) were different in kind than natural sciences (*Naturwissenschaften*), with the goal of the latter being scientific explanation, and the goal of the former being grasping or understanding (*Verstehen*) of the ''meaning'' of social phenomena (Schwandt, 1994: p. 119). The issue was once again forcefully articulated in the US by Clifford Geertz (1973) who called for a new paradigm for social science inquiry, stipulating that it should be ''not an experimental science in search of law but an interpretive one in search of meaning,'' and called for ''thick description'' of social phenomena instead of law-like generalizations of observed relationships between phenomena.

2. For different types of evaluation designs, employing qualitative and quantitative methods, see Patton (1987: Ch. 4).
3. Critical theory is not limited to ''local knowledge'' only. In fact, Jurgen Habermas (e.g., 1971), who is perhaps, the leading authority on critical or ''emancipatory social science,'' has produced one of the most important critiques of modern society. The reference to ''local knowledge'' in the text should be understood within the context of example.
4. This is of course, too broad a generalization, but still, I believe, a valid one. There have been studies of voter psychology, and numerous books and articles have been devoted to consumer preferences and spending behavior, but by and large, the concept of rational man has remained the premise of analysis. In economics especially, with more easily quantifiable phenomena, and well-articulated methodological foundations (e.g., Friedman 1953), positivism still reigns. Of course, there are the well-established schools of institutional (evolutionary) economics (e.g., Samuels, 1995) and newly emerging school of socio-economics (e.g., *The Journal of Socio-Economics*), and critical studies (e.g., McCloskey, 1985) have been widely recognized in the field of economics, but as James March (1992) has aptly put, ''the war is over, and the victors have lost''—the rhetoric of rationality still is predominant. March (1992: p. 264) writes: ''contemporary microeconomics is a rhetoric of rationality surrounding a rich, behavioral interpretation attentive to limited rationality, conflict, ambiguity, history, institutions, and multiple equilibria. It has adopted most of the substance of many of the early critiques of the theory and seems prepared to do the same with many of the later critiques.'' For a very insightful critique of modern economic thought, see also Heilbroner and Miller's *The Crisis of Vision in Modern Economic Thought* (1995). The issue of rationality in public administration is in similar state. For example, Argyris's (1973) famous polemic with Simon in the pages of *Public Administration Review* was recognized as a powerful critique, and the discipline today sees people in organizations as more multi-dimensional, but Argyris's critique only supplemented rather than substituted Simon's ''administrative man.''
5. For more broader perspective on use of metaphor in thought in general, and in science in particular, see Ortony (1993).

REFERENCES

Adams, G.B. (1994). ''Enthralled with Modernity: The Historical Context of Knowledge and Theory Development,'' in *Research in Public Administration: Reflections on Theory and Practice*, J.D. White and G.B. Adams, (eds.), Newbury Park, CA: Sage, pp. 25–41.

Adler, P.A. and P. Adler (1994) ''Observational Techniques.'' in *Handbook of Qualitative Research*, N.K. Denzin and Y. S. Lincoln. (eds.), Thousand Oaks, CA: Sage, pp. 377–392.

Argyris, Chris, R. Putnam, and D.M. Smith (1985). *Action Science*, San Francisco: Jossey-Bass.
Argyris, C. (1973). ''Organization Man: Rational and Self-Actualizing,'' *Public Administration Review, 33*: 253–267.
Argyris, C. (1973). ''Some Limits of Rational Man Organizational Theory,'' *Public Administration Review, 33*: 253–267.
Asbury, J.-E. (1995). ''Overview of Focus Group Research,'' *Qualitative Health Research, 5* (4): 414–420.
Atkinson, P., S. Delamont and M. Hammersley (1988). ''Qualitative Research Traditions: A British Response to Jacob,'' *Review of Educational Research, 58*: 231–250.
Atkinson, P. and M. Hammersly (1994). ''Ethnography and Participant Observation,'' in *Handbook of Qualitative Research*, Denzin, Norman K. and Yvonna S. Lincoln, (eds.), Thousand Oaks, CA: Sage, pp. 248–261.
Atkinson, P. (1992) *Understanding Ethnographic Texts*, Newbury Park, CA: Sage.
Babbie, E. (1990). *Survey Research Methods*, 2nd edition, Belmont, CA: Wadsworth Publishing.
Bailey, M.T. (1994). ''Do Physicists Use Case Studies? Thoughts on Public Administration Research,'' in *Research in Public Administration: Reflections on Theory and Practice*, J.D. White and G.B. Adams (eds.), Newbury Park, CA: Sage, pp. 183–196.
Behn, R.D. ''Management and the Neutrino: The Search for Meaningful Metaphors,'' *Public Administration Review, 52*: pp. 409–419.
Boyle, J. (1994). ''Styles of Ethnography,'' in *Critical Issues in Qualitative Research Methods*, Morse, Janice (ed.), Newbury Park, CA: Sage, pp. 159–185.
Brink, P. and M.J. Wood (eds.) (1989). *Advanced Design in Nursing Research*, Newbury Park, CA: Sage.
Buban, S.L. (1986). ''Studying Social Process: The Chicago and Iowa Schools Revisited,'' in *Studies in Symbolic Interaction: Supplement 2. The Iowa School*, (Part A), C.J., Couch, S. Saxton, and M.A. Katovisch (eds.) Greenwich, CT: JAI Press, pp. 25–38.
Caporoso, J.A. 1995. ''Research Design, Falsification, and the Qualitative-Quantitative Divide,'' *American Political Science Review, 89* (2): 457–460.
Cassell, C. and G. Symon. (eds.) 1994. *Qualitative Methods in Organizational Research: A Practical Guide*, London: Sage.
Chein, I., S. Cook, and J. Harding (1948). ''The Field of Action Research,''*American Psychologist, 3*: 43–50.
Chenail, R. (1993). ''A Case for Clinical Qualitative Research,'' *The Qualitative Report: An Online Journal Dedicated to Qualitative Research and Critical Inquiry*. Vol. 1, No. 4, Summer 1993. WWW document at URL: http://www.nova.edu/ssss/QR.
Cohen, M.Z. and A. Omery. 1994. ''Schools of Phenomenology: Implications for Research,'' in *Critical Issues in Qualitative Research Methods*, J. Morse, (ed.) Newbury Park, CA: Sage, pp. 136–157.
Denhardt, R.B. (1981) . *In the Shadow of Organization*, Lawrence, KS: University of Kansas Press.
Denzin, N.K. (1978). *The Research Act: A Theoretical Introduction to Sociological Methods*, 2nd edition, New York: McGraw Hill.
Denzin, N.K. (1989). *Interpretive Interactionism, Newbury Park, CA: Sage.*
Denzin, N.K. (1994). ''The Art and Politics of Interpretation,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage. pp. 500–515.
Denzin, N.K. and Y.S. Lincoln (1994a). ''Introduction: Entering the Field of Qualitative Research,'' in *Handbook of Qualitative Research*, Denzin N.K. and Y.S. Lincoln, (eds.) Thousand Oaks. CA: Sage, pp. 1–17.
Denzin, N.K. and Y.S. Lincoln (eds.) (1994b). *Handbook of Qualitative Research*, Thousand Oaks, CA: Sage.
Deshler, D. and E. Merrill (1995). *Participatory Action Research: Traditions and Major Assumptions*, Internet WWW page at URL: http://munex.arme.cornell.edu/parnet/tools/tools1.html, May 25, 1995.
Dexter, L.A. (1970). *Elite and Specialized Interviewing*, Northwestern University Press.
Diesing, P. (1991). *How Does Social Science Work? Reflections on Practice*. University of Pittsburgh Press.

Douglas, J.D. and J. Johnson (1977). *Existential Sociology*, Cambridge: Cambridge University Place.
Erickson, F. (1986). ''Qualitative Methods in Research on Teaching,'' in *Handbook of Research on Teaching*, 3rd edition, M.C. Winrock (ed.), New York: MacMillan, pp. 119–161.
Erlandson, D.A., L.E. Harris, B.L. Skipper, and S.D. Allen (1993). *Doing Naturalistic Inquiry: A Guide to Methods*, Newbury Park, CA: Sage.
Farmer. D.J. (1995). The *Language of Public Administration: Bureaucracy, Modernity and Postmodernity, University* of Alabama Press, 1995.
Fischer, E. 1995. *Evaluating Public Policy*, Chicago: Nelson Hall.
Fontana, A. and J.H. Frey (1994) ''Interviewing: The Art of Science,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln, (eds.), Thousand Oaks, CA: Sage, pp. 361–376.
Fowler, F.J., Jr. (1984) *Survey Research Methods*, Newbury Park, CA: Sage Publications.
Fox, C.J., and H.T. Miller (eds.) (1996) ''Symposium: Modern/Postmodern Public Administration: A Discourse About What is Real,'' *Administrative Theory and Praxis, 18* (1): 41–138.
Fox, C.J., and H.T. Miller (1995). *Postmodern Public Administration*, Thousand Oaks, CA: Sage.
French, W.L. and C. Bell (1995). *Organizational Development: Behavioral Science Interventions for Organization Improvement*, 5th edition, Englewood Cliffs, NJ: Prentice Hall.
Friedman, M. (1953). ''The Methodology of Positive Economics,'' in *Essays in Positive Economics*, M. Friedman, Chicago: University of Chicago Press, 1953, pp. 3–43.
Garfinkel, H. (1967). *Studies in Ethnomethodology*, Englewood Cliffs, NJ: Prentice Hall.
Geertz, C. (1973). ''Thick Description: Toward an Interpretive Theory of Culture,'' in *Interpretation of Cultures*, C. Geertz, New York: Basic Books.
Glasser, B. and A. Strauss (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago: Aldine.
Glasser, B. (1978). *Theoretical Sensitivity*, Mill Valley, CA: Sociology Press.
Glasser, B. (1992). *Basics of Grounded Theory Analysis*, Mill Valley, CA: Sociology Press.
Goffman, E. (1971). *Relations in Public*, New York:Basic Books.
Guba, E.S. and Y.S. Lincoln (1989). *Fourth Generation Evaluation*, Newbury Park, CA: Sage.
Guba, E.S. and Y.S. Lincoln (1994). ''Competing Paradigms in Qualitative Research,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 105–117.
Gummeson, E. (1991). *Qualitative Methods in Management Research*, Newbury Park, CA: Sage.
Habermas, J. (1971). *Knowledge and Human Interests*, Boston: Beacon Press.
Hamilton, D. (1994). ''Traditions, Preferences, and Postures in Applied Qualitative Research,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 60–69.
Heidegger, M. (1972) *On Time and Being*, New York: Harper and Row.
Heilbronner, R. and W. Milberg (1995). *The Crisis of Vision in Modern Economic Thought*, Cambridge University Press.
Hodder, I. (1994). ''The Interpretation of Documents and Material Culture,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 393–402.
Holstein, J.A. and J.F. Gubrium, (1994). ''Phenomenology, Ethnomethodology and Interpretive Practice,'' in *Handbook of Qualitative Research*, N.K. Denzin and S. Lincoln (eds), Thousand Oaks, CA: Sage, pp. 262–273.
Huberman, A.M. and M.B. Miles, (1994). ''Data Management and Analysis Methods,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 428–444.
Hummel, R. (1994a). *The Bureaucratic Experience*, 4th. ed., New York: St. Martin's Press.
Hummel, R. (1994b). ''Stories Managers Tell: Why They Are as Valid as Science,'' in *Research in Public Administration: Reflections on Theory and Practice*, J.D. White, and G.B. Adams, (eds.), Newbury Park, CA: Sage, pp. 225–245.
Husserl, E. (1970). *The Crisis of European Sciences and Transcendental Phenomenology*, Evanston. IL: Northwestern University Press.
Jacob, E. (1987). ''Qualitative Research Traditions: A Review,'' *Review of Educational Research, 51*: 1–50.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Jacob, E. (1988). ''Clarifying Qualitative Research: A Focus on Traditions,'' *Educational Researcher*, *17*: 16–24.
Janesick, V. (1994). ''The Dance of Qualitative Research Design: Metaphor, Methodolatry and Meaning.'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln, (eds.), Thousand Oaks, CA: Sage, pp. 209–219.
Johnson, J.B. and R.A. Joslyn, (1991). *Political Science Research Methods*, 2nd edition, Washington, D.C.: Congressional Quarterly Press.
Jorgensen, D.L. (1989). *Participant Observation: A Methodology for Human Studies*, Thousand Oaks, CA: Sage.
Kass, H.D. and B.L. Catron (eds.) (1990). *Images and Identities in Public Administration*, Newbury Park, CA: Sage.
Kelle, U. (ed.) (1995). *Computer-Aided Qualitative Data Analysis: Theory, Methods, and Practice*, Newbury Park, CA: Sage.
King, G., R. Keohane, and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton University Press.
King, G., R. Keohane, and S. Verba (1995). ''The Importance of Research Design in Political Science,'' *American Political Science Review*, *89* (2): 475–481.
Kritzer, H.M. (1996). ''The Data Puzzle: The Nature of Interpretation in Qualitative Research,'' *American Journal of Political Science*, *40* (1): 1–32.
Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*, Chicago University Press.
Lakatos, I. (1970). ''Falsification and the Methodology of Scientific Research Programmes,'' in *Criticism and the Growth of Knowledge*, I. Lakatos, and A. Musgrave (eds.), Cambridge: Cambridge University Press. pp. 91–196.
Levine, H.G. (1985). ''Principles of Data Storage and Retrieval for Use in Qualitative Evaluations,'' *Educational Evaluation and Policy Analysis*, *7* (2): 179–186.
Lijphart, A. (1975). *The Politics of Accomodation: Pluralism and Democracy in the Netherlands*, Berkeley, CA: University of California Press.
Lincoln, Y.S. (ed.) (1985). *Organizational Theory and Inquiry: The Paradigm Revolution*, Newbury Park, CA: Sage.
Lincoln, Y.S. and E.S. Guba (1985). *Naturalistic Inquiry*, Beverly Hills, CA: Sage.
Lincoln, Y. and N.K. Denzin (eds.) (1994). ''The Fifth Moment,'' in *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage, pp. 575–587.
Lofland, L. (1989). ''Social Life in the Public Realm,'' *Journal of Contemporary Ethnography*, *17*: 453–482.
Manning, P.K. and B. Cullum-Swan (1994). ''Narrative, Content, and Semiotic Analysis,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln, (eds.), Thousand Oaks, CA: Sage, pp. 463–478.
March, J.G. (1992). ''The War is Over and the Victors Have Lost,'' *The Journal of Socio-Economics*, *21* (3): 261–267.
Marshall, C. and G.B. Rossman, (1995). *Designing Qualitative Research*, 2nd edition, Thousand Oaks, CA: Sage.
Marton, F. (1994). ''Phenomenography,'' in *International Encyclopedia of Education*, T. Husen and T.N. Postlethwaite, (eds.), London: Pergamon Press, pp. 4424–4429.
McCloskey, D. (1985). *The Rhetorics of Economics*, Madison: The University of Wisconsin Press.
McGrath, J. (1981). ''Dilemmatics: The Study of Research Choices and Dilemmas,'' *American Behavioral Scentist*, *25* (2): 179–210.
Meier, K. and J. Brudney (1993). *Applied Statistics for Public Administration*, 3rd edition, Belmont, CA: Wadsworth Publishing.
Miles, M.B. and A.M. Huberman (1984). *Qualitative Data Analysis: A Sourcebook of New Methods*, Beverly Hills, CA: Sage.
Miles, M.B. and A.M. Huberman (1994). *Qualitative Data Analysis: A New Sourcebook of Methods*, 2nd edition, Beverly Hills, CA: Sage.
Mintzberg, H. (1973). *The Nature of Managerial Work*, New York: Harper and Row.
Mitroff, I.I. and L.R. Pondy (1974). ''On the Organization of Inquiry: A Comparison of Some

Radically Different Approaches to Policy Analysis,'' *Public Administration Review*, *34*: 471–479.
Morse, J.M. (1996). ''Is Qualitative Research Complete?'' *Qualitative Health Research*, *6* (1): 3–5.
Morse, J.M. (1995). ''The Significance of Saturation,'' *Qualitative Health Research*, *5* (2): 147–149.
Morse, J.M. (1994a). ''Designing Funded Qualitative Research,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage. pp. 220–235.
Morse, J.M. (1994b). '''Emerging from the Data': The Cognitive Process of Analysis of Qualitative Inquiry,'' in *Critical Issues in Qualitative Research Methods*, J. Morse (ed.), Newbury Park, CA: Sage, pp. 23–43.
Muecke, M. (1994). ''On the Evaluation of Ethnographies,'' in *Critical Issues in Qualitative Research Methods*, J. Morse (ed.), Newbury Park, CA: Sage, pp. 187–209.
Myers, M.D. (1996). *Qualitative Research in IS*. Internet WWW page, at URL: http://comu2.auckland.ac.nz/~isworld/quality.htm, May 1996.
Nau, D.S. 1995. ''Mixing Methodologies: Can Bimodal Research be a Valuable Post-Positivist Tool?'' *The Qualitative Report: An Online Journal Dedicated to Qualitative Research and Critical Inquiry*, Vol. 2, No. 3, December 1995. WWW document at URL: http://www.nova.edu/ssss/QR.
O'Sullivan, E. and G. Rassel (1994). *Research Methods for Public Administration*, 2nd edition, White Plains, NY: Longman Publishers USA.
Olson, H. (1995). ''Quantitative 'Versus' Qualitative Research: The Wrong Question,'' Internet WWW page at URL: http://www.alberta.ca./dept/slis/cais/olson.htm.
Ortony, A. (ed.) (1993) *Metaphor and Thought*, Cambridge University Press.
Patton, M.Q. (1987). *How to Use Qualitative Methods in Evaluation*, Newbury Park, CA: Sage.
Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods*, Newbury Park, CA: Sage.
Pfiffner, J.M. (1940). *Research Methods in Public Administration*, New York: The Ronald Press Company.
Proceedings of the Nineteenth National Conference on Teaching Public Administration, February 16–17, 1997, Savannah, Georgia.
Punch, M. (1994). ''Politics and Ethics in Qualitative Research,'' in *Handbook of Qualitative Research*, N.K. Denzin and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 83–98.
QUALIDATA (1996). World Wide Web document. Available at URL: http://www.essex.ac.uk/qualidata, September 1996.
Qualitative Inquiry. A Quarterly Journal, Sage Periodicals.
Qualitative Research (1995). World Wide Web document. Available at URL:http://lispstat.alcd.soton.ac.uk/am306/qiualitative.text.
Ray, M.C. (1994). ''The Richness of Phenomenology: Philosophic, Theoretic and Methodologic Concerns,'' in *Critical Issues in Qualitative Research Methods*, J. Morse (ed.), Newbury Park, CA: Sage, pp. 117–133.
Reason, P. (1994). ''Three Approaches to Participative Inquiry,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 324–339.
Richards, T.J. and L. Richards, (1994) ''Using Computers in Qualitative Research,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Yvonna S. Lincoln (eds), Thousand Oaks, CA: Sage, pp. 445–463.
Rogowski, R. (1995). ''The Role of Theory and Anomaly in Social-Science Inference.'' *American Political Science Review*, *89* (2): 467–470.
Rosaldo, R. (1989). *Culture and Truth: The Remaking of Social Analysis*, Boston: Beacon.
Rosenau, P.M. (1992). *Post-Modernism And The Social Sciences: Insights, Inroads And Intrusions*, Princeton University Press.
Samuels, W.J. (1995). ''The Present State of Institutional Economics,'' in *Cambridge Journal of Economics*, Vol 19, pp. 569–590.
Schutz, A. (1967). *The Phenomenology of Social World*. Evanston, IL: Northwestern University Press.
Schwandt, T.A. and E.S. Halpern (1988). *Linking Auditing and Metaevaluation: Enhancing Quality in Applied Research*, Newbury Park, CA: Sage.
Schwandt, T.A. (1994). ''Constructivist, Interpretivist Approaches to Human Inquiry,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln (eds), Thousand Oaks, CA: Sage, pp. 118–137.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Stake, R.E. 1994. ''Case Studies,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln, (eds.), Thousand Oaks, CA: Sage, pp. 236–247.

Stern, P.N. (1994). ''Eroding Grounded Theory,'' in *Critical Issues in Qualitative Research Methods*, J. Morse (ed.), Newbury Park, CA: Sage, pp. 212–223.

Strauss, A. and J. Corbin (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*, Newbury Park, CA: Sage.

Thayer, F. (1980). ''Organization Theory as Epistemology: Transcending Hierarchy and Objectivity,'' in *Organization Theory and Public Administration*, C.J. Bellone (ed.), Boston, MA: Allyn and Bacon.

The Journal of Socio-Economics (1960). Previously, *The Journal of Behavioral Economics*.

The Qualitative Report: An Online Journal Dedicated to Qualitative Research and Critical Inquiry, Available at URL: http://www.nova.edu/ssss/QR. May 1996.

Torbert, W.R. (1991). *The Power of Balance: Transforming Self, Society, and Scientific Inquiry*, Newbury Park, CA: Sage.

Urban Wallace and Associates (1995). ''Should I Use Qualitative or Quantitative Research to Answer My Marketing Question?'' World Wide Web document. Available at URL: http://www.uwa.com/marketing/consultants/research.htm.

van Maanen, M. (1990). *Researching Lived Experience*, Albany, NY: State University of New York Press.

Vidich, A. and S.M. Lyman (1994). ''Qualitative Methods: Their History in Sociology and Anthropology,'' in *Handbook of Qualitative Research*, N.K. Denzin, and Y.S. Lincoln (eds.), Thousand Oaks, CA: Sage, pp. 23–59.

Weinholtz, D., B. Kacer, and T. Rocklin (1995). ''Salvaging Quantitative Research With Qualitative Data,'' *Qualitative Health Research*, *5*(3): 388–397.

Weitzman, E.A. and M.B. Miles (1995). *Computer Programs for Qualitative Data Analysis: A Software Sourcebook*, Newbury Park, CA: Sage.

Werner, O. and G.M. Schopfle (1987). *Systematic Fieldwork*, Vol. 1–2, Newbury Park, CA: Sage.

Whelan, R.K. (1989). ''Data Administration and Research Methods in Public Administration,'' in *Handbook of Public Administration*, J. Rabin, W.B. Hildreth, and G.J. Miller (eds.), New York, NY: Marcel Dekker, pp. 657–682.

White, J.D. (1994). ''On Growth of Knowledge in Public Administration,'' in *Research in Public Administration: Reflections on Theory and Practice*, J.D. White and G.B. Adams, (eds.), Newbury Park, CA: Sage, pp. 42–59.

White, J.D. and G.B. Adams (eds.) (1994). ''Making Sense with Diversity: The Context of Research, Theory and Knowledge Development in Public Administration,'' in *Research in Public Administration: Reflections on Theory and Practice*, Newbury Park, CA: Sage, pp. 1–22.

Whyte, W.F. (ed.) (1991). *Participatory Action Research*, Newbury Park, CA: Sage.

Wolcott, H.F. (1992). ''Posturing in Qualitative Inquiry,'' in *The Handbook of Qualitative Research in Education*, M.D. LeCompte et al. (eds.), New York: Academic Press, pp. 3–52.

Yeager, S.J. (1989). ''Classic Methods in Public Administration Research,'' in *Handbook of Public Administration*, J. Rabin, W.B. Hildreth, and G.J. Miller (eds.), New York, NY: Marcel Dekker. 1989, pp. 683–794.

Yin, R.K. (1984). *Case Study Research: Design and Methods*, Newbury Park, CA: Sage.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

11

Statistics for Nominal and Ordinal Data

Michael Margolis
University of Cincinnati, Cincinnati, Ohio

I. OVERVIEW

Despite the growth of sophisticated techniques for multivariate analysis, contingency tables remain a common means for reporting results in both the popular and professional literature. There is good reason for this: contingency tables give readable overviews of the data, and they also illuminate differences among nominal categories or trends across ordinal categories. Furthermore, through cross-tabulation within categories of control variables, they can be used to characterize multivariate relationships.

This chapter takes the perspective that the principal purpose for using these measures should be to enhance or otherwise clarify information about relationships among variables of interest. Statistical data analysis is not an end in itself. Rather, public administrators and policy analysts use it to help explain political events, public policies, or other political phenomena of concern to themselves, their clientele, elected officials, or the citizenry in general. It is useful for some purposes to test whether or not statistically significant relationships exist among variables. Often, however, the samples which contain the variables fail to satisfy the formal assumptions required for applying the statistical test. Moreover, even when such relationships prove to be statistically significant, they may be substantively insignificant. In many cases, therefore, it is more useful to employ measures of association rather than statistical tests to characterize and compare the nature and strength of bivariate relationships

The discussion covers statistical measures of association commonly used to characterize relationships among nominal and ordinal variables in contingency tables. It also includes Kendall's Tau-A and Spearman's Rho, two rank order statistics frequently used to characterize bivariate relationships among ordinal variables that have few or no ties in rank. The sections that follow present first nominal and then ordinal measures of association. Nominal measures presented include: Percentage difference, Chi-square, Contingency coefficient, Phi, Cramer's V, Lambda, and Goodman-Kruskal Tau. Ordinal measures include Tau-A, Tau-B, Tau-C, Gamma, Somer's D, Wilson's E, and Spearman's Rho.

We will discuss each measure in the context of the types of questions it seems most suitable to answer. In addition, we will discuss its null and perfect conditions and will attempt to give readers both a formal basis and a more intuitive feel for interpreting the range of values each measure can assume. The final sections will introduce multivariate considerations, review the main advantages and disadvantages of the measures discussed, and make some suggestions for their application.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

II. NOMINAL MEASURES OF ASSOCIATION

A. Introduction

Nominal variables are divided into mutually exclusive and exhaustive categories for purposes of measurement. The categories have no natural order. A city, for example, can be divided into north, south, east, and west sides, and every residence in the city can be reliably placed in one and only one of the four categories. But absent some theory or hypothesis that imposes a direction: e.g., the West and South sides have larger proportions of owner-occupied residences than do the North and East sides, the order in which the four categories are presented makes no difference.

It also follows that the magnitude of any desirable nominal measure of association will be unaffected by the way the categories are ordered. Indeed, as we shall demonstrate, when a theory or hypothesis imposes order or direction on nominal variables, it is usually more appropriate to use ordinal rather than nominal measures of association to characterize their association.

The measures discussed below are used to examine bivariate associations. By examining bivariate relationships within categories of control variables, however, they can also be used to characterize multivariate relationships.

B. Percentage Difference

There are many ways to describe data. A good general rule is to start simply. For any variable of interest, answering the question, ''What have we here?'' should normally precede examining the relationship of that variable to any other variables. (See Kaplan, 1964; also see Chapters 3 and 4 of this volume for a discussion of levels of measurement and univariate descriptive measures). After researchers have familiarized themselves with the univariate distributions of nominal or ordinal variables of interest, they normally begin analyzing the data by comparing the proportions (percentages) of observations that fall into various categories.

There are also many ways to display data. While we will focus on bivariate tables here, standard statistical programs, such as SPSS and Excel, feature excellent graphics that can be used to display bivariate relationships among nominal and ordinal variables. (See Norusis, 1995: Chapter 7; Microsoft, 1993–1994: parts 3 and 4). However researchers decide to display their data on the relationships among two or more nominal or ordinal variables, they should first consider the nature of the underlying hypothesized relationship.

For bivariate relationships, a primary consideration is whether one or more variables are hypothesized as independent (possible causes) and one or more as dependent variables (possible consequences) or whether the variables are simply expected to have an association of an (as yet) unspecified causal nature. When the former condition holds, it normally makes the most sense to display and then compare percentages across the categories of the independent variable. When the latter condition holds, analysts may choose to display and compare percentages across the categories of either or both row and column variables or to display and compare categories as proportions of the entire number of cases in the set of data.

Consider, for instance, the west-east-north-south side variable mentioned above. Suppose we took a random sample of 1000 residences stratified to replicate the proportion of residences on each side of town and that we interviewed a knowledgeable adult at each residence. If we asked interviewees whether the dwelling unit was owner-occupied, rented, or an institutional residence, such as a halfway house, senior citizen home, or residential treatment center, a simple cross-tabulation of the responses by side of town might look like the data seen in Table 1.

We hypothesized that the West and South sides would have larger proportions of owner-occupied residences than would the North and East sides. The univariate marginal (''Total'')

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 1 Type of Residence by Side of Town (Factitious Raw Data)

	West	East	North	South	Total
Owner-occupied	142	73	145	150	510
Rental	68	147	120	55	390
Institution	15	30	35	20	100
Total	225	250	300	225	1000

distributions show that ''owner-occupied'' is the modal row category and that the East and North sides have more residences than do the West and South sides. The 73 owner-occupied residences on the East side seem below average, but what about the 145 owner-occupied residences on the North side? The unequal number of cases in the cells and marginal totals hamper direct comparisons. The data in the table may support the hypothesis, but it is difficult to tell just from examining the raw number of cases. If we consider the column variable to be the predictor (independent variable) here, however, then calculating percentages down the columns, the table can be rewritten as Table 2.

By comparing the percentages of owner-occupied residences on the West and South sides against those on the East and North it becomes abundantly clear that the data do indeed indicate that the former have greater proportions of owner-occupied residences than do the latter. Moreover, by taking the percentage differences across rows, we can make statements like ''Approximately two-thirds of South Side residences are owner occupied. This proportion (66.7%) exceeds the proportion on the East Side (29.2%) by nearly 38 percentage points.'' We also observe that not only are rentals the modal residence for East Side (58.8%), in contrast to all other regions of the city, but that this proportion exceeds the city average (39.0%) by nearly 20 percentage points.

Depending upon the principal concerns of our research, we might choose to collapse together some of the rows and columns of the table. For instance, as our original hypothesis suggested that the West and South sides have larger proportions of owner-occupied residences than do the North and East sides we might combine the first and fourth and then the second and third columns of Table 2 to produce a table with two columns and three rows that simply contrasts residential patterns on the West and South sides with those on the North and East.[1] And as the hypothesis distinguished only between owner-occupied and other residences, we might further collapse the table by combining the second and third rows as shown in Table 3.

The data in Table 3 highlight nearly a 25 percent difference between the proportion of

TABLE 2 Type of Residence by Side of Town (Percentages)[a]

	West	East	North	South	Total%
Owner-occupied	63.1%	29.2%	48.3%	66.7%	51.0%
Rental	30.2	58.8	40.0	24.4	39.0
Institution	6.7	12.0	11.7	8.8	10.0
Total	100.0%	100.0%	100.0%	99.9%	100.0%
(N)	(225)	(250)	(300)	(225)	(1000)

[a] Figures are presented to 10ths of a percent for illustrative purposes in the discussion below. Given 95% confidence intervals of as much as ±3% for estimates of proportions based on a simple random sample of 1000, a researcher would ordinarily round back to whole percentages when presenting the data in a report.
Source: Table 12.1

TABLE 3 Type of Residence by Sides of Town (Combined Percentages)

	West and South	North and East	(Total%)
Owner-occupied	64.9%	39.6%	51.0%
Not owner-occupied	35.1	60.4	49.0
Total%	100.0%	100.0%	100.0%
(N)	(450)	(550)	(1000)

Source: Table 2.

owner-occupied dwellings on the West and South sides in comparison to owner-occupied dwellings on the North and East sides. In addition, the West and South sides have 15 percent more owner-occupied dwellings than the average of 51 percent for the city, while the North and East sides have over 11 percent less than the average.

Overall, the data illustrate how the percentage differences appear to support the hypothesis that the West and South sides have proportionately more owner-occupied dwellings than do the North and East sides. The proportions of owner-occupied dwellings seem similar on the West and South sides, and these proportions contrast sharply with those on the North and East sides. The difference between the proportion of owners on the West and South sides (combined) from the East side seems especially stark: over 35 percentage points. Indeed, the proportion of owner-occupied dwellings on the North side is as close or closer to the proportions on the West and South sides than it is to the proportion on the East side.

C. Nominal Measures Based on Chi-Square

Even though analysts can make good use of percentage differences to highlight or otherwise contrast patterns of relationship(s) among two (or more) nominal variables in a contingency table, the descriptions can become prolix. This becomes more apparent as the number of cells in a table or the number of variables under consideration increase. A two by two table has six possible percentage comparisons that can be made among its cells (though some will be redundant) plus up to eight additional comparisons that can be made between cells and their row or column marginal totals. For a three by three table the number of cell comparisons jump to 36 and 18 respectively. Introducing a third (control) variable increases the number of percentage difference comparisons as many fold as the control variable has categories. For example, splitting the data in Table 3 to control for dwelling units with above and below the town's median household income would double the number of possible percentage differences; controlling the same data for number of children in household—none; one or two; more than two—would treble the number of possible percentage difference comparisons. Clearly, meaningful statistics that summarize patterns of data in a table would save analysts and their readers both time and space.

Statistics that summarize patterns among nominal variables in a contingency table conventionally range between 0 and 1. ''Zero'' represents a null relationship (no association) and ''1'' represents a perfect one. The most common null condition for ''no association'' between two variables is statistical independence. Under this condition we can determine the expected values for each cell (i, j):

$$\text{Expected (i, j)} = [\text{total for row (i)} * \text{total for column (j)}]/N \quad (1)$$

where N is the total number of cases in the table.

The familiar Chi-square statistic for testing the null hypothesis of statistical independence is based upon the comparison of the expected number of cases in each cell of an r by k table with the actual number observed. Specifically

$$\text{Chi-square} = \sum [\text{Observed (i, j)} - \text{Expected (i, j)}]^2/\text{Expected (i, j)} \quad \text{for i} = 1 \text{ to r; j} = 1 \text{ to k.} \quad (2)$$

where r is the number of rows and k is the number of columns.

As this statistic's distribution approximates the chi-square distribution when the row and column variables are independent, values for Chi-square that indicate a significant deviation from that distribution allow us to reject the null hypothesis of no association. While a Chi-square of 0 indicates that the data support the null hypothesis of statistical independence, the Chi-square statistic itself can range to large positive numbers. Thus, it fails to satisfy the conventional criterion of having ''1'' stand for a perfect relationship between two variables in a contingency table. Chi-square indicates the presence of a relationship, but not its strength. Indeed, if one doubles the cases in each cell of a table, the Chi-square statistic will double, even though the proportion of cases in each cell, and hence the percentage differences among cells, will not change at all.

Statisticians have developed several measures of association that adjust for these deficiencies but still retain the familiar Chi-square formula. These include:[2]

$$\text{Phi} = \text{Square root of (Chi-square/N)} \quad (3)$$

Pearson's Contingency coefficient,

$$\text{C} = \text{Square root of [Chi-square/(Chi-square + N)]} \quad (4)$$

and

$$\text{Cramer's V} = \text{Square root of [Phi}^2/(\text{min} - 1)] \quad \text{where min} = \text{r or k, whichever is smaller.}^3 \quad (5)$$

Each of these statistics, like Chi-square itself, has a value of zero when the two variables it describes are statistically independent. Using the formulas given above readers can verify that Tables 1 through 3 generate the statistics found in Table 4.

The statistics show several desirable characteristics. Each normally varies between zero and one, and their magnitudes tend to remain within relatively small portions of their possible ranges as the original data is compressed from a larger (four by three) to a smaller (two by two) table. Moreover, they are unaffected by proportionate increases or diminution of the number of cases in the cells of the table. If we double the cases in the cells of the original tables, only the Chi-squares double to 175.0 and 126.4 respectively. The other statistics remain unchanged.

TABLE 4 Comparing Statistical Measures

	Tables 1 and 2	Table 3
Chi-square	87.5*	63.2*
Phi	.296	.251
C	.284	.244
V	.209	.251

* $p < .001$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Finally, as we would expect of nominal level measures, changing the order of the rows or columns does not affect their values.

Unfortunately, these Chi-square based statistics also have some undesirable characteristics. In tables with more than two rows or columns the value of Phi can exceed 1 for strong relationships. Conversely, regardless or the strength of relationship among the two variables, C can never reach a maximum value of 1. Its maximum value changes from .707 for a two by two table to values that approach (but never reach) 1 as the number of rows and columns increase. Only V always remains within the zero to one range and can attain a value of 1 for tables where $r \neq k$. But for all the statistics, including V, no simple interpretations exist for most magnitudes in their ranges.

In short, the most that can be said about the data in Tables 1 through 3 is that owner-occupied dwelling units are not distributed proportionately across town. The West and South sides have disproportionately more units, the North has just under the expected average and the East side has disproportionately fewer units. The probability that a full census of housing would reveal no relationship between side of town and owner-occupation of dwelling units is less than one in a thousand. The strength of association is between the types of dwelling unit occupation and side of town is shows a range of .209 to .251 as measured by Cramer's V. While this is low in the possible range of values $0 \leq V \leq 1$, its strength must be compared relative to the strength of other nominal relationships.[4] Finally, one can compare the expected to the actual values to discern the substantive content of the cells that make the largest contributions to the statistics. For example in Table 5, West, East, and South Owner-occupied, and East and South Rental, show the largest deviations from expected values.

D. Nominal Measures of Proportion Reduction of Error

Because of the difficulty of interpreting the magnitude of Chi-square based measures, some analysts recommend using statistical measures of association that indicate proportion reduction of error (PRE). For strength of association among nominal variables these measures again run between 0 and 1, but in this case, their magnitude indicates an easily interpretable improvement in predicting a case's category on a dependent variable based on new information about the case's category for an independent variable.

Two of the most common PRE measures for nominal variables are Lambda and Goodman-Kruskal Tau. Unlike the Chi-square measures each has a symmetric and asymmetric form, the latter in each case being the more easily interpretable. Each of the asymmetric measures has a guessing rule based on the univariate distribution of the dependent variable. The measure's value represents the improvement or proportion reduction of error in making new guesses when

TABLE 5 Type of Residence by Side of Town (Raw Data and Expected Values)

	West	East	North	South	Total
Owner-occupied	142	73	145	150	510
(expected)	(114.75)	(127.5)	(153)	(114.75)	
Rental	68	147	120	55	390
(expected)	(87.75)	(97.5)	(117)	(87.75)	
Institution	15	30	35	20	100
(expected)	(22.5)	(25)	(30)	(22.5)	
Total	225	250	300	225	1000

information about the case's value on the independent variable is made known. Both Lambda and Tau usually take on different values depending upon which of two variables under consideration is dependent. When no choice can be made regarding the causal order of the variables, the symmetric forms of the measures are used. The symmetric forms can be interpreted as a weighted average of the asymmetric measures.

Lambda's asymmetric forms are conceptually the most simple. They begin by counting up the number of errors we make if we guess the modal category on the dependent variable. We then compare these errors with the number of errors we make using information about the category (column or row) of the independent variable into which each case falls. The statistical value represents the improvement $0 \leq \text{Lambda} \leq 1$ (proportion reduction of error) of the second set of predictions over the first. The formulas are:

Lambda for Columns Independent =

$$1 - [N - \sum \text{Largest cell in column (j)}]/[N - (\#\text{ of cases in largest row})] \quad (6)$$

where r = number of rows, k = number of columns,

N = # of cases in the table, and j = 1, 2 . . . , k

Lambda for Rows Independent =

$$1 - [N - \sum \text{Largest cell in row (i)}]/[N - (\#\text{ of cases in largest column}] \quad (7)$$

where r = number of rows, k = number of columns,

N = # of cases in the table, and i = 1, 2 . . . , r.

Consider again the data in Table 1. If we choose ''Side of Town'' as the independent variable, then ''Owner-occupied'' becomes the modal category on the dependent variable, ''Type of Residence.'' If we guess this category every time for the dependent variable we will make 510 correct predictions and 490 errors. The 490 errors equal [N − 510], that is [N − # of cases in largest row] in Equation 6.

Now suppose someone tells us the side of town of each dwelling unit before we guess. For West, North, and South we will still guess ''Owner-occupied,'' the modal cell category for each of these columns. We will make (142 + 145 + 150) = 437 correct predictions and (68 + 15) + (120 + 35) + (55 + 20) = 313 errors. For the East side, however, we will guess ''Rental.'' This will result in an additional 147 correct predictions and (73 + 30) = 103 errors. The new total correct is (437 + 147) = 584 with (313 + 103) = 416 errors. This is [N − $\sum$ Largest cell in column (j)] in Equation 6.

Plugging these values into the formula we get Lambda for Columns independent = 1 − [(1000 − 584)/(1000 − 510)] = 1 − (416/490) = 1 − .849 = .151, or approximately a 15 percent reduction in errors.[5]

If we chose to look at Type of Residence (rows) as the independent variable, our guessing rule would be the same. We would guess that every residence, regardless of its type, was on the North Side. We would make 700 errors. If we knew the residence type, however, we would now guess that ''Owner-occupied'' was on the South Side, Rental on the East Side and ''Institution'' was on the North. We would now make 668 errors. Plugging these values into the Equation 7 we get Lambda for Rows Independent = 1 − (668/700) = .046, or about a 4.5 percent reduction in errors.

It is conceivable, however, that we might have no basis for arguing that one of the other of our variables was independent. For instance, we might theorize that type of housing is really a function of some third variable, such as income. In this case we would employ Lambda symmetric to characterize the association between the variables. Its formula is:

Lambda symmetric =

$$1 - [2*N - \sum \text{Largest cell in column (j)} - \sum \text{Largest cell in row (i)}]/$$
$$[2*N - (\#\text{ of cases in largest row}) - (\#\text{ of cases in largest column})] \quad (8)$$

where r, k, N, i, and j are defined as in Equations 6 and 7.

Applying this formula to the data in Table 1 results in a Lambda symmetric of .089. This value is essentially a weighted average of the two asymmetric Lambdas. Its magnitude is always less than one of the asymmetric Lambdas and greater than the other (except when all three Lambdas are equal).

In addition to their simplicity of interpretation, the Lambdas are relatively easy to calculate even by hand. Moreover, unlike some of the Chi-square based statistics, none can ever exceed 1. Nor can a ''perfect'' relationship—one that has no errors of prediction—fail to equal 1.

Lambda does have a major drawback, however. It is insensitive to percentage differences in tables where the modal category on the dependent variable is sufficiently large to dominate all other categories even when information about the category on the independent variable is available. For example, consider the race of the 300 informants who gave us information about the types of residence on the North Side displayed in Table 1. Suppose there were 50 blacks and 250 whites who identified themselves politically as shown in Table 6.

Examining the percentage differences suggests that ''Black(s)'' are more likely than ''White(s)'' to declare affiliation with the Democratic party. Indeed, Chi-square = 7.67 ($p < .05$); V = Phi = .160; and C = .158. But because there are so many more whites than blacks and so many more Democrats than others, ''White'' and ''Democrat'' remain the modal categories even when new information about party affiliation or race is given. As a result, Lambda Row Indep = Lambda Col. Indep = Lambda Symmetric = 0. This illustrates that unlike the Chi-square based measures, statistical independence is not the null condition for Lambda. In short, Lambdas' focus on modal categories can cause them to ignore relationships that other more sensitive measures will pick up and even find statistically significant at the commonly used $p \leq .05$ level.[6]

The Goodman-Kruskal Taus are PRE measures that are more sensitive to percentage differences than are the Lambdas. The logic of the Taus is similar to that of the Lambdas, but the Taus are more complex and more tedious to calculate by hand. Nonetheless, Goodman-Kruskal Taus can provide summaries of cross-tabulated data that are more sensitive to some percentage differences than are the Lambdas. And with the advent of computerized routines, the tediousness of the calculations is no longer a serious problem.

The formulas for the asymmetric Taus are:

Goodman-Kruskal Tau for Columns Independent =

[(Expected mistakes based on row totals)

− (Mistakes made knowing cell values by column)]/

(Expected mistakes based on row totals) (9)

TABLE 6 Race by Party Affiliation (Factitious Data: North Side)

	Republican	(N)	Independent	(N)	Democrat	(N)	Total
Black	20%	(10)	10%	(5)	70%	(35)	50
White	40%	(100)	10%	(25)	50%	(125)	250
Total	36.7%	(110)	10%	(30)	53.3%	(160)	300

where expected mistakes =

$\sum$ Rowtotal(i) $*$ [(N $-$ Rowtotal(i))/N], for i = 1 to r;

and mistakes knowing cell values by column

= [$\sum$[Cell value(i, j)] $*$ [Coltotal(j) $-$ Cell value(i, j)]/Coltotal(j)],

for i = 1 to r; j = 1 to k.

Goodman-Kruskal Tau for Rows Independent =

[(Expected mistakes based on column totals)

$-$ (Mistakes made knowing cell values by row)]/

(Expected mistakes based on column totals) (10)

where expected mistakes =

$\sum$ Coltotal(j) $*$ [(N $-$ Coltotal(j))/N], for j = 1 to k;

and mistakes knowing cell values by row

= [$\sum$[Cell value(i, j)] $*$ [Rowtotal(i) $-$ Cell value(i, j)]/Rowtotal(i)],

for i = 1 to r; j = 1 to k

and where r, k, N, i, and j are defined as in Equations 6 and 7.

Essentially, the asymmetric Goodman-Kruskal Taus differ from the asymmetric Lambdas only by the application of their more complex guessing rule. Instead of choosing the modal category of the dependent variable as the first guess, we make our guesses in proportion to the distribution of categories on the dependent variable. This leads to more errors than does the guessing rule for asymmetric Lambdas. Similarly, instead of guessing the modal cell category on the dependent variable when we are told the category on the independent variable, we now guess in proportion to the cases in each cell category on the dependent variable.

To illustrate, consider once again the data in Table 1 with ''side of town'' as the independent variable. We begin by making 510 guesses of ''owner-occupied,'' 390 guesses of ''rental'' and 100 guesses of ''institution.'' While it is possible that all these guesses could be right, the number of erroneous guesses expected by chance is [510 $*$ (490/1000)] + [390 $*$ (610/1000)] + 100 $*$ [900/1000] = 249.9 + 237.9 + 90 = 577.8. We note that this guessing rule generates 87.8 more errors than the guessing rule used for the corresponding asymmetric Lambda.

Once we are told the column from which our case is drawn, we begin to guess in proportion to numbers of cases in each cell of that column. Thus, for West side we guess 142 owner-occupied, 68 rental and 15 institution. We make [142 $*$ (83/225)] + [68 $*$ (157/225] + [15 $*$ (210/225)] = 113.83 errors. Continuing across the columns we make an additional 136.65 + 177.83 + 109.77 = 424.25 errors for a total of 538.08. Substituting in Equation 9, the formula reduces to 1 $-$ (538.08/577.8) = .069. This makes Goodman-Kruskal Tau for columns independent smaller than any of the other nominal measures of association for these data. Similarly, Goodman-Kruskal for Rows independent comes out to be only about .028, again smaller than any of the previous measures.

Because the asymmetric Taus have statistical independence as their null condition, they will even detect an association, albeit a weak one for the data in Table 6. Tau for columns independent equals .026 while Tau for rows independent equals .019.

The symmetric form of Goodman-Kruskal Tau, like the symmetric form of Lambda, can be thought of as a weighted average of the asymmetric measures. Its formula is:

Goodman-Kruskal Tau symmetric =

[(Expected mistakes based on row totals) + (Expected mistakes based on column totals)]
− [(Mistakes made knowing cell values by column)
+ (Mistakes knowing cell values by row)]/
[(Expected mistakes based on row totals)
+ (Expected mistakes based on column totals)] (11)

where expected mistakes are calculated as in Equations 9 and 10.

Inserting data from Tables 1 and 6 into these into this formula yields symmetric Taus of .044 and .021 respectively. As was the case with asymmetric forms, Tau is less than Lambda for Table 1 but greater than Lambda for Table 6.

At this stage it should be apparent that there is no one nominal statistic among those we have reviewed that is superior to others in all desired aspects. The Chi-square based statistics are sensitive to percentage differences, but they are difficult to interpret and their some of their ranges can exceed 1. The PRE measures are easier to interpret but Lambda is insensitive to certain percentage differences. Goodman-Kruskal Tau, while more sensitive to some weak relationships than Lambda, uses a PRE guessing rule that tends to generate measures of association whose magnitudes are smaller than all the others for tables where no category of a row or column variable has so many cases as to dominate the distribution of cases.

The choice of which statistic(s) to use to characterize a relationship will depend upon the questions a policy analyst or researcher has in mind. We shall have more to say about this matter after we have discussed common ordinal statistics.

III. ORDINAL MEASURES OF ASSOCIATION

A. Introduction

Ordinal variables, like nominal variables, are divided into mutually exclusive and exhaustive categories for purposes of measurement. The difference is that the categories have a natural or a theoretical order. Places in the finish of a horse race, for instance, can be thought of as having a natural order. Place of finish, therefore, would be an ordinal variable: it runs from first through last, but it says nothing about the distance between any two places in the order.

To continue with our example of sides of town, we could impose a theoretical order on sides of city. Suppose we developed a scale that averaged the z-scores for annual family income, per capita years of formal education of adult residents, and school taxes paid per capita. We could then order ''Side of Town'' on this variable from highest to lowest scale score and investigate the extent to which this order is associated with the pattern of owner-occupation of residential dwellings.[7]

In contrast to nominal measures of association, the magnitudes of desirable ordinal measures of association should be sensitive to the order of the categories of variables in a cross-tabulation. Once again, we would like to standardize our measures so that ''0'' represents no relationship between the variables and ''1'' represents a perfect relationship. We can add an additional piece of information to characterize ordinal relationships, however. We can use a negative sign to indicate that when a case falls into the higher ordinal categories on one variable, it tends to fall into the lower ordinal categories on the other; and vice-versa: lower categories on the first variable are associated with higher categories on the second. A positive sign would denote that the order of categories would tend to vary in the same direction: higher categories

on one variable associated with higher on a second; and similarly, lower categories on one variable associated with lower values on a second.

B. Kendall's Tau-A, Tau-B and Tau-C

Table 7 contains data on the rank order of the dates of the Republican presidential primaries or caucuses of selected states in 1992 and 1996. Low numbers indicate early primaries, high numbers indicate late ones. In recent years states have changed their primary election (or caucus) dates to enhance their influence on the selection of the presidential nominees. Between 1992 and 1996, for example, California and Delaware moved their primaries to earlier dates. We can use the data in Table 7 to estimate the extent to which the order of primaries and caucuses among the selected states changed due to these sorts of moves.

Kendall's Tau-A is a statistic that can be used to help answer this question. It consists of the ratio of concordant minus discordant paired comparisons between the ranks of all cases on two variables, to the total number of possible paired comparisons. Specifically:

$$\text{Tau-A} = 2 * (C - D)/(N^2 - N) \tag{11}$$

where C and D respectively are the total number of Concordant and Discordant paired comparisons and N is the total number of cases;

$$C = \sum [\text{Celltotal } (i, j) * [\sum \text{Celltotal } (m,n)]]$$

$$\text{for } i = 1 \ldots r - 1;\ j = 2 \ldots k \text{ and } m > i \text{ and } n > j;$$

$$D = \sum [\text{Celltotal } (i, j) * [\sum \text{Celltotal } (m, n)]]$$

$$\text{for } i = 1 \ldots r - 1;\ j = 2 \ldots k \text{ and } m < i \text{ and } n > j.$$

Essentially, Tau-A looks at the order of rankings of two cases or observations on the first variable and compares it to the order of rankings on the same two cases or observations on the second variable. For clarity and simplicity of calculation the first variable is ranked in as perfect order as possible. If the rankings run in the same direction on both variables, the pairs of observations are considered concordant. If they run in opposite directions, they are considered discor-

TABLE 7 Republican Delegate Selection Schedules for Select States, 1992–1996

State	1992 rank (date)	1996 rank (date)
Iowa	1 (2/10)	1 (2/12)
New Hampshire	2 (2/18)	2 (2/20)
South Dakota	4 (2/25)	5 (2/27)
Georgia	5.5* (3/3)	10.5* (3/5)
Maryland	5.5* (3/3)	10.5* (3/5)
Wyoming	8 (3/7)	33 (5/4)
Delaware	15.5* (3/10)	3 (2/24)
Texas	15.5* (3/10)	17 (3/12)
Illinois	21 (3/17)	23.5* (3/19)
California	43 (6/2)	26 (3/26)

* If more than one state has the same date (has a tied ranking) each is given the midpoint of the ranks (e.g., Georgia and Maryland are tied for ranks 5 and 6 in 1992 and 8 through 13 in 1996).
Source: Based on Polsby and Wildavsky 1996: pp. 131–132.

dant. If the rankings are tied on either variable, no decision as to concordant or discordant pairs can be made.[8] The paired comparisons for New Hampshire and South Dakota (2–4 for 1992 and 2–5 for 1996) are concordant whereas the paired comparisons for Wyoming and California (8–43 for 1992 and 33–26 for 1996) are discordant. Delaware and Texas are tied at rank 15.5 for 1992, so no paired comparison can be made against their rankings for 1996.

There are 36 concordant pairs and only seven pairs that are discordant. $C - D = 37 - 7 = 29$. Dividing by 45 (all possible paired comparisons $= (N)(N - 1)/2 = (10 * 9)/2 = 45$), we get $29/45 = .64$ as Tau-A for Equation 11. For the set of states chosen, this indicates a fairly strong consistency in the order of delegate selection dates between 1992 and 1996, changes made by California, Delaware, and Wyoming notwithstanding.

Tau-A has a straightforward interpretation and a range of $-1 \leq \text{Tau-A} \leq 1$. It is most useful for describing relationships among two ordinal variables with many ranks and relatively few ties. Its major drawback comes when there are a substantial number of ties in the order of cases on one or both of the variables being compared. By maintaining all possible comparisons in the denominator, Tau-A cannot reach 1 (or -1) if there are any ties. Indeed, because tied comparisons are excluded from both C and D, the numerator is diminished relative to the denominator, and Tau-A effectively counts ties as weakening the relationship.

As most cross-tabulations have many cases in each cell, it follows that there are many ties on one or both of the variables. This makes Tau-A impractical as a measure of association for variables in most cross-tabulations.

Kendall's Tau-B and Tau-C are more forgiving regarding ties. Both maintain C–D in their numerators. The former diminishes the denominator to adjust for ties. The latter is an estimator of Tau $-$ B that adjusts for unequal numbers of rows and columns (Liebetrau, 1983: pp. 72–74). The formulas are:

$$\text{Tau-B} = (C - D)/\text{Square root of Denomsq}; \tag{12}$$

where $\text{Denomsq} = (1/2 * (N^2 - N) - \text{TieRow}) * (1/2 * (N^2 - N) - \text{TieCol})$;

$\text{TieRow} = \sum ((1/2) * [\text{RowtotalN}(i)^2] - \text{RowtotalN}(i))$ for $i = 1 \ldots r$;

$\text{TieCol} = \sum ((1/2) * [\text{ColtotalN}(j)^2] - \text{ColtotalN}(j))$ for $j = 1 \ldots k$.

RowtotalN(i) and ColtotalN(j) = the number of cases in the ith row and jth column respectively; and N = total number of cases.

$$\text{Tau-C} = 2 * \min * (C - D)/(N^2 * (\min - 1)) \tag{13}$$

where min = r or k, whichever is smaller, as in Equation 5.

Tau-B and Tau-C also range from -1 to $+1$ with 0 as a null relationship. Tau-B can reach its maximum and minimum values, however, only when r = k. Tau-C, analogous to Cramer's V, can reach 1 or -1 for a non-square table. As it is an estimator of Tau-B, however, Tau-C generally will be close to (and often less than) Tau-B; and for all cross-tabulations, Tau-A $\leq$ Tau-C, and Tau-A $\leq$ Tau-B.

Returning to the data in Tables 1 and 2, we would like to use the Kendall's Tau statistics to measure the strength of any underlying ordinal relationship between the variables. We can calculate the Tau values for the data as presented, but it would make little sense to do so, for there is no natural order to the sides of the town. Consider once again, however, the z-score scale discussed in the opening paragraph of this section. If we hypothesized that higher scores

TABLE 8 Type of Residence by Side of Town Ordered by Z-Score Scale

	Higher ⟵ Scale Scores ⟶ Lower				
	South	West	North	East	Total
Owner-occupied	66.7%	63.1%	48.3%	29.2%	51.0%
Rental	24.4	30.2	40.0	58.8	39.0
Institution	8.8	6.7	11.7	12.0	10.0
Total%	100.0%	100.0%	100.0%	99.9%	100.0%
(N)	(225)	(225)	(300)	(250)	(1000)

on the z-score scale were associated with greater likelihood of owner-occupied housing and lesser likelihood of institutional housing, with rentals in-between, then we could reorder the columns from highest to lowest on average scale scores as shown in Table 8.

For these data, the Tau-C = .226, Tau-B = .229, and Tau-A = .151. The nominal statistics of course are unaffected by changing the order of the columns. The Chi-square based statistics remain the same as listed in Table 4, and the Lambdas and the Goodman-Kruskal Taus remain steady also. The ordinal measures indicate a positive association of modest size, comparable in magnitude to the nominal association, but now giving information about the direction of association.[9]

C. Gamma and Wilson's E

Tau-B and Tau-C are bracketed by two additional ordinal measures of association which maintain Concordant minus Discordant pairs in their numerator: Gamma and Wilson's E. Gamma is the most forgiving regarding ties: it ignores rather than counts them in its denominator. Wilson's E, a less frequently used statistic, forgives ties only when a case is tied on both the x and y variable. The formulas for these statistics are:

$$\text{Gamma} = (C - D)/(C + D) \tag{14}$$

$$\text{Wilson's E} = 2 * (C - D)/(N^2 - N - \text{Tieboth}) \tag{15}$$

where $\text{Tieboth} = \sum (1/2) * ([\text{Celltotal}(i, j),]^2 - \text{Celltotal}(i, j))$,

$i = 1$ to r and $j = 1$ to k.

Returning to Table 8, Gamma = .342 and E = .170. Reviewing the magnitudes of the ordinal measures discussed so far: Tau-A ≤ Wilson's E ≤ Tau-C and Tau-B ≤ Gamma. Each of these measures is symmetrical. None presumes that the x or y variable is independent. The choice of which to use depends on the questions researchers ask. We shall have more to say about this later.

D. Somer's Ds

Somer's D_{yx} and D_{xy}, in contrast to ordinal statistics presented above, presume that either the x or y variable respectively has been hypothesized as independent. The statistics once again have C − D in their numerator, but they are asymmetric. They forgive ties on the independent variable, leaving in the denominator only those cases that can't be distinguished on the dependent

TABLE 9 Magnitudes of Ordinal Measures of Association

	Table 1 (and 2)	Table 8	Table 3
Tau-A	−.033	.151	.125
Tau-B	−.050	.229	.251
Tau-C	−.049	.226	.250
Gamma	−.074	.342	.476
Wilson's E	−.037	.170	.171
Somer's D_{yx}	−.056	.260	.250
Somer's D_{xy}	−.044	.202	.253

variable. Somer's Ds produce values that bracket Tau-B. One value is greater than (or equal to) Tau-B; the other is less than (or equal to) Tau-B. The formulas are:

$$\text{Somer's } D_{yx} \text{ (for Rows Independent)} = (C - D)/\text{RowDenom}; \quad (16)$$

where $\text{RowDenom} = (1/2 * (N^2 - N) - \text{TieRow})$

and x is the row variable and y is the column variable

$$\text{Somer's } D_{xy} \text{ for Cols. Independent} = (C - D)/\text{ColDenom}; \quad (17)$$

where $\text{ColDenom} = (1/2 * (N^2 - N) - \text{TieCol})$

and x is the row variable and y is the column variable.

The Somer's D_{yx} and D_{xy} for Table 8 are .260 and .202 respectively.[10] Comparing the formulas, readers may also verify that $D_{yx} * D_{xy} = (\text{Tau-B})^2$

We have seen that changing the order of the categories affects the magnitudes of ordinal measures of association. Collapsing the categories also affects these magnitudes. Table 9 summarizes the values of the measures of association for the data in Tables 1 (and 2), 8 and the collapsed categories presented in Table 3.

It should be clear that because the columns of Tables 1 and 2 have no natural or theoretical order, the negative values in the first column represent inappropriate applications of the ordinal measures. They are essentially meaningless; only nominal measures like those in column 1 of Table 4 should be applied.

Once we set the columns in an appropriate theoretical order, however, as done in Table 8, the measures do yield some useful information. If we hypothesize that higher z-score ratings are associated with greater proportions of owner occupied dwelling units and lesser proportions of rental and institutional units, the measures indicate that the relationship is positive though not particularly strong. Collapsing the rows and columns as we did in Table 3 increases the numbers of observations tied within the categories of each variable. Tau-A, which is unforgiving of ties decreases in magnitude. All the other symmetric measures forgive ties to a greater or lesser degree. Gamma, which essentially ignores ties, shows the greatest increase; Wilson's E shows the least. Tau-B and Tau-C show modest increases. The product of the asymmetric Somer's Ds increases though D_{yx} decreases slightly due to the increased ties on the y variable. Somer's D is forgiving of ties on the independent variable, however, and this same increase in tied observations on y, therefore, increases rather than decreases the magnitude of D_{xy}.

E. Spearman's Rho

Spearman's Rho is a popular symmetric rank order statistic, most commonly applied when variables being compared have few or no tied ranks. The logic of its derivation is such that

Rho would equal Pearson's product-moment correlation (r), were the rankings actually interval level measures. Thus, Rho can be defined as (Liebetrau, 1983: pp. 48, 56–58):

$$\text{Rho} = (\sum[R_i - \bar{R}] * [S_i - \bar{S}])/\text{square root of } (\sum[R_i - \bar{R}]^2 * \sum [S_i - \bar{S}]^2) \quad (18)$$

where R_i is the rank of the X variable, S_i is the rank on the Y variable,

$i = 1$ to N for a sample of N paired observations (X_i, Y_i) on each variable,

and $\bar{R}$ and $\bar{S}$ are the mean ranks on X and Y.

It can be shown that this reduces to the calculation formula (Gibbons, 1993: pp. 3–5):

$$\text{Rho} = 1 - \{(6\sum d_i^2)/(N^3 - N)\} \quad (19)$$

where d_i is the difference in ranks between the paired observations, X_i and Y_i.

Rho ranges between -1 (perfect negative relationship) and 1 (perfect positive relationship) with 0 representing no relationship. As in Table 7, when ties occur, the tied observations are assigned the mean of the set of ranks that they would otherwise have occupied. The calculation formula yields perfect relationships only when there are no ties. To account for ties, the calculation formula can be modified to:

$$\text{Rho} = \{(N^3 - N - 6\sum d_i^2 - 6(t' + u')\}/ \{\text{square root of } (N^3 - N - 12t') * \text{square root of } (N^3 - N - 12u')\} \quad (20)$$

where $t' = (\sum t_i^3 - \sum t_i)/12$;

$u' = (\sum u_i^3 - \sum u_i)/12$; and t_i and u_i are the number of ties at any given rank i.

Spearman's Rho is most commonly used when there are few if any ties relative to the number of ranks observed for each of the variables being compared. When there are many ties relative to the ranks, we ordinarily produce a cross tabulation and use the ordinal measures of associations introduced in the previous section to indicate the strength of the relationship.

To calculate d and apply Spearman's Rho to the data in Table 7, we must first rank the data from 1 to 10 as shown in Table 10. Applying Equation 20, we find two sets of ties in 1992

TABLE 10 Republican Delegate Selection Schedules for Select States, 1992–1996**

State	1992 rank (date)	1996 rank (date)	|d|
Iowa	1 (2/10)	1 (2/12)	0
New Hampshire	2 (2/18)	2 (2/20)	0
South Dakota	3 (2/25)	4 (2/27)	1
Georgia	4.5* (3/3)	5.5* (3/5)	1
Maryland	4.5* (3/3)	5.5* (3/5)	1
Wyoming	6 (3/7)	10 (5/4)	4
Delaware	7.5* (3/10)	3 (2/24)	4.5
Texas	7.5* (3/10)	7 (3/12)	0.5
Illinois	9 (3/17)	8 (3/19)	1
California	10 (6/2)	9 (3/26)	1

* To calculate d for the ten states under consideration, the order of the data is ranked from 1 to 10. If more than one state has the same date (has a tied ranking) each is given the midpoint of the ranks. E.g., Georgia and Maryland are tied for ranks 4 and 5 in 1992 and for ranks 5 and 6 in 1996. Delaware and Texas are tied for ranks 7 and 8 in 1992.

** Ranked for Spearman's Rho.

Source: Table 7.

and one set in 1996. Each set has two observations tied with the same rank. Therefore, $t' = 2(\sum 2_i^3 - \sum 2_1)/12 = 2\ (8 - 2)/12 = 1$; and $u' = (\sum 2_i^3 - \sum 2_1)/12 = .5$. Plugging these values into Equation 20, we get Rho = .746. This compares to a Rho of .774 if we use Equation 19, and the Tau-A of .64 that we calculated earlier.[11] Generally, when there are few ties relative to the number of ranks, the unadjusted Rho calculated from 19 will not differ significantly from the adjusted Rho. Modern computer packages, such as SAS and SPSS, will calculate Rho for any cross-tabulation, even when there are many ties relative to ranks (Gibbons 1993: pp. 62–63). Before the advent of computerized routines removed the tediousness of applying Equation 20, ties were often ignored when calculating Rho and the statistic was not commonly used for cross tabulations (Kerlinger 1964: pp. 260–61).

IV. MULTIVARIATE ANALYSIS

A. Control Variables

The measures of association we have discussed throughout this chapter characterize the strength of relationships between two variables. While these bivariate relationships can be interpreted and their strengths can be compared with one another, researchers and practitioners often are interested in theories or problems that require consideration of more than two variables.

The most straightforward method of carrying out multivariate analysis involving nominal and ordinal variables is to introduce ''control'' variables. Essentially, for each category (or relevant categories) of the control variables, we examine the bivariate relationships that had been originally measured in order to determine the extent to which these relationships remain unchanged.

We know, for example, that a greater proportion of women than men voted to re-elect President Clinton in November 1996. We might suspect, however, that this ''gender gap'' could be attributed to differences in men's and women's opinions toward the degree of involvement the federal government should have in resolving social problems, such as access to good day care for children of working mothers, provision of health insurance or medical services for children, or improvement of local public schools. We could test this suspicion by comparing the proportions of men and women who voted for Dole or Clinton within separate categories of a survey variable that measured the extent to which respondents favored such involvement by the federal government. If the control variable had three substantive categories regarding such involvement: (1) more; (2) same as now; and (3) less; we would generate three tables, one for each category. Our suspicion would be affirmed if, for each table generated, the strength of association between sex and presidential vote dropped to nearly zero. This would happen if it turned out that the percentage differences in presidential voting between men and women in each category were small, but that women tended to be clustered in categories (1) and (2) while the men were clustered in categories (2) and (3).

An advantage of this method of control is that the resultant measures of association provide estimates of the strength of the bivariate relationship under three separate circumstances. It is conceivable that the same measures could be different from one another under these separate circumstance. For instance, we might discover that differences between the sexes disappeared within category (1) but that women voted disproportionately for Clinton within categories (2) and (3). This would suggest that men and women differed in their choices of presidential candidates even when they agreed that the role the federal government should play in resolving social problems should either remain the same or shrink. But the gender gap disappeared among those who favored a greater role for the federal government.

We could further extend the multivariate analysis by separating the cases in the tables by race. Such a separation would allow us to check the circumstances (if any) under which a gender gap existed among blacks, who otherwise gave overwhelming support to Clinton over Dole. We would now obtain six separate measures of the association between sex and presidential choice: two comparisons of the sexes—(1) blacks and (2) whites—within each of the three categories regarding role of the federal government.

The disadvantages to this method of analysis stem from two factors. First, as control categories multiply, the presentation and descriptions of the tables and their respective measures of association can become complicated and prolix. This can be thought of as analogous to the problem of always using percentage differences to describe the relationships among variables. Second, as the numbers of tables increase, the cases upon which they are based decrease, and the estimates of strength of association thereby become less reliable. Blacks comprised approximately 10 percent of the voters in 1996 presidential election. As we compare presidential voting choices of men and women, controlling for race and opinion on the federal government's role in the example above, we would expect to find only about 75 black men and 75 black women in a sample of 1500 who actually voted. When we sort these men and women into categories according to their opinions on the degree of involvement in social problems they prefer for the federal government, the numbers of black men and women included in the relevant tables could fall to fewer than 30 cases each. This sparsity of cases would be hardly conducive for making reliable estimates of measures of association, but researchers could view the results as exploratory or preliminary, despite their unreliability. In order to achieve sufficient cases to make reliable estimates, new data then could be collected or perhaps found in other independent surveys.

B. Partial Correlation

An approach that attempts to overcome the above described disadvantages for ordinal variables employs ordinal measures analogous to those used in calculating partial correlations for interval level data (Gibbons, 1993: Chapter 5; Garson, 1976: pp. 361–63). While this method avoids the problems of small numbers of cases and can facilitate shorter, less complicated explanations, it still lacks the richness of using partial correlations in conjunction with ordinary least squares (OLS) regression models. The single partial correlation coefficient allows the researcher to comment on the strength of the bivariate relationship under the designated controls without worrying about the paucity of cases in the cells of particular combinations of variables. There is no related regression equation, however, that can be used to provide an estimate of the impact that a unit change in the independent variable has on the dependent variable.

Additional problems arise. When more than one control variable is introduced, the sampling distributions of the partial correlation coefficients are generally unknown (Gibbons, 1993: p. 50). And when only one control variable is used, examining the bivariate relationship in separate categories often yields a richer analysis, for the table associated with each category of the control variable provides separate measures of association that are unique to that category. Finally, when only one control variable with a limited number of categories is used, the problem of small numbers of cases is unlikely to appear.

An argument can be made, therefore, that if we really want to employ partial correlations to examine the relationships among ordinal variables, we might do better to presume that a known distribution underlies the observations and that the variables themselves can be treated as interval rather than ordinal (Weisberg, Krosnick, and Bowen, 1996: pp. 182–183, 313–315).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

IV. SUMMARY AND CONCLUSIONS

This chapter has reviewed a number of common nominal and ordinal measures of association that public administrators and policy analysts may find useful for describing relationships among two or more variables of interest. The nominal measures covered included Percentage difference, Chi-square, Contingency coefficient, Phi, Cramer's V, Lambda, and Goodman-Kruskal Tau. Ordinal measures included Tau-A, Tau-B, Tau-C, Gamma, Somer's D, Wilson's E, and Spearman's Rho. Besides presenting the formulas for these measures, the chapter has discussed their relation to one another and has given some examples of the research or policy problems to which they can be applied.

Although these measures have known distributions (Gibbons, 1993: Appendix A), the discussion has focused mostly on their substantive rather than statistical significance. Even though computerized data analysis programs normally produce the statistical significance of these measures, their magnitudes—or substantive significance, if you will—for a table or graph are often of more interest to a policy analyst or administrator than are their levels of statistical significance. If the magnitude of an association is sufficiently large and the question of concern is sufficiently important, then, regardless of the level of statistical significance, a good argument can be made for collecting new data or for seeking new evidence from data collected independently by others. Weak relationships, characterized by low magnitudes of association, however, may achieve statistical significance, even when substantively they are of dubious importance. The gender gap in presidential voting, which varied in magnitude between four and seven percent in the four preceding elections before it jumped to 11 percent in 1996, can serve as an example. While the gap had been of statistical significance since the 1980 election, not until 1996 did the majority of men and women who voted Democratic or Republican differ in their presidential choice (Connelly, 1996).[12]

The discussion has suggested imposing a theoretical order that permits movement from nominal to ordinal (and possibly from ordinal to interval) levels of measurement is often a reasonable research strategy. Ordinal and interval measures of association generally allow for richer, more meaningful interpretations of relationships among variables than do nominal measures. If a set of data has a sufficient number of cases, however, successive examination of the separate measures of bivariate association within the categories of the control variable(s) can yield insights that the single value produced by an ordinal (or interval) partial correlation or each single partial regression coefficient of an OLS multiple regression equation may not reveal (Norusis, 1995: p. 472).

In the end, there is no single measure of association that can be applied uniformly to characterize the relationship among two or more nominal or ordinal variables. The choice depends upon the problems or questions the policy analyst or administrator has in mind. Which variables are independent and which are dependent? Or does the theory or hypothesis under investigation provide no definitive guidance as to the possible causal relationships among the variables? To what extent are bivariate relationships uncovered expected to hold across various categories of control variables? Is the null condition independence, or does the theory require a stronger condition before the relationship under investigation assumes substantive significance?

The discussion has attempted to illustrate how the measures relate to one another and how their magnitudes are affected by marginal distributions of the data and by the presence of tied rankings of cases within the categories of ordinal variables. It is hoped that this discussion will provide the basis for making an informed and defensible choice of measures of association suitable to the particular problems or questions of concern.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

NOTES

1. Given the 19% difference between the proportion of owner-occupied dwellings on the North and East sides, a researcher might choose to combine only the West and South columns.
2. Tshchuprow's T = square root of (phi^2/square root of [(r − 1) ∗ (c − 1)]), found in some textbooks, is equal to Cramer's V when r = k (Blalock, 1972; Liebetrau, 1983). Cramer's V can equal 1 when r ≠ k, however, and T cannot. Finally, V, Phi, and C are found in statistical routines such as SPSS crosstabs, but T is not.
3. Note that V = Phi for a two by two table.
4. V directly adjusts for the number of rows and columns in a table, but Phi and C do not. It is advisable, therefore, to use Phi or C to compare tables (1 to N) that have the same number of rows and columns, i.e., where $r_1 = r_2 \ldots = r_N$ and $k_1 = k_2 = \ldots = k_N$.
5. Routines are available in standard statistical packages, such as SPSS, to calculate the various forms of Lambda and Goodman-Kruskal Tau automatically.
6. See Weisberg, 1974 and Bruner, 1976 for more elaborate discussions of the sensitivities of measures in detecting relationships and the effects of the proportions of cases in marginal categories on the magnitudes of the relationships detected.
7. We are ignoring here the ''distance'' measured by the z-scores.
8. There are N(N − 1)/2 possible comparisons that can be made taking combinations of N observations two at a time. Using simple algebra to expand this expression leads to (N^2 − N)/2, which is divided into (C − D) in Equation 11.
9. See Gibbons, 1993 for tests of significance for Kendall's Taus.
10. SPSS also produces a Somer's D symmetric, which I have never found particularly useful. It is essentially a weighted average of the asymmetric D's, whereas Tau-B is the geometric mean of the two (Garson, 1976: p. 295).
11. Note that the new rankings will not affect the value of Tau-A as we have not changed their order relative to one another.
12. Reagan and Bush were the popular choice of both men and women in the presidential elections of the 1980s, as was Clinton in 1992. The Republicans, so to speak, simply were even more popular among men than women, and the opposite was true for Clinton in 1992.

REFERENCES

Blalock, H.M., Jr. (1972). *Social Statistics*, 2nd ed., New York: McGraw-Hill.

Bruner, J. (1976). ''What's the Question to That Answer? Measures and Marginals in Crosstabulation,'' *American Journal of Political Science*, XX: 781–804.

Connelly, M. (1996). ''Portrait of the Electorate,'' *New York Times*, (National Edition), November 10, p. 16.

Garson, G.D. (1976). *Political Science Methods*, Boston: Holbrook Press.

Gibbons, J.D. (1993). *Nonparametric Measures of Association*, Newbury Park, CA: Sage Publications (Quantitative Applications in the Social Sciences, V. 91).

Kaplan, A. (1964). *The Conduct of Inquiry: Methodology for Behavioral Science*, San Francisco: Chandler Publishing Company.

Kerlinger, F. (1964). *Foundations of Behavioral Research: Educational and Psychological Inquiry*, New York: Holt, Rinehart and Winston, Inc.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Liebetrau, A.M. (1983). *Measures of Association*, Beverly Hills, CA: Sage Publications (Quantitative Applications in the Social Sciences, V. 32).
Microsoft Corporation. (1993–1994). *User's Guide: Microsoft Excel*, (Version 5), Redmond, WA: Microsoft Corporation.
Norusis, M.J. (1995). *SPSS 6.1: Guide to Data Analysis*, Englewood Cliffs, NJ: Prentice-Hall..
Polsby, N.W. and A. Wildavsky, (1996). *Presidential Elections: Strategies and Structures of American Politics*, 9th ed., Chatham, NJ: Chatham House.
Upton, G.J.G. (1978). *The Analysis of Cross-Tabulated Data*, New York: John Wiley & Sons.
Weisberg, H. (1974). ''Models of Statistical Relationship,'' *American Political Science Review*, LXVIII: (December) 1638–1655.
Weisberg, H.F., J.A. Krosnick, and B.D. Bowen, (1996). *An Introduction to Survey Research, Polling and Data Analysis*, 3rd ed., Thousand Oaks, CA: Sage Publications.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

12
Analysis of Variance

Carmen Cirincione
University of Connecticut, Storrs, Connecticut

I. INTRODUCTION

A. What is ANOVA?

ANalysis Of VAriance (*ANOVA*) is a set of statistical methods used to assess the mean differences across two or more groups. As others (Iverson and Norpoth, 1987) have said, a more appropriate name might be ''analysis of means,'' but the name refers to the fact that *ANOVA* evaluates mean differences across groups by partitioning sources of variance in the dependent variable. In its simplest form, the variance is isolated into two distinct sources, that due to group membership and that due to chance; sometimes the latter is called error, residual, or within-group variance. For example, a researcher may be interested in the importance of monetary rewards for mid-level managers serving in three different sectors: public, private, and a hybrid (Wittmer, 1991). *ANOVA* could be used to determine whether the differences in attitudes among managers across sectors is simply due to chance.

B. Applications

Public administration and policy scholars have applied *ANOVA* methods to issues of public management, public finance, and public policy. In the realm of public management, Brown and Harris (1993) investigated the influence of workforce diversity in the U.S. Forest Service. They examined attitude differences based on gender while controlling for age, years of experience, education, and professional identification. Edwards, Nalbandian, and Wedel (1981) examined the espoused values of students and alumni from four graduate programs at the University of Kansas: public administration, business administration, law, and social welfare. The purpose of the study was to assess differences in attitude based on program affiliation. Emmert and Crow (1988) attempted to identify characteristics that would distinguish four types of organizations: public-governmental, private-industrial, cooperative, and mixed. Herman and Heimovics (1990) compared the leadership skills of chief executive officers of nonprofit organizations who had been prejudged as effective with those of chief executives of nonprofit organizations who had been prejudged to be less effective. Newell and Ammons (1987) surveyed 527 city managers, mayors, mayoral assistants, and assistant city managers, and found that respondents in each position differed with regard to the perceived emphasis on the management, policy, and political roles played by people in their positions.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Applications of *ANOVA* also can be found in public budgeting research. For example, Frank (1990) and Gianakis and Frank (1993) used *ANOVA* to assess the accuracy of various revenue forecasting methods. Klammer and Reed (1990) conducted an experiment to determine the effects of different formats of cash flow reports on decision making. They found that bank analysts were much more consistent when a direct method was used than when an indirect method was employed.

Similarly, *ANOVA* is used in public policy research in a number of substantive areas. In studies of the criminal justice system, Samuel Nunn (1994) assessed the effects of installing mobile digital terminals in police cars on vehicle theft recovery in Texas. Wells, Layne, and Allen (1991) investigated whether learning styles differed for supervisory, middle, upper middle, upper, and executive managers in the Georgia Department of Corrections. In mental health research, Warner and colleagues (1983) compared the efficacy of three client follow-up methods, and found that face-to-face interviews were more effective than either telephone or mail interviews.

ANOVA is one of the many statistical methods used by scholars in testing their theories with regard to public administration. The goals of this chapter are to introduce the reader to the fundamental principles underlying *ANOVA*, to illustrate the computational steps required to conduct an *ANOVA*, and to illustrate the links between *ANOVA* and other commonly used methods, such as *t*-tests of mean differences and multiple regression analysis.

II. APPROACHES TO ANOVA

There are many methods of employing *ANOVA*, but the fixed effects completely randomized design is most familiar to researchers. This design involves one dependent variable, *Y*, and one or more independent variables, also called factors. If the analysis involves only one independent variable, *X*, it is typically called a oneway *ANOVA*; if it involves multiple independent variables, it is known as a factorial design analysis of variance. The independent variables identify discrete groups of presumably homogenous subjects. They may be qualitative (e.g., participate in a job training program or not) or quantitative (e.g., amount of time in job training program: one month, two months, or three months). Qualitative independent variables measure differences in type or kind; quantitative independent variables measure variations in amount or degree. When analyzing qualitative variables, a researcher wishes to uncover differences in the dependent variable associated with the various groups or kinds. The researcher attempts to determine the nature of the relationship between the dependent and quantitative independent variables. For example, is this relationship best characterized by a linear, a quadratic, or some other, higher-order polynomial function? Qualitative factors are used much more frequently than quantitative factors in *ANOVA* and therefore are the focus of this chapter.

The choice of the groups, also called levels or treatment conditions, determines whether a fixed or a random factors model is chosen. Some authors refer to the former as a Model I and the latter as a Model II (Hays, 1994). In a fixed factor model, the levels of the independent variable represent the comparisons of interest. In a random effects model, the levels have been sampled from a large pool of potential levels, and the researcher wishes to generalize to the population of levels. Examples will clarify this distinction.

First, assume that a researcher is interested in the efficacy of a proposed job training program as compared with the current program. The researcher could design a study containing one independent variable—program type—in which the two levels represent the precise question to be addressed. This variable would be considered fixed. If another researcher were to replicate the original investigator's work, he or she would use the same two groupings. Second, assume that a researcher is interested in the impact of trainers in a job training program. Further

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

assume that 50 individuals conduct the training sessions and that the researcher cannot include all of the trainers in the study. Instead she takes a random sample of 10 trainers and then conducts an analysis of the impact of trainers. This investigator wishes to draw inferences about the impact of the trainers not included in the study because the theoretical construct is the overall impact of trainers. If other researchers replicated this study, they would likely choose a different set of trainers. The trainer effect would be termed a random factor or effect; a random effects model should be used for the analysis.

In fixed effects models, inferences are limited to the specific levels chosen for the study; in random effects models, inferences go beyond the included levels. The method of analysis can differ, and different types of inferences can be drawn in the two situations. Mixed models include more than one independent variable; at least one is a fixed effect and at least one is a random effect. Because of space limitations, this chapter addresses only the most commonly used approach, the fixed effects model.

A completely randomized design is one in which subjects are assigned randomly to one and only one treatment level in the case of one independent variable. If two or more independent variables are involved, subjects are assigned randomly to one of the treatment combinations. Random assignment is a method of eliminating systematic effects other than the independent variables of interest. It does so by converting systematic sources of variability into random sources. Random assignment is not essential to the use of *ANOVA*, but it influences the interpretation of the results. Throughout the chapter, assume that random assignment has been used and that the design is balanced; that is the number of subjects is the same in each treatment condition.

Among completely randomized fixed effect designs, several models are possible. Those discussed here are oneway *ANOVA*, multiple comparison tests, and completely crossed factorial designs.

III. ONEWAY ANOVA

A. Two Groups

We begin with a completely randomized fixed effects model in which the investigator wishes to determine whether there is a difference between two groups. For instance, assume that a team of researchers wishes to determine the efficacy of a new job training program relative to the one currently in use. Five subjects are assigned randomly to a demonstration program of the new approach and five are assigned randomly to the current program. Upon completion of the program, the research team records the hourly wage rate for the first job placement. Let Y_{ij} represent the starting hourly wage rate for the ith person in the jth group, where i ranges from 1 through 5 and j ranges from 1 (demonstration program) through 2 (current job training program). Let J represent the total number (2) of groups, and n_j represent the number of subjects in the jth group. The total number of subjects, $n_.$, in the study is 10. The sample mean for the first group is denoted by $\overline{Y}_1$: that for the second by $\overline{Y}_2$. The sample variances for two groups are denoted by s_1^2 and s_2^2. In the entire sample of 10 persons, the grand mean and the variance are denoted by $\overline{Y}$ and s^2 respectively. Table 1 displays the starting hourly wages for all 10 subjects.

Thus far the problem sounds like an example of an independent sample t-test. The null hypothesis in such a t-test is that the population means for the two groups are equal; the alternative is that they are not. To assess the significance of the sample means, one computes the average for each of the two groups and then the difference between these two averages. The

TABLE 1 Starting Hourly Wage: Two Group Case

Program	Hourly wage
Demonstration	$Y_{11} = 12.00$
	$Y_{21} = 11.00$
	$Y_{31} = 10.00$
	$Y_{41} = 13.00$
	$Y_{51} = 11.00$
Current	$Y_{12} = 7.00$
	$Y_{22} = 8.50$
	$Y_{32} = 5.50$
	$Y_{42} = 8.25$
	$Y_{52} = 8.50$

ratio of the difference between the means to the standard error is distributed as *t* (see Equation 1). One then can determine the statistical significance of the sample result.

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \tag{1}$$

Table 2 displays the results for the sample statistics of both groups and for the *t*-value. The results are statistically significant; therefore the null hypothesis of equal means is rejected. The hourly wages for participants who completed the demonstration program are $3.95 higher on average than for people trained in the current program. This approach to testing for mean differences should be familiar to all social scientists. We now recast the problem as an *ANOVA*.

The null and alternative hypotheses are identical to those for the independent sample *t*-test, but *ANOVA* focuses on the partitioning of the variance of the dependent variable. To observe this emphasis, let us first calculate the sample mean and variance for the 10 subjects. Equation 2 is used to calculate the grand mean:

$$\overline{Y} = \frac{\displaystyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij}}{\displaystyle\sum_{j=1}^{J} n_j}. \tag{2}$$

The first step in computing the variance for the set of 10 wages is to subtract the average wage from each of the values and to square these differences. The results are summed over all subjects. The result is termed the total sum of squares, *TSS*, and is represented in Equation 3.

TABLE 2 Sample Statistics and *t*-Value for Job Training Example

Sample statistics for the demonstration program	Sample statistics for the current program	*t*-Value	*p*-Value
$\overline{Y}_1 = 11.50$	$\overline{Y}_2 = 7.55$	$t = 4.76$	.001
$s_1^2 = 1.75$	$s_2^2 = 1.70$		

$$TSS = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2. \tag{3}$$

The variance of Y thus becomes:

$$s^2 = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2}{n_. - 1}. \tag{4}$$

ANOVA partitions this variability into two sources: that due to the differences among the means and that due to the variability within each group. The latter reflects the dispersion of the values in a group around that group mean.

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 + \sum_{j=1}^{J} n_j(\bar{Y})^2 \tag{5}$$

The total sum of squares equals the error sum of squares plus the between groups sum of squares.

$$TSS = ESS + BSS \tag{6}$$

Each term then can be transformed into a mean square, a variance, by dividing by the appropriate degrees of freedom.

$$\frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2}{n_. - 1} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n_. - J} + \frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{J - 1} \tag{7}$$

$$\frac{TSS}{n_. - 1} = \frac{ESS}{n_. - J} + \frac{BSS}{J - 1} \tag{8}$$

To test the null hypothesis of equal means, one calculates the ratio of the mean square between groups to the mean squared error. This ratio follows an F distribution with $J - 1$ and $n_. - J$ degrees of freedom. Table 3 is an *ANOVA* table that displays the sources of variation and the results of the F-test for the job training example. As with the t-test, one would reject the null of equality of means. The p-value is the same for both tests because the two tests are equivalent, F equals t^2.

The computational steps for a oneway completely randomized fixed factor (also called fixed effects) *ANOVA* are straightforward. First, calculate the between groups sum of squares. Second, calculate the error sum of squares. Third, divide each sum of squares by the appropriate degrees of freedom to calculate the mean squares. Fourth, divide the mean squared between groups by the mean squared error to calculate F. Then determine the significance of the result. Table 4 displays the equations for calculating the sums of squares and the mean squares in the format of an *ANOVA* table.

TABLE 3 Oneway ANOVA for Job Training Example

Source	df	Sum of squares	Mean square	F	p-Value
Between	1	39.01	39.01	22.61	.001
Error	8	13.80	1.72		
Total	9	52.81			

TABLE 4 ANOVA Formulae

Source	df	Sum of squares	Mean square	F
Between	$J - 1$	$\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2$	$\frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{J - 1}$	$\frac{\frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{J - 1}}{\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2}{n_. - J}}$
Error	$n_. - J$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2$	$\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2}{n_. - J}$	
Total	$n_. - 1$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2$		

B. More than Two Groups

Thus far the coverage of oneway completely randomized fixed effects *ANOVA* has addressed only cases in which the independent variable consists of two groups. In such cases, the independent sample *t*-test and *ANOVA* are equivalent, but what if the number of groups, J, is greater than 2? Let us build on the job training example. In addition to the demonstration program and the current program, assume that people also are assigned randomly to a third category in which they receive no job training. The data are displayed in Table 5.

In this case, the omnibus or overall *F*-test in the oneway *ANOVA* assesses the null hypothesis that all group means are the same. The alternative implies that the mean for at least one

TABLE 5 Starting Hourly Wage: Three Group Case

Program	Hourly wage
Demonstration	$Y_{11} = 12.00$
	$Y_{21} = 11.00$
	$Y_{31} = 10.00$
	$Y_{41} = 13.50$
	$Y_{51} = 11.00$
Current	$Y_{12} = 7.00$
	$Y_{22} = 8.50$
	$Y_{32} = 5.50$
	$Y_{42} = 8.25$
	$Y_{52} = 8.50$
None	$Y_{13} = 6.00$
	$Y_{23} = 6.00$
	$Y_{33} = 8.50$
	$Y_{43} = 5.00$
	$Y_{53} = 7.00$

TABLE 6 Sample Statistics for Job Training Example

Demonstration program	Current program	No job training	Whole sample
$\overline{Y}_1 = 11.5$	$\overline{Y}_2 = 7.55$	$\overline{Y}_3 = 6.50$	$\overline{Y} = 8.52$
$s_1^2 = 1.75$	$s_2^2 = 1.70$	$s_3^2 = 1.75$	$s^2 = 6.45$

group is different from the mean for at least one other group or combination of groups. The partitioning of the sums of squares and the steps taken in conducting the *F*-test are identical whether the number of groups is 2 or more than 2. The computational equations in Table 4 still apply. In this specific example, *J* becomes 3 and *n*. equals 15.

Table 6 presents the summary statistics for the sample; Table 7 displays the results of the *ANOVA*. The between group sum of squares represents the squared differences of all three group means from the grand mean, an the sum of squared error now adds the variability within the third group to the sum of the other two groups. The value of *F* is 20.55, and this result is statistically significant: the starting hourly wage differs across the groups. The example could readily be extended to independent variables containing even more groups.

C. Assumptions

Inferences based on the omnibus *F*-test in a fixed effects *ANOVA* rely on three assumptions: (1) the residuals are distributed normally; (2) the variances of the errors are the same for all groups; and (3) the residuals are independent of one another. The omnibus *F*-test in a oneway fixed factor *ANOVA* is robust to violations of the normality assumption. In other words, the true Type I and Type II error rates are insensitive to such violations. The omnibus *F*-test in a fixed factor *ANOVA* is also robust to violations of the homogeneity of variance assumption when the design is balanced and the samples are not small (Maxwell and Delaney, 1990).

In the case of unbalanced designs, the conclusions are quite different: even small departures from homogeneity can have a large impact on the Type I error rate. Hays (1994) also concludes that the F-test is not robust to simultaneous violations of the normality and the homogeneity of variance assumptions. The effects of the two violations appear to be additive (Glass and Hopkins, 1984); thus it is possible for one violation to cancel out the other. Prior work indicates that the omnibus *F*-test in a fixed effects *ANOVA* is sensitive to violations of the independence assumption. That conclusion appears unanimous. Hays (1994) finds that violations of this assumption can lead to true Type I error rates substantially different from the nominal .05 level. Maxwell and Delaney (1990) state that the difference can be dramatic.

A number of solutions may be adopted when one or more of these assumptions have been violated. For violations of the normality and/or homogeneity of variance assumptions, one might transform the data for the dependent variable to a more ''normal'' or at least symmetric distribution with stable variances. A transformation is a re-expression of each data value. This re-expression is achieved by exponentiating each data value to some power, *P*, that ranges from

TABLE 7 ANOVA Table for Job Training Example

Source	df	Sum of squares	Mean square	*F*	*p*-Value
Between	2	69.51	34.75	20.05	.0001
Error	12	20.80	1.73		
Total	14	90.31			

positive infinity to negative infinity. This sequencing of *P* is known as the ladder of power. In this context, the value of 0 requires taking the natural logarithm of the values. The farther *P* deviates from 1, the greater the effect of the transformation on the distribution of the values (Velleman and Hoaglin, 1981).

Although the use of transformations is quite common, the appropriateness of the approach is disputed. Interpretation of the results is a fundamental issue, and the equality of means in one metric does not guarantee equality in another. Maxwell and Delaney (1990) discuss these issues and cite studies that have contributed to the debate.

Another alternative is the use of nonparametric statistical procedures; these tests do not rely on normality assumptions. A nonparametric procedure that may be used in place of a oneway fixed factor *ANOVA* is the Kruskal-Wallis one-way analysis of ranks test which assesses the equality of medians across groups. First the data on the dependent variable are ranked and then analysis is performed on the ranked data. For the computational procedures, see *Nonparametric Statistics for the Behavioral Sciences*, by Siegal and Castellan (1988).

Cirincione et al. (1994) used the Kruskall-Wallis test to assess differences in arrest rates among four groups: prison inmates with no prior mental health history, prison inmates with prior mental health history, mental health patients with prior arrest history, and mental health patients with no prior arrest history. The dependent variable was the arrest rate—number of arrests per year at risk—following release into the community. The distributions of arrest rates for the four groups were highly skewed. The rates were extremely low for most of the subjects, but were high in some cases. The sample sizes were not the same across groups: the smallest sample size was 50 and the largest was 315. As a result of these properties, confidence in the validity of the omnibus *F*-test for a fixed factor *ANOVA* was quite low and the Kruskall-Wallis procedure was employed. Although nonparametric tests are a potential alternative to a fixed effects *ANOVA*, their primary drawback is their lack of power.

Another alternative to fixed factor *ANOVA* is the use of robust procedures such as the Brown and Forsythe test (Maxwell and Delaney, 1990). This procedure relies on an estimate of the mean squared error that accounts for different within group variances. The equation for *F* becomes:

$$F^* = \frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^{J} \left(1 - \left(\frac{n_j}{n_.}\right)\right) s_j^2}, \tag{9}$$

where s_j^2 is the error variance for the jth group. Other robust procedures are available as well. These methods have not been applied widely because they were developed recently (Maxwell and Dalaney, 1990).

IV. MULTIPLE COMPARISON TESTS

A. Overview

The omnibus *F*-test for a oneway *ANOVA* evaluates the claim that the population means for all groups are equal. Rejection of the null hypothesis implies that they are not all equal, but does not locate differences. An investigator, having determined that at least one difference exists, might wish to locate the significant difference or differences. Concurrently, when designing a study, a researcher might have theoretical reasons for focusing on particular differences rather

than on all possible differences. In the job training example cited above, which involves the demonstration program, the current program, and no job training, the researcher might be concerned with two questions: (1) Are the hourly wages for people participating in a job training program (either the demonstration or the current program) significantly different from the wages for people who receive no job training? and (2) Are the hourly wages for people participating in the demonstration program significantly different from the wages for people in the current program?

In either case, the researcher would wish to test for hypotheses regarding differences between specific subsets of means. These tests of subsets of groups are commonly termed comparisons or contrasts. Simple contrasts refer to pairwise comparisons in which one group is tested against another. The number of possible pairwise comparisons equals $J(J-1)/2$, where J refers to the total number of groups. The comparison between the demonstration program and the current program is an example. The null hypothesis is

$$\mu_1 - \mu_2 = 0. \tag{10}$$

General or complex contrasts involve more than two means. for example, a researcher might wish to compare the wages of people who receive no job training (Group 3) with the wages of those in a job training program (Groups 1 and 2). In this example, one group is compared to a combination or an average of two groups. The null hypothesis is

$$\frac{\mu_1 + \mu_2}{2} - \mu_3 = 0. \tag{11}$$

The symbol designated to represent the ith contrast is ψ_i. Each ψ_i addresses a research question. In the job training example, two comparisons are identified:

$$\psi_1 = \mu_1 - \mu_2 = 0 \tag{12}$$

and

$$\psi_2 = \frac{\mu_1 + \mu_2}{2} - \mu_3 = 0. \tag{13}$$

One also should view each contrast as a linear combination of group means. Equations 12 and 13 can be rewritten as

$$\psi_1 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 = 0 \tag{14}$$

and

$$\psi_2 = \left(\frac{1}{2}\right)\mu_1 + \left(\frac{1}{2}\right)\mu_2 - (1)\mu_3 = 0. \tag{15}$$

Both equations can be written as

$$\psi_i = \sum_{j=1}^{J} c_j\mu_j. \tag{16}$$

A coefficient, c_j, is associated with the population mean of a given group, j. The coefficients sum to 0 and not all coefficients are 0.

$$\psi_1 = 1 - 1 + 0 = 0 \tag{17}$$

and

$$\psi_2 = \frac{1}{2} + \frac{1}{2} - 1 = 0. \tag{18}$$

Therefore

$$\sum_{j=1}^{J} c_j = 0. \tag{19}$$

Researchers have a wide selection of approaches for testing these contrasts. The choice depends primarily on three factors:

- Type of control over Type I error desired,
- Planned versus post hoc comparisons, and
- Orthogonal versus nonorthogonal sets of contrasts.

Researchers must determine an acceptable approach to control the probability of committing a Type I error. For example, one could conduct a series of *t*-tests and set *ALPHA* to .05. This would be the per comparison Type I error rate, for which the symbol is $ALPHA_{PC}$. The experimentwise error rate, $ALPHA_{EW}$, is the probability of committing at least one Type I error. If one were to test a number of contrasts, the Type I error rate for a given test would be .05 but the probability of making at least one Type I error would be higher. The experimentwise error rate is a function of the number of contrasts to be tested and the per comparison error rate chosen. If the *K* contrasts are independent of one another, it can be shown that

$$ALPHA_{EW} = 1 - (1 - \text{ALPHA}_{pc})^K. \tag{20}$$

The experimentwise Type I error rate rises quickly with an increase in the number of independent comparisons. Assuming a per comparison Type I error rate of .05, the experimentwise Type I error rate for 3, 10, and 20 independent contrasts would be .14, .40, and .64 respectively.

In conducting an *ANOVA* and investigating a set of specific comparisons, researchers must choose an $ALPHA_{PC}$, which in turn determines $ALPHA_{EW}$. The choice is based on the determination of the appropriate balance between Type I and Type II errors. If a low $ALPHA_{PC}$ is chosen to produce a low $ALPHA_{EW}$, then the probability of at least one Type II error—incorrectly failing to reject a null hypothesis—increases. The various statistical approaches for assessing multiple contrasts address this tradeoff differently; researchers must be aware of these differences in choosing a method.

Multiple comparison approaches also can be characterized as either planned or post hoc. Planned or a priori comparisons are made when researchers wish to test specific contrasts—research questions—before conducting the analysis. The planned tests should be essential to the study design and the purpose of the study. Planned comparisons usually are performed instead of the omnibus *F*-test in the *ANOVA*. Post hoc, a posteriori, or incidental methods for assessing multiple are conducted after rejection of the null of equal population means by the overall *F*-test. The post hoc approaches usually assess all or a large number of possible contrasts.

An advantage of the planned comparison approach is that researchers are less likely to capitalize on chance. They conduct only a small number of tests rather than searching for differences wherever they might be. In the comparisons between planned and post hoc procedures that are governed by the data, planned comparisons offer greater control over experimentwise Type I error rates. Researchers sometimes allow the data to suggest the comparison to be tested. In these data-driven cases, they are likely to focus on the groups with the greatest observed mean difference. This approach is not appropriate, however, because the sampling distribution

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

of the difference between the highest and the lowest group means is not the same as the sampling distribution for any two groups—for example, groups 1 and 2.

Sets of contrasts can be orthogonal or nonorthogonal. If the normality and homogeneity of variance assumptions hold, orthogonality means independence (Hays, 1994). If two contrasts are independent, the conclusion of rejection or failure to reject the null hypothesis in one contrast is not related to the conclusion reached in the other. Orthogonal contrasts offer greater control over the number of Type I errors one may make. In a set of orthogonal contrasts in which a Type I error has been made for a given contrast, the chances of making another Type I error are unchanged. In the case of nonorthogonal contrasts, multiple Type I errors are likely if at least one such error has been made.

In choosing between orthogonal and nonorthogonal sets of contrasts, researchers must weigh statistical and substantive concerns. Although orthogonal comparisons offer control over the number of Type I errors and although interpretation may thus be enhanced, the research questions of interest may not be orthogonal. These questions should determine the analysis.

B. Multiple Comparison Methods

Researchers may use a number of methods in assessing the contrasts of interest, such as planned contrasts, Dunn's test, Fisher's least significant Difference (LSD), and the Scheffe test.

1. Planned Contrasts

When researchers wish to test a small number of contrasts of theoretical interest, they should employ planned contrasts. Planned contrasts are performed instead of the omnibus *F*-test. In the example of the job training programs, assume that the researcher wishes to test two research questions: ''Are hourly wages for people in the demonstration program the same as the wages for people in the current program?'' and ''Are hourly wages for people in a job training program, either the demonstration or the current program, the same as those for people who receive no job training?'' The two null hypotheses to be tested are

$$H_0: \psi_1 = 1\mu_1 - 1\mu_2 + 0\mu_3 = 0 \tag{21}$$

and

$$H_0: \psi_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - 1\mu_3 = 0. \tag{22}$$

Testing each hypothesis requires estimation of the contrast values, $\hat{\psi}_1$ and $\hat{\psi}_2$. The test statistic used follows a *t*-distribution with $n_\cdot - J$ degrees of freedom. The equation is

$$t = \frac{\hat{\psi} - \psi}{\sqrt{MS_e \sum_{j=1}^{J} \frac{c_j^2}{n_j}}}, \tag{23}$$

where

$\hat{\psi}$ is the value of the contrast estimated in the sample;
ψ is the value hypothesized in the null;
MS_e is the mean squared error, the same as that computed in the omnibus *F*-test;
c_j is the coefficient for the *j*th group; and
n_j is the sample size for the *j*th group.

On the basis of the job training example, the resulting *t*-values for the two contrasts are 4.74 and 4.20 respectively. Both contrasts are statistically significant. From the first contrast, one may conclude that wages on average are $3.95 per hour more for people in the demonstration program than for people in the current program. From the second contrast, one may conclude that wages for people in a job training program, either the demonstration or the current program, are on average $3.02 per hour higher than for those who received no job training.

2. *Dunn's Test*

Dunn's test, also known as a Bonferroni test, requires researchers first to specify the desired experimentwise error rate—say, .05. One then distributes this error rate evenly across the comparisons:

$$\frac{ALPHA_{EW}}{K} = ALPHA_{PC}. \tag{24}$$

The job training example involves two contrasts. In the Dunn test, one would set the per comparison Type I error rate to .025 rather than .05. The critical values of the *t*-statistics also would change. With this adjustment, the Dunn test would set a .05 limit to the experimentwise error rate.

On the basis of the results of the planned contrasts for the job training example, the researcher still would reject both null hypotheses because both *p*-values are less than .025. The Dunn test is a commonly used post hoc procedure. If the number of contrasts is large, however, the Dunn test is quite conservative. For example, if a researcher wished to test 10 contrasts with an experimentwise error rate of .05, the per comparison rate would be .005. This multiple comparison test is less powerful than other multiple comparison approaches.

To address the conservatism of the Dunn test, modifications to the procedure have been developed (Keppel, 1982; Toothaker, 1991). One approach is to arrange the contrasts on the basis of *p*-values. The test of the contrast with the lowest *p*-value is the same equation as used in the unmodified Dunn test. If the null is rejected, one then moves to the contrast with the next smallest *p*-value. The second test is based on the equation

$$\frac{ALPHA_{EW}}{K - 1} = ALPHA_{PC}. \tag{25}$$

The cutoff for this test is less stringent than for the first. If the result is significant, one proceeds to the contrast with the next smallest *p*-value and uses

$$\frac{ALPHA_{EW}}{K - 2} = ALPHA_{PC}. \tag{26}$$

If the result is significant, one continues in this way until the full set of contrasts has been tested or until the null has not been rejected. Other modifications to the Dunn test are possible as well (Keppel, 1982; Toothaker, 1991); each is based on partitioning $ALPHA_{EW}$ in some way other than evenly across the full set of contrasts.

3. *Fisher's LSD Test*

Fisher's least significant difference (LSD) test, also known as a protected *t*-test, requires researchers to first test the null of equal population means with the overall omnibus *F*-test. If the results are not significant, testing stops. If the results are significant, further post hoc tests are

conducted. To test all pairwise comparisons, one simply conducts the necessary number of *t*-tests using Equation 23. $ALPHA_{PC}$ is set to the desired level—say, .05—for each test.

The logic of the test is that the use of the overall *F*-test is a control for experimentwise Type I error. In other words, the procedures contains a set of ''protected *t*-tests.'' Carmer and Swanson (1973) conclude that the procedure provides a balance between Type I and Type II error rates. Maxwell and Delaney (1990) disagree, however; they argue that the procedure provides no control over Type I error in the second stage and thus the method should not be used.

4. The Scheffe Test

The Scheffe test is an exploratory approach that may be used to test both simple and complex contrasts. It controls the experimentwise error rate by adjusting the critical value for the hypothesis test. The *t*-test for a contrast is shown in Equation 23. In a typical *t*-test, one then would refer to the *t*-distribution with $n_{\cdot} - J$ degrees of freedom to determine the cut-off or critical value of *t*. As discussed earlier, the relationship between the *t*-distribution and the *F*-distribution is such that *t*, with *v* degrees of freedom, equals the square root of *F* with (1, *v*) degrees of freedom. Therefore, in the usual *t*-test, one rejects the null hypothesis if $t_{n-J} > \sqrt{F_{1,n-j}}$. In the Scheffe test, however, the critical value of *F* is adjusted to reflect the number of means tested. The appropriate decision rule is to reject null if the sample based value of *t* exceeds the adjusted critical value of *F*, as given in this equation:

$$t > \sqrt{(J-1)F_{J-1,n-J}}. \tag{27}$$

In the job training example, the number of groups is 3 and the critical value of *F* is 3.89. The appropriate critical value of *t* is the square root of 7.78, or 2.79. One then would conduct the various *t*-tests.

In general, the Scheffe test is the most conservative approach for pairwise comparisons. Hays (1994) states that this test is insensitive to departures from normality, but Maxwell and Delaney (1990) conclude that it is not robust to violations of the homogeneity of variance assumption. In such cases, the procedure may be modified.

V. FACTORIAL DESIGNS

A. Introduction

ANOVA is not limited to studies involving only one independent variable. When the study is based on two or more independent variables, a factorial design *ANOVA* can be used. If there are two independent variables, the researcher conducts a two factor or two-way *ANOVA* while a *z*-factor or *z*-way *ANOVA* involves *z* independent variables. Factorial designs also are referenced by the patterns of groups or levels contained in the factors. For example, a 2×3 (read ''2 by 3'') factorial design includes two factors; the first has two groups and the second has three. In a completely randomized fixed effects factorial design, all independent variables are fixed and all subjects are assigned randomly to one and only one treatment combination. Although designs exist to handle studies in which each treatment combination contains only one subject, each treatment combination usually contains at least two subjects.[1] The independent variables are completely crossed; all possible combinations of groups or treatments exist for all independent variables. If each treatment combination contains an equal number of subjects, the design is said to be balanced.

TABLE 8 Data Layout and Means for Factorial Design ANOVA

	Factor 2												
	Group 1				Group 2				Group 3				Total
Group 1	Y_{111}	Y_{711}	Y_{1311}	Y_{1811}	Y_{112}	Y_{712}	Y_{1312}	Y_{1812}	Y_{113}	Y_{713}	Y_{1313}	Y_{1813}	$\overline{Y}_{1\cdot}$
	Y_{211}	Y_{811}	Y_{1411}	Y_{1911}	Y_{212}	Y_{812}	Y_{1412}	Y_{1912}	Y_{213}	Y_{813}	Y_{1413}	Y_{1913}	
	Y_{311}	Y_{911}	Y_{1511}	Y_{2011}	Y_{312}	Y_{912}	Y_{1512}	Y_{2012}	Y_{313}	Y_{913}	Y_{1513}	Y_{2013}	
	Y_{411}	Y_{1011}	Y_{1611}	Y_{2111}	Y_{412}	Y_{1012}	Y_{1612}	Y_{2112}	Y_{413}	Y_{1013}	Y_{1613}	Y_{2113}	
	Y_{511}	Y_{1111}	Y_{1711}	Y_{2211}	Y_{512}	Y_{1112}	Y_{1712}	Y_{2212}	Y_{513}	Y_{1113}	Y_{1713}	Y_{2213}	
	Y_{611}	Y_{1211}			Y_{612}	Y_{1212}			Y_{613}	Y_{1213}			
		$\overline{Y}_{11}$				$\overline{Y}_{12}$				$\overline{Y}_{13}$			
Factor 1													
Group 2	Y_{121}	Y_{721}	Y_{1321}	Y_{1821}	Y_{122}	Y_{722}	Y_{1322}	Y_{1822}	Y_{123}	Y_{723}	Y_{1323}	Y_{1823}	$\overline{Y}_{2\cdot}$
	Y_{221}	Y_{821}	Y_{1421}	Y_{1921}	Y_{222}	Y_{822}	Y_{1422}	Y_{1922}	Y_{223}	Y_{823}	Y_{1423}	Y_{1923}	
	Y_{321}	Y_{921}	Y_{1521}	Y_{2021}	Y_{322}	Y_{922}	Y_{1522}	Y_{2022}	Y_{323}	Y_{923}	Y_{1523}	Y_{2023}	
	Y_{421}	Y_{1021}	Y_{1621}	Y_{2121}	Y_{422}	Y_{1022}	Y_{1622}	Y_{2122}	Y_{423}	Y_{1023}	Y_{1623}	Y_{2123}	
	Y_{521}	Y_{1121}	Y_{1721}	Y_{2221}	Y_{522}	Y_{1122}	Y_{1722}	Y_{2222}	Y_{523}	Y_{1123}	Y_{1723}	Y_{2223}	
	Y_{621}	Y_{1221}			Y_{622}	Y_{1222}			Y_{623}	Y_{1223}			
		$\overline{Y}_{21}$				$\overline{Y}_{22}$				$\overline{Y}_{23}$			
Total		$\overline{Y}_{\cdot 1}$				$\overline{Y}_{\cdot 2}$				$\overline{Y}_{\cdot 3}$			$\overline{Y}$

B. Two-Factor ANOVA

The discussion of factorial design begins with the simplest case, a two factor *ANOVA*. Table 8 displays the factorial layout of a 2×3 completely crossed and randomized fixed effects factorial design *ANOVA*. Each of the two rows refers to a group or level of the first independent variable, Factor 1; each of the three columns refers to a level of Factor 2. Each cell or box represents a treatment or group combination. For example, the cell in the upper left hand corner refers to subjects in Group 1 of Factor 1 and Group 1 of Factor 2. Each level of Factor 1 occurs in each level of Factor 2, and vice versa. That is the design is completely crossed. The researcher randomly assigns 132 subjects such that each treatment combination contains 22 subjects. The value of the dependent variable for the ith person in the jth group of Factor 1 and in the kth group of Factor 2 is represented by Y_{ijk}.

The goal of factorial design *ANOVA* is to test the effect of the independent variables, both individually and taken together. The *ANOVA* decomposes the variability in the dependent variable into four sources: (a) the effect of Factor 1 alone; (b) the effect of Factor 2 alone; (c) the interaction effect of Factor 1 and Factor 2 taken together; and (d) error variance.

The main effect for a factor is the effect of membership in a level or group of that factor. The main effect of Factor 1 in Table 8, for example, can be regarded as a row effect while the main effect of Factor 2 can be considered as a column effect. An interaction effect is present when the impact of one factor depends on the level of the other factor. In other words, the effect of one factor is not the same for all levels of the second factor. The interaction is the effect of belonging to the j, kth group combination over and above the effect of being situated in the jth row and the kth column.

One can think of interaction effects as nonadditive or contingent. For example, one can modify the job training example used in the discussion of oneway *ANOVA* to be a 2×3 factorial design. Assume that Factor 1 is job training program (demonstration or current) and Factor 2 is length of time in training (one, two, or three months). One might find that the effect of time in training depends on the program. For the demonstration program, higher wages might be

TABLE 9 Equations for Two-Way, Fixed Effects ANOVA

Source	df	Sum of squares*	Mean square	F
Main effect Factor 1	$J - 1$	$nK \sum_{j=1}^{J} (\bar{Y}_{j\cdot} - \bar{Y})^2$	$\frac{SS_1}{J - 1}$	$\frac{MS_1}{MS_e}$
Main effect Factor 2	$K - 1$	$nJ \sum_{k=1}^{K} (\bar{Y}_{\cdot K} - \bar{Y})^2$	$\frac{SS_2}{K - 1}$	$\frac{MS_2}{MS_e}$
Interaction effect	$(J - 1)(K - 1)$	$n \sum_{k=1}^{K} \sum_{j=1}^{J} [\bar{Y}_{jk} - \bar{Y}_{j\cdot} - \bar{Y}_{\cdot K} + \bar{Y}]^2$	$\frac{SS_{1\times 2}}{(J - 1)(K - 1)}$	$\frac{MS_{1\times 2}}{MS_e}$
Error	$JK(n - 1)$	$\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} (Y_{ijk} - \bar{Y}_{jk})^2$	$\frac{SS_e}{JK(n - 1)}$	
Total	$n_{\cdot} - 1$	$\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} (Y_{ijk} - \bar{Y}_{jk})^2$		

* n refers to the number of subjects in each cell in the factorial design. A balanced design is assumed.

associated with longer periods of training. For the current program, the average wage might be the same regardless of the number of months in training.

In two-way *ANOVA*, one typically tests three null hypotheses: one for each of the two main effects and one for the interaction effect. One must partition the variability of the dependent variable to test these hypotheses. Table 9 displays the equations necessary to compute the various sums of squares, mean squares, and *F*-ratios. The main effect of Factor 1 is based on the differences between the row means and the grand mean, while the main effect of Factor 2 is based on the differences between the column means and the grand mean. Derivation of the sum of squares for the interaction is less intuitive: One starts with the concept of a cell effect, which is based on the difference between the each cell mean and the grand mean weighted by the number of subjects in each cell. The cell effect represents the combined or composite effects (the two main effects and the interaction effect taken together) of the factors. To isolate the interaction effect, one must remove the influence of the two main effects by subtracting them from the composite effect. The *F*-test for each of the three hypotheses then is based on a ratio of the mean square due to the effect of interest to the mean squared error.

An example will clarify the use of two-way *ANOVA*; this is an adaptation of an experiment conducted by Cirincione (1992). Assume that a research team wishes to conduct a needs assessment for treatment of drinking problems. The assessment is made for U.S. counties, and the methodology rests on the construction of a model based on expert judgment. To construct the model, experts first identify the important need indicators, namely a county's alcohol-related death rate and annual sales of alcoholic beverages per capita (to persons age 21 and over). The research team then wishes to determine the relative importance of each indicator in predicting need for treatment. To make this determination, each expert judges vignettes that describe hypothetical counties. Each description includes the per capita consumption rate and the alcohol-related death rate for the county. The research team next asks the experts to evaluate the need for treatment for the county in each vignette, and then statistically estimates the relative importance of the two indicators.

The research team is concerned with the potential impact of the way in which information

TABLE 10 Sample Means for Relative Importance of Alcohol-Related Death Rates in Needs Assessment

		Response scale			
		1–25 Anchored	1–25 unanchored	1–25 Percent	Totals
Definition of death rate	Rates	43.909	54.182	47.045	48.379
	Likert	54.318	60.409	51.500	55.405
	Totals	49.114	57.295	49.273	51.894

is presented in the vignettes and with the response scale used by the experts. Accordingly, the researchers conduct an experiment that tests these two factors. The first factor, definition of indicators, consists of the use of two different definitions for a alcohol-related death rates. For half of the experts, this cue is defined as the number of alcohol-related deaths (including cirrhosis, alcohol psychosis, alcohol-related homicide or suicide, and alcohol-related vehicular fatalities) per 100,000 persons in a county; the values range from 15 to 75. For the other half of the experts, the alcohol-related death rate is defined on a Likert scale with 1 representing a very low rate and 7 representing a very high rate. For all vignettes, alcoholic beverage sales are measured in dollars; the values range from $50 to $250 per resident age 21 and older.

The second factor tested by the research team is response scale. The research team assigns the experts randomly to one of three response scale conditions: 1–25 anchored (Group 1), 1–25 unanchored (Group 2), and 1%–25% (Group 3). For the first response scale, the subjects are asked to evaluate the need for treatment on a scale of 1 to 25: For the set of vignettes they are to judge, they must assign a 1 to the county they believe has the lowest need and a 25 to the county they believe has the highest need. The second response scale is a 25 point Likert scale with no anchor requirement. The third response scale is based on the estimated percentage of each county's population (21 and over) in need of treatment of drinking problems. Estimates of the relative importance of alcohol-related death rates are derived from these judgments.[2] A total of 132 experts participated, with 22 assigned to each of the six combinations in the 2×3 factorial design. Table 10 displays the sample statistics; Table 11 contains the results for the *ANOVA*.

Both main effects are significant. When the description is rendered on a Likert scale, the relative importance of alcohol-related death rates is 7.026 point higher on average than when it is presented on a ratio scale. Concurrently, the average relative weight of alcohol-related death rates is 8.102 points higher when the 1–25 unanchored scale is used. The interaction effect is not significant; the impact of indicator definition does not depend on the response scale used, and vice versa. The *ANOVA* results suggest that when information is presented and elicited in a format that may not be solidly anchored or well grounded, the importance of alcohol-related death rates relative to alcoholic beverage sales increases. The findings suggest that in tasks

TABLE 11 Two-way ANOVA Table for Needs Assessment Study

Source	df	SS	MS	*F*	*p*-Value
Indicator definition	1	1631.03	1631.030	5.587	.020
Response scale	2	1926.20	963.098	3.299	.040
Interaction	2	205.65	102.826	.352	.704
Error	126	36783.64	291.934		

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

involving somewhat ambiguous information researchers should use multiple methods, provide feedback to the experts, and resolve any differences that arise before employing models developed in this manner for needs assessment. Otherwise the conclusions may be an artifact of the method of presentation and/or of elicitation.

C. Multiple Comparisons

As with oneway *ANOVA*, an investigator may wish to test a subset of specific comparisons when the study includes more than one independent variable. As in the one factor case, control of Type I error rate is a concern, but a different method is typically used to handle the problem. In the oneway *ANOVA*, the two types of error rates are the probability of making a Type I error for a given comparison, $ALPHA_{PC}$ (the per comparison error rate), and the probability of making at least one Type I error across all hypothesis tests in the study, $ALPHA_{EW}$ (the experimentwise error rate).

In a two factor *ANOVA*, three hypotheses usually are tested with *F*-tests. Each of the main effects and the interaction effect can be regarded as a representation of a ''family'' of comparisons. In the needs assessment example, only one comparison—rate versus Likert scale—exists for Factor 1 because there are only two groups. Factor 2 offers three possible pairwise comparisons and three complex comparisons. The interaction effects can be tested through several contrasts. In factorial design *ANOVA*, researchers usually control Type I error rates based on families of tests, $ALPHA_{FW}$ (the familywise error rate), rather than on an experimentwise basis. This approach results in an upper limit of $ALPHA_{EW}$ equal to αE where E is the number of families in the *ANOVA*.

Each of the multiple comparison methods discussed in the connection with oneway *ANOVA* can be extended to factorial design *ANOVA*. Tests for the main effects are straightforward. One may treat them as if the experiment were broken down into a series of oneway *ANOVA*s with each series representing a family. To test interaction effects, the investigator can assess the consistency of the effect of one given factor for two or more levels of the second factor.

D. Factorial Designs: Assumptions and Unequal Cell Sizes

The assumptions regarding the residuals for the two factor *ANOVA*, as well as for higher-order factorial designs, are the same as those for the oneway *ANOVA*. The errors must be distributed normally with a mean of 0 and a constant variance. Furthermore, the error term in any given treatment combination must be unrelated to all other residuals in that treatment combination and to the residuals in other treatment combinations.

This discussion has been limited to studies employing balanced factorial design; that is, each treatment combination contains the same the number of cases. Balanced designs guarantee that the main and interaction effects are orthogonal. If they are not orthogonal, the effects are correlated. The formal model for the unbalanced factorial design is the same for as balanced designs, but it is not possible to assess each effect separately from the others. To address this problem, a number of alternative approaches (types of sums of squares) are possible.

E. Higher Order Factorial Designs

Factorial design *ANOVA* is not limited to two independent variables. The addition of each new factor significantly increases the complexity of the design and the number of hypotheses to be tested. Three factors entail three main effects (Factor 1, Factor 2, and Factor 3), three two-way

interactions (Factor 1 × Factor 2, Factor 1 × Factor 3, and Factor 2 × Factor 3), and one three-way interaction (Factor 1 × Factor 2 × Factor 3). A four factor *ANOVA* involves four main effects, six two-way interactions, four three-way interactions, and one four-way interaction. This pattern continues for each additional factor. The computational procedures are not much more complicated than in the case of two factors. The ease of interpretation however, is often lost, and the required sample sizes are much greater.

VI. ANOVA AND MULTIPLE REGRESSION

Here we examine briefly the parallels between *ANOVA* and ordinary least squares linear regression. In investigating the relationship between a dependent variable and one or more independent variables, researchers must develop a theory of this relationship. The general linear model is a mathematical expression positing that the value of the dependent variable is a linear combination of the independent variables and random error. Equations 28 and 29 display a typical layout of this model.

$$Y_i = a + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_J X_{iJ} + \varepsilon_i \tag{28}$$

and

$$= a + \sum_{j=1}^{J} \beta_j X_{ij} + \varepsilon_i, \tag{29}$$

where

- Y_i is the observed value of the dependent variable for the ith person in the study;
- α is a constant representing the y-intercept. It can be viewed as the effect of independent variables held constant in the study;
- β_j is the slope associated with the jth independent variable. It represents the expected change in the dependent variable for a one unit change in the value of X_j holding all other independent variables constant;
- X_{ij} is the value for the ith person on the jth independent variable; and
- ε_i is the error or residual for the ith person.

One can use ordinary least squares to estimate the equation.

The reader should recognize Equation 29 as the regression model. Using the linear regression model to conduct an *ANOVA* requires appropriate coding of the variables and appropriate calculation of the *F*-ratios for the test statistic.

An example will clarify this parallel. Suppose a researcher is interested in comparing attitudes about the importance of monetary rewards across three sectors of the economy: public, private, and nonprofit.[3] She collects a random sample of 6 public sector, 6 private sector, and 6 nonprofit sector supervisors, and then conducts a oneway *ANOVA* to assess group differences. The grand mean is 33.3; the means for the public, private, and nonprofit sector supervisors are 33.0, 43.0, and 24.0 respectively. The value of the omnibus *F*-test is 3.8937, which is statistically significant. Thus the importance of monetary rewards is not the same across the three sectors.

To conduct the same test using a regression model, one must determine how to represent the factor levels as independent variables. In regression models that include an intercept, the number of independent variables needed to represent a factor with J levels is $J - 1$. Thus, in this example, two independent variables are required. Also one must choose a mathematical

TABLE 12 Dummy Coding for Single Factor

		Dummy coding	
Y	Sector	X_1	X_2
48	Public	1	0
42	Public	1	0
36	Public	1	0
30	Public	1	0
21	Public	1	0
21	Public	1	0
61	Private	0	1
49	Private	0	1
52	Private	0	1
40	Private	0	1
31	Private	0	1
25	Private	0	1
30	Nonprofit	0	0
42	Nonprofit	0	0
21	Nonprofit	0	0
21	Nonprofit	0	0
18	Nonprofit	0	0
12	Nonprofit	0	0

coding scheme for the variables. Many such schemes exist, but the most commonly used is dummy variable coding. Table 12 displays an example of this approach.

Dummy variable coding, also known as indicator variable coding, requires the investigator to represent $J - 1$ of the groups as independent variables, each based on a 0/1 coding. The group not represented by an independent variable is called the omitted or reference group. The value 1 indicates that the subject is a member of the level represented by a given independent variable; the value 0 indicates that the subject is not a member of that group. The cases in the omitted group are coded 0 for all $J - 1$ independent variables. All other cases are coded 1 for one and only one independent variable. Any of the three groups can be chosen to be the omitted category; in this example, the nonprofit sector is omitted. The variable X_1 represents public sector supervisors; the variable X_2 represents private sector supervisors.

Table 13 displays the results of the ordinary least squares regression model. The overall test in multiple regression is a test of whether the coefficient of determination, R^2, is equal to 0. In this case, it determines whether a relationship exists between sector of employment and the importance of monetary rewards. The relationship is captured by a linear combination of

TABLE 13 Regression Results

	β_j	*p*-Value
X_1	9.00	.2062
X_2	19.00	.0138
Constant	24.00	.0002
R^2	.342	
F-Ratio	3.89	
p-Value	.04	

the $J - 1$ independent variables. In this case, sector of employment can predict 34.2% of the variability in the dependent variable. Testing the null ''no relationship'' results in an *F*-ratio of 3.89 and a *p*-value of .04. These results are identical to those for the omnibus test in the *ANOVA* model, because the two tests are the same.

One can use the equation to estimate the group means. For instance,

Public Sector:
$9.00(1.0) + 19.00(0.0) + 24.00 = 33.00$

Private Sector:
$9.00(0.0) + 19.00(1.0) + 24.00 = 43.00$

Nonprofit Sector
$9.00(0.0) + 19.00(0.0) + 24.00 = 24.00$

From these results, one can see that the *y*-intercept represents the average value of the omitted group and that each regression weight represents the average difference between the group identified by the independent variable and the omitted group. The tests of the regression weights are pairwise multiple comparison tests of each group against the omitted group. In this case, we find no statistically significant difference between supervisors in the public and the nonprofit sectors. The difference between the private and the nonprofit sectors, however, is statistically significant.

This presentation of the use of regression analysis to perform a oneway *ANOVA* assumes a balanced design. As in the case of oneway *ANOVA*, balanced designs lead to more robust regression models. Among unbalanced designs, unless the lack of balance is due to some extraneous variables not addressed by the investigator, the internal validity of the study is not affected. Unequal sample sizes that result from random sampling and that represent the population proportions of group membership indeed may be preferable. As Pedhazur (1982) argues, such designs are better suited to measure the magnitude of the relationship between the dependent variable and the factor. If one wishes to compare groups, the balanced design is preferable. In either case, the statistical procedures are the same.

Regression analysis also can be used to assess factorial designs. Because main effects and interaction effects are represented by multiple independent variables, one cannot readily assess the significance of the effects according to the significance of the regression weights. Instead, a model comparison approach can be used.[4]

The logic of model comparison is straightforward. First run the regression model with all the independent variables included, and calculate the coefficient of determination, R_f^2. Then run a second model with the following restriction: Do not include in the model the independent variables associated with the effect being tested, and calculate the coefficient of determination, R_r^2. The test of the effect is a test of the change in the coefficient of determination due to the effect and the test statistic is

$$F = \frac{(R_f^2 - R_r^2)}{(1 - R_f^2)} \times \frac{n - p_f - 1}{p_f - p_r}. \tag{30}$$

Where:

R_f^2 is the coefficient of determination for the full model;
R_r^2 is the coefficient of determination for the restricted model;
n is the total number of subjects in the model;
p_f is the number of independent variables in the full model; and
p_r is the number of independent variables in the restricted model.

This brief discussion should demonstrate that *ANOVA* and regression models, when coded properly, can be used to test many of the same research questions.

VII. CONCLUSIONS

The proper use of *ANOVA* forces researchers to grapple with several issues when designing studies. The limitations regarding the internal validity of correlation based conclusions from observational studies are much greater than those for conclusions drawn from experimental research. Although the procedures and statistics in *ANOVA* are the same in each case, the conclusions that one may properly draw depend on the research design. *ANOVA* also highlights the issues of Type I error, Type II error, and statistical power; it was developed as a means of dealing with inflated Type I error rates due to multiple *t*-tests. These issues must be addressed whenever a researcher tests multiple null hypotheses, not only when an *ANOVA* procedure is used.

Knowledge of *ANOVA* methods also should improve one's comprehension of regression analysis. The use of categorical independent variables and the proper coding of such variables can be understood in the *ANOVA* framework. The overall test of fit of the regression model is the same as the test of the composite effects in the *ANOVA* model. The tests of the regression weights are the same as multiple comparison tests.

Although public administration and public policy scholars use *ANOVA* less often than some other statistical methods, they would be well served by a knowledge of *ANOVA* methods. If public administration research is to move toward theory testing and more rigorous work, the principle of *ANOVA* methods and designs must be understood.

NOTES

1. Readers interested in single subject designs may consult Iverson and Norpoth (1987) and Montgomery (1991).
2. In the case of only two indicators, researchers need derive the relative importance of only one indicator to fully identify the system of relative weights.
3. This example is based on two studies (Edwards et al., 1981; Emmert and Crow, 1988).
4. In factorial designs that account for interaction effects, coding schemes (e.g., effect and contrast coding) other than dummy variable coding are preferable. Dummy variable coding leads to artificial correlations between main and interaction effects.

REFERENCES

Brown, G. and C.C. Harris (1993). ''The Implications of Work Force Diversification in the U.S. Forest Service,'' *Administration and Society, 25*(1): 85–113.

Carmer, S.G. and M.R. Swanson (1973). ''An Evaluation of Ten Multiple Comparison Procedures by Monte Carlo Methods,'' *Journal of the American Statistical Association*, 68: 66–74.

Cirincione, C. (1992). *The Integrated Contingency Model: Range-Sensitive Decision Models*, Unpublished Doctoral Dissertation, State University of New York at Albany, Albany.

Cirincione, C., H.J. Steadman, P.C. Robbins, and J. Monahan (1994). ''Mental Illness as a Factor in Criminality: A Study of Prisoners and Mental Patients,'' *Journal of Criminal Behavior and Mental Health, 4*(1): 33–47.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Edwards, J.T., J. Nalbandian, and K.R. Wedel (1981). ''Individual Values and Professional Education: Implications for Practice and Education,'' *Administration and Society*, *13*(2): 123–143.

Emmert, M.A. and M.M. Crow (1988). ''Public, Private and Hybrid Organizations: An Empirical Examination of the Role of Publicness,'' *Administration and Society*, *20*(2): 216–244.

Frank, H. (1990). ''Municipal Revenue Forecasting with Time Series Models: A Florida Case Study,'' *American Review of Public Administration*, *20*(1): 45–57.

Gianakis, G.A. and H.A. Frank (1993). ''Implementing Time Series Forecasting Models: Considerations for Local Governments,'' *State and Local Government Review*, *25*(2): 130–144.

Glass, G.V. and K.D. Hopkins (1984). *Statistical Methods in Education and Psychology*, Second ed., Englewood Cliffs: Prentice-Hall, Inc.

Hays, W.L. (1994). *Statistics*, Fifth ed, Fort Worth: Harcourt Brace College Publishers.

Herman, R.D. and R.D. Heimovics (1990). ''An Investigation of Leadership Skill Differences in Chief Executives of Nonprofit Organizations,'' *American Review of Public Administration*, *20*(2): 107–124.

Iverson, G.R. and H. Norpoth (1987). *Analysis of Variance*, Vol. 1, Newbury Park: Sage Publications.

Keppel, G. (1982). *Design & Analysis*: *A Researcher's Handbook*, Second ed., Englewood Cliffs: Prentice-Hall.

Klammer, T.P. and S.A. Reed, (1990). ''Operating Cash Flow Formats: Does Format Influence Decisions?'' *Journal of Accounting and Public Policy*, *9*: 217–235.

Maxwell, S.E. and H.D. Delaney (1990). *Designing Experiments and Analyzing Data: A Model Comparison Approach*, Belmont: Wadsworth Publishing Company.

Montgomery, D.C. (1991). *Design and Analysis of Experiments*, Third ed., New York: John Wiley & Sons, Inc.

Newell, C. and D.N. Ammons (1987). Role Emphasis of City Managers and Other Municipal Executives, *Public Administration Review*, *47*(3): 246–253.

Nunn, S. (1994). ''How Capital Technologies affect Municipal Service Outcomes: The Case of Police Mobile Digital Terminals and Stolen Vehicle Recoveries,'' *Journal of Policy Analysis and Management*, *13*(3): 539–559.

Pedhazur, E.J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*, Second ed., New York: Holt, Rinehart and Winston.

Siegel, S. and J.N. John Castellan (1988). *Nonparametric Statistics for the Behavioral Sciences*, Second ed., New York: McGraw-Hill Book Company.

Toothaker, L.E. (1991). *Multiple Comparisons for Researchers*, Newbury Park: Sage Publications.

Velleman, P.F. and D.C. Hoaglin (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*, Boston: Duxbury Press.

Warner, J.L., J.J. Berman, J.M. Weyant, and J.A. Ciarlo (1983). ''Assessing Mental Health Program Effectiveness: A Comparison of Three Client Follow-up Methods,'' *Evaluation Review*, *7*(5): 635–658.

Wells, J.B., B.H. Layne, and D. Allen (1991). ''Management Development Training and Learning Styles,'' *Public Productivity and Management Review*, *XIV*(4): 415–428.

Wittmer, D. (1991). ''Serving People or Serving for Pay: Reward Preferences Among Government, Hybrid Sector, and Business Managers,'' *Public Productivity and Management Review*, *4*(4): 369–383.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

13
Linear Correlation and Regression

Leslie R. Alm
Boise State University, Boise, Idaho

I. INTRODUCTION

The primary emphasis of social science research, including public administration, is to evaluate relationships between variables in search of causation. As researchers, we want to know how variables are related to each other; that is, we want to know which variable is influencing (causing) the other. For instance, public administrators may want to know if gubernatorial leadership is causally linked to the effectiveness of employment training programs in states, or if the percentage of African Americans affects the amount of money spent by cities on minority set-aside programs, or if public education strategies influence participation in recycling programs. To be sure, establishing that one variable is having a causal effect on another variable requires meeting a rigorous set of standards which includes showing that there is a strong theoretical reason for believing that one variable is the cause of the other, that the relationship between the two variables is not the result of another variable that is related to both variables, that one of the variables precedes the other in time, and that the two variables covary (move or change in relation to each other) (Welch and Comer, 1988).

Linear correlation and regression analysis are two widely accepted statistical techniques designed to help the researcher establish these criteria for causal linkages by providing an estimation of exactly how variables are related to each other. Regression analysis provides an equation which describes the exact relationship between two interval level variables and is used to predict the value of one variable based on the value of the other. Correlation analysis produces a measure of association that not only indicates the strength and direction of the relationship, but also provides a measure of how accurate the regression equation is in predicting the relationship.

It is important to note that while these two statistical techniques are most powerful when used in multivariate analysis (analysis involving more than two variables), the purpose of this chapter is to illustrate and explain correlation and regression as they apply to the linear relationship between *two* variables at the interval level of measurement. It is common in the social sciences to describe correlation and regression between two variables as bivariate correlation and simple (or bivariate) regression.

Bivariate correlation and simple regression provide the foundation for multivariate regression (the subject of the following chapter) in that the logic and principles that underlie their understanding are identical to the logic and principles that underlie the more complex multivariate techniques. The path we take to understanding linear correlation and regression begins with the idea that variables are related to each other in some linear fashion. In fact, the definition, calculation, and interpretation of both correlation and regression coefficients are directly tied

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

to the concept of linearity. If one is to understand linear regression and correlation, one must first come to terms with linearity.

II. VARIABLES, RELATIONSHIPS, AND LINEARITY

The concept of linearity is framed within a discussion of variables and how these variables are related to each other. Typically, the researcher wants to know if one variable is influencing another variable. For example, a researcher might want to know if people with higher levels of education vote more often. The variables of concern would be education and voting. As researchers, we could illustrate this relationship in terms of independent and dependent variables where the independent variable is the one that provides the influence and the dependent variable is the one that receives the influence. This general relationship would be portrayed by the symbols

$$\mathbf{X} \rightarrow \mathbf{Y}$$

where **X** represents the independent variable and **Y** represents the dependent variable. The arrow points from **X** to **Y** indicating that **X** is influencing **Y**. It should be recognized that the direction of the arrow is chosen by the researcher based on theoretical considerations, good judgment, and past research (Lewis-Beck, 1980). For our example involving education and voting, the relationship would be illustrated as follows:

$$\text{EDUCATION} \rightarrow \text{VOTING}$$

where education would be the independent variable and voting would be the dependent variable.

Linear correlation and regression are tools that are used with interval levels of measurement and are based on the assumption of linearity.[1] The level of measurement requirement means that in order to effectively use linear correlation and regression, our variables must be measured such that they permit either comparisons of quantitative differences among cases on a scale (e.g., time: 1950, 1990) or in absolute distances between cases (e.g., money: $10, $20) (Hoover and Donovan, 1995). The assumption of linearity (that our relationship follows the path of a straight line) is justified on the grounds that it is generally considered our most parsimonious alternative and it provides a starting point when our theory is weak and inspection of the data themselves fails to provide a clear alternative. Furthermore, in the ''real'' world, numerous relationships have been found to be empirically linear (Lewis-Beck, 1980).

III. THE RESEARCH SETTING

Research typically begins with a research hypothesis of the form ''as one variable (X) increases (decreases) in value, the other variable (Y) increases (decreases) in value.'' Researchers want to know two things about the relationship of these two variables—the direction and strength. Direction is either positive (represented by a ''+'' sign and meaning that the variables change in the same direction; e.g., as one increases in value, the other increases in value) or negative (represented by a ''−'' sign and meaning that the variables change in the opposite direction; e.g., as one increases in value, the other decreases in value). Strength shows how closely the variables covary (change together). For linear correlation and simple regression, strength is determined by how close the relationship comes to being a straight line.

A good way to view relationships between two variables (and check for linearity) is through the use of scatter diagrams. Scatter diagrams show the distribution of scatter points (ordered pairs of values associated with each of the variables) along an X-Y continuum in such

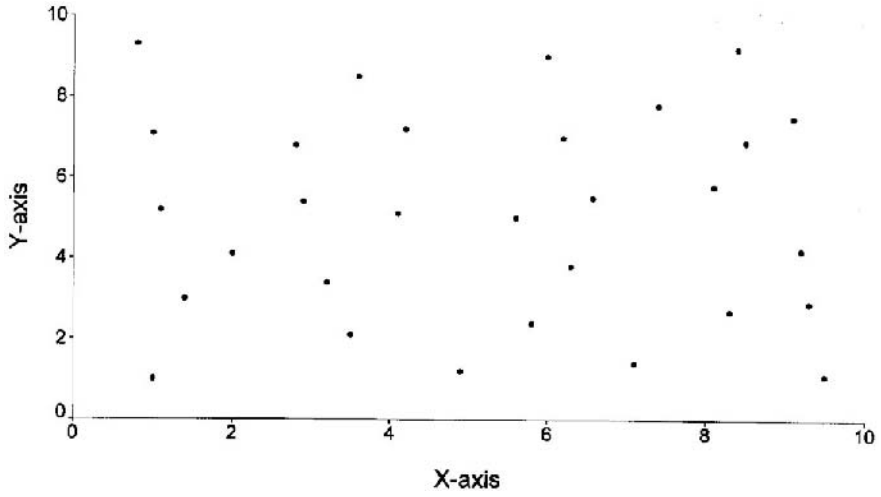


FIGURE 1 Scatter diagram indicating no linear relationship.

a way that one can visualize how variables covary. Furthermore, scatter diagrams provide an excellent means of conceptualizing the ideas of strength and direction.

To get a general feel of scatter diagrams and how they can help us grasp the concepts of strength and direction, look at Figure 1–4. There does not appear to be a discernable relationship in Figure 1, whereas Figure 2 appears to be curvilinear (nonlinear) in nature. For now, we will not concern ourselves with nonlinear relationships, but will focus on Figures 3 and 4, which appear to be linear in nature even though they exhibit different characteristics. For instance, the scatter points in Figure 4 appear to be more closely grouped together than the scatter points in Figure 3. In fact, this grouping of variables signifies the strength of the relationship—the closer the scatter points come to being grouped (about an imaginary line), the stronger the association between the variables. In this case, the strongest association would be represented by Figure 4 and the weakest by Figure 3.

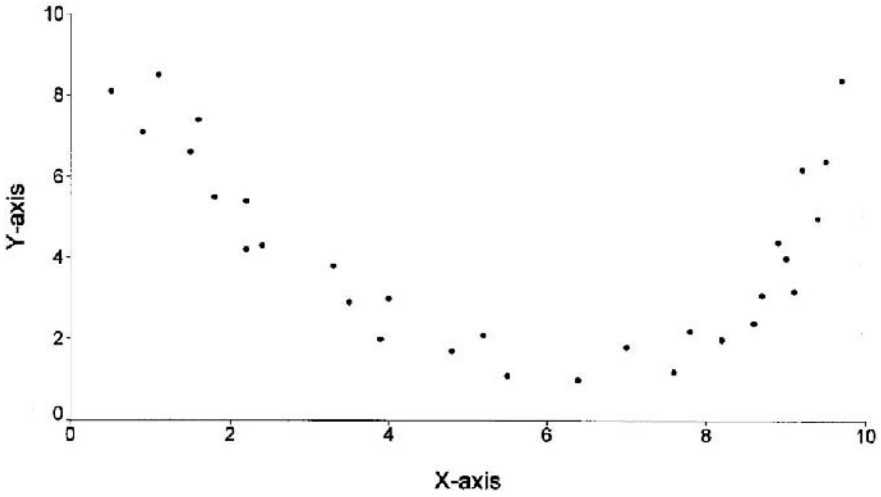


FIGURE 2 Scatter diagram indicating nonlinear relationship.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

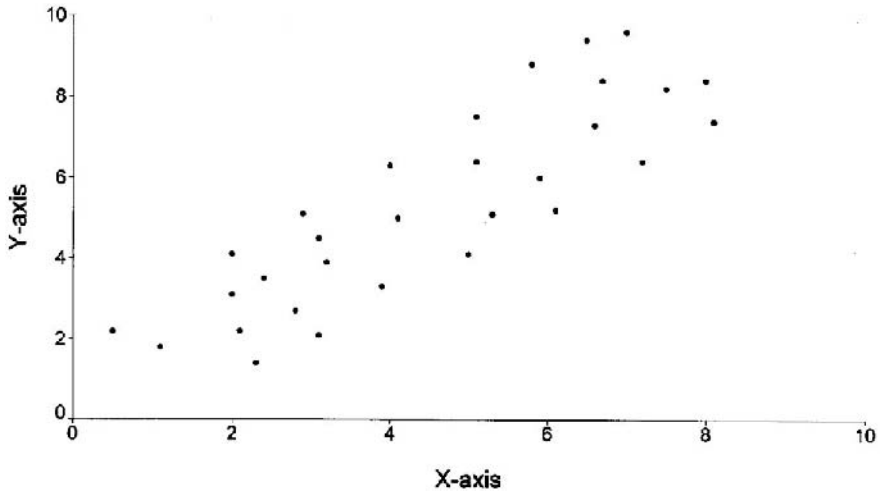


FIGURE 3 Scatter diagram indicating weak, positive linear relationship.

Something else that is noticeable about the scatter plots in Figures 3 and 4 is that each of the plots has a different slant; that is, the scatter points in Figure 4 appear to be much flatter than those in Figure 3. In addition, the scatter points in Figure 3 appear to move upward as one moves right along the X-axis (the horizontal axis), while the scatter points in Figure 4 appear to move downward. In fact, those characteristics signify both the direction and the nature of the relationship. We would say that the scatter points in Figure 3 would be positive in nature (as X increases, Y increases) and that the scatter points in Figure 4 would be negative in nature (as X increases, Y decreases). Furthermore, we could also make a distinction between the relationships shown in Figure 3 and 4 in that the scatter points in Figure 3 change at a much higher rate than do the scatter points in Figure 4; that is, for Figure 3, as we move along the X-axis, Y changes at a much greater rate than it does in Figure 4. This change in Y against the change in X is known as the slope of the line and is formally defined as the change in Y for a one-unit change in X.

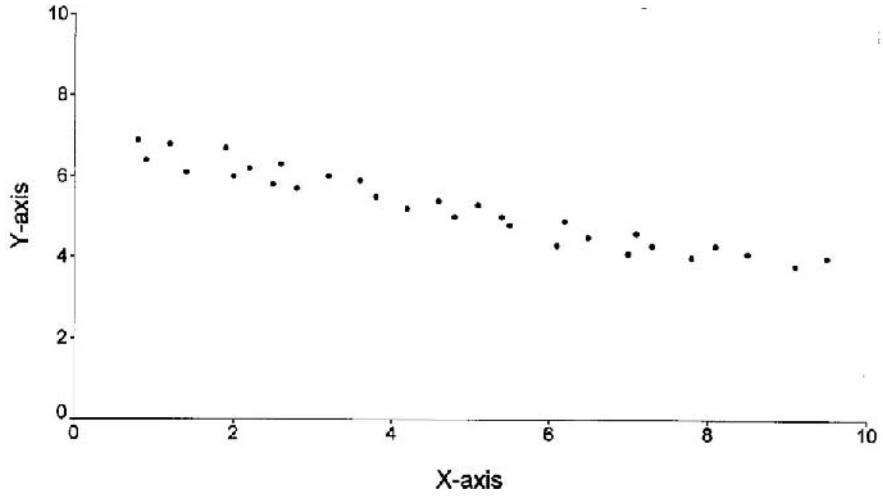


FIGURE 4 Scatter diagram indicating strong, negative linear relationship.

One thing that is important to note right away is that the measure of direction and strength are distinct in nature. One measure has to do with whether the variables change in a direct or inverse manner (direction), while the other measure has to do with how closely the scatter points come to forming a line (strength). Moreover, we have two distinct ways of approaching an explanation and analysis of these relationships. To determine the direction and nature of the relationship we use regression analysis and to determine how strong the relationship is we use correlation analysis.

IV. LINEAR REGRESSION ANALYSIS

As mentioned earlier, as researchers we are often interested in the relationship of one variable to another, usually viewed in the format of $\mathbf{X} \rightarrow \mathbf{Y}$. However, to determine the exact linear nature of the relationship, we express the relationship in the format of a simple regression equation,

$$\mathbf{Y = a + bX},$$

where **Y** represents the dependent variable, **a** represents the Y-intercept of the line (the point where the line crosses the Y-axis), **b** represents the slope of the line, and **X** represents the independent variable.[2]

The idea of linear regression analysis is to fit our research data (as represented by a scatter diagram) to a straight line and then use the equation of that line to predict the nature of the relationship between the two variables. A real-life example may prove helpful in illustrating and explaining the usefulness of this statistical technique.

Much research has been conducted recently involving the unique aspects of the American West (Farley, 1995), especially as it involves a deep historical conflict among competing values that has resulted in a 'new environmental West' where a new environmental movement is challenging, and changing, the values established by an older, natural resource based West (Hays, 1991). In particular, researchers have been investigating the relationship between the amount of federal land in a county (comparatively speaking, the American West contains a much higher percentage of federal lands than do the other regions of the United States) and people's attitudes and beliefs (Alm and Witt, 1995).

In this context, suppose a researcher chose the individual counties in Oregon ($n = 36$) as the unit of analysis for a study investigating the relationship between the amount of federal land in a county and how the people in that county voted for President in the 1992 general election. Previous research suggested that the amount of federal land in a county was linked to more conservative environmental attitudes (Alm and Witt, 1997), hence, the researcher suspected that the higher the percentage of federal land in the county, the lower the percentage vote would be for the Democratic candidate—Bill Clinton. After she gathers her data (see Table 1), the first thing the researcher did was produce a scatter diagram relating percent federal land to percent vote for Clinton. The scatter diagram is shown in Figure 5. A glance at the scatter diagram suggested to the researcher that the relationship between these two interval-level variables was indeed linear.

The researcher's next step was to use a mathematical process known as the method of least squares to estimate the slope (b) and the Y-intercept (a) for the equation of the line that best represents the scatter points as displayed in Figure 5.[3] Fortunately for today's researchers, there are many statistical packages available to calculate these values. However, if the reader is interested, an overview of the hand calculations for these values is presented in Procedure 1 of the Appendix.

For this particular example, the researcher ends up with a value for the slope (b) equal

TABLE 1 Variables for Oregon Study

County	Percent federal land	Percent vote for Clinton
Baker	51.40	31.81
Benton	17.00	47.37
Clackamas	47.80	39.03
Clatsop	.01	45.80
Columbia	2.70	42.77
Coos	21.50	40.70
Crook	49.40	34.49
Curry	59.40	34.76
Deschutes	75.90	35.73
Douglas	48.10	30.83
Gilliam	2.90	36.03
Grant	60.00	28.47
Harney	70.60	28.86
Hood River	62.50	39.61
Jackson	45.10	37.80
Jefferson	16.70	36.59
Josephine	57.90	32.80
Klamath	51.10	29.77
Lake	67.70	26.80
Lane	54.40	48.78
Lincoln	30.20	44.41
Linn	37.50	34.00
Malheur	71.50	23.81
Marion	29.40	37.28
Morrow	13.70	33.79
Multnomah	26.30	55.34
Polk	8.80	37.29
Sherman	8.20	32.44
Tillamook	19.70	43.89
Umatilla	20.00	34.55
Union	47.70	34.43
Wallowa	57.60	29.53
Wasco	16.00	42.50
Washington	2.60	40.39
Wheeler	23.20	31.05
Yamhill	14.80	35.50

Source: U.S. Department of Interior, Bureau of Land Management, Oregon State Office, Portland, Oregon, 1994; The Election Data Book, Bernan Press, Lanham, Maryland, 1992.

to −.14 (rounded off) and a value for the intercept (a) equal to 41.53 (rounded off). The equation of the line then becomes

$$\mathbf{Y = a + bX} \text{ or percent vote for Clinton} = 41.53 + [-.14 \text{ (percent federal land)}]$$

The actual depiction of this line can be viewed in Figure 6. What is important to the researcher is the interpretation of this equation; that is, within the framework of the research project, how

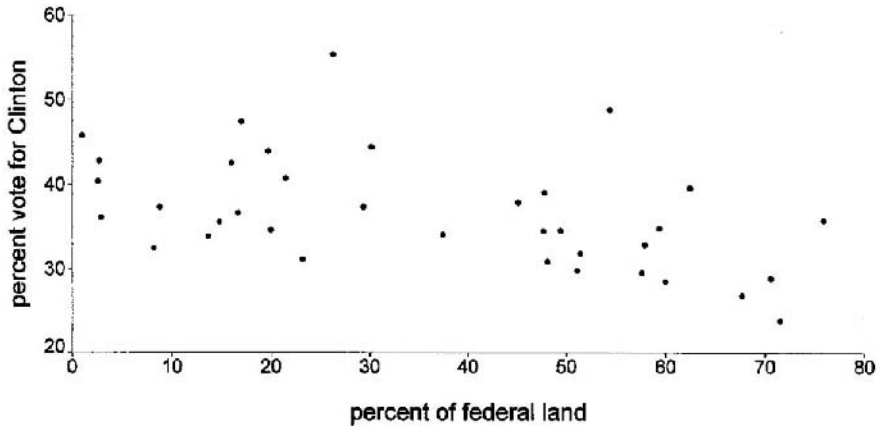


FIGURE 5 Scatter diagram of 1992 Oregon Vote For Clinton by County. From *The Election Data Book*, 1992.

do the calculated values of the slope and intercept help to define the relationship between percent vote for Clinton and percent federal land.

The more important of the two coefficients is **b**, which is known as the unstandardized regression coefficient and represents not only the slope of the line, but tells us the exact relationship (on average) between percent federal land and percent vote for Clinton in Oregon counties. The interpretation of the unstandardized regression coefficient is fairly straight forward. In general, the value for **b** indicates the average change in the dependent variable (Y) associated with a unit change in the independent variable (X). In this case, for $\mathbf{b} = -.14$, the interpretation would be: for a one percent increase in percent federal land in Oregon counties, on average, there would be a corresponding decrease of .14 percent in the vote for Clinton.

Several things are important to note here. First, it is very important that the researcher operationalizes each of the variables in a manner that is conducive to interpretation. Straight forward and concrete measures serve this purpose well. While there may be instances when the researcher has no choice in the type of units used, it is important that the researcher put a good

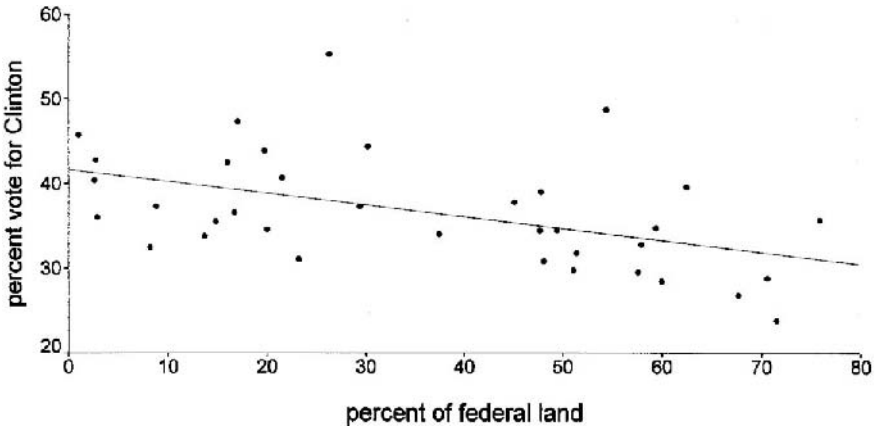


FIGURE 6 Scatter diagram of 1992 Oregon Vote For Clinton by County. From *The Election Data Book*, 1992.

amount of thought and consideration into how he/she will measure variables before the actual data is gathered. Second, the sign in front of the slope indicates the direction of the relationship. If the sign is positive (+), then the dependent variable increases as the independent variable increases. If the sign is negative (−), then the dependent variable decreases as the independent variable increases. For this example, the sign was negative, so as percent of federal land increases, the percent vote for Clinton decreases. Third, we are talking about average change; that is, on average, for a one unit change in the independent variable, the dependent variable changes so much. For this example, the researcher would not make the claim that in every instance, a one percent increase in federal land leads exactly to a decrease of .14 percent vote for Clinton. Rather, the value of the unstandardized regression coefficient represents the average decrease in percent vote for Clinton we would expect over a large number of counties.

The interpretation of the intercept (a) is also quite straight forward. It simply estimates the value of the dependent variable (Y), when the independent variable (X) equals zero. In this example, with **a** = 41.53, the interpretation would be that for a county that has zero percent federal land (excluding all other variables), 41.53 percent of the people would vote for Clinton. However, there are two major problems that can occur that make the interpretation of the intercept unusable (Lewis-Beck, 1980). First, if the range of values for the independent variable (X) does not include the intercept, making generalizations about its meaning is quite risky. For example, if the actual range of the independent variable for our example was from 60 to 90 percent federal land, our calculated value of **a** (equal to 41.53) would not be very representative of our actual values and it would be risky to put much faith in its meaning. Second, if the intercept turns out to be a negative value, the meaning would not make sense. For instance, if we had a negative value for **a** in our example, that would mean for a value of zero percent federal land in a county, percent vote for Clinton would be negative, which we know is impossible. When all is said and done, while the intercept is necessary to complete the regression equation, researchers generally ignore its interpretation.

As in calculating all statistics, the researcher must insure that the unstandardized regression coefficient (b) is statistically significant. Simply put, statistical significance tells the researcher whether he/she can have faith that the calculated value for **b** is representative of a relationship that exists in the population from which the sample was taken. In practical terms, the calculated value for the unstandardized regression coefficient is only descriptive in nature; that is, it merely describes the set of data from which it was computed. In actuality, we want to test this relationship (the value for **b**) to insure that it exists in the real world and was not derived simply by chance.[4] To complete this test, we use the procedure described in Chapter 3 for testing of null hypotheses. The actual calculations are presented in Procedure 2 of the Appendix.

For our example, the value of the unstandardized regression coefficient was statistically significant; hence, the researcher can have confidence that the value for the unstandardized regression coefficient exists in the ''real world'' and that the calculated value for **b** (−.14) is representative of what it would be in the population from which it came. As with the calculations of the coefficients, statistical significance can be easily calculated today by using any number of statistical packages designed to do exactly that.

V. LINEAR CORRELATION

While the regression coefficient provides us with the exact nature of the relationship between two variables, the correlation coefficient provides us with a measure of how strong the relationship is. As mentioned earlier, strength is a measure of linearity; that is, it measures how close the relationship comes to being a straight line. Looking again at Figures 1–4, it is clear that Figures 1

and 2 do not exhibit linearity; however, Figures 3 and 4 do. In fact, visual inspection allows us to notice that the scatter points displayed in Figure 4 appear to be bunched more closely along a linear continuum than the scatter points depicted in Figure 3. We could therefore conclude that the ''strongest'' relationship is that depicted in Figure 4.

The correlation coefficient is the statistical measure that provides an exact measure of how closely the relationship between two variables comes to being linear. The most commonly used measure of linear correlation is Pearson's Correlation Coefficient, often referred to as Pearson's **r**. As a measure of linear associations, Pearson's **r** varies from -1 to $+1$ in value, with a value of zero meaning that no linear association exists between the variables (Norusis, 1991). The sign provides the direction of the association, while the actual value provides the indication of strength. As the value of Pearson' **r** moves from 0 to $+1$, it indicates a stronger positive association, with a $+1$ indicating a perfect positive linear association. As the value of Pearson's **r** moves from 0 to -1, it indicates a stronger negative association, with a -1 indicating a perfect negative association. Generally, in the real world we do not find perfect associations between variables, hence the values for Pearson's **r** will fall somewhere between ± 1.

Returning to our earlier example depicting the relationship between percent federal land and percent vote for Clinton in Oregon counties, we use the same mathematical principles established to calculate the unstandardized regression coefficient (b) to calculate Pearson's **r**. The actual calculation of Pearson's **r** is provided in Procedure 1 of the Appendix. Most often, however, these calculations are now left to computers and advanced statistical programs readily available to today's researchers.

For our example, Pearson's $\mathbf{r} = -.4700$ and would be interpreted as follows: there is a moderately strong, negative, linear association between the percent federal land and percent vote for Clinton (for the 1992 election). As we did for the regression coefficient, we must test for statistical significance of Pearson's **r**. We find that the calculated value for Pearson's **r** is statistically significant (see Procedure 3 in the Appendix), allowing us to reject the null hypothesis that Pearson's **r** is equal to zero and accept the fact that for our data, the value of the correlation coefficient equals $-.4700$.

There are several important points that must be made regarding the coefficients discussed above. First, in bivariate regression and correlation analysis, the direction established by the calculations for Pearson's **r** and the regression coefficient will necessarily be the same. While this may seem an obvious observation, as you continue your study of multivariate techniques (which control for many factors) the direction of the initial bivariate relationships may change. Second, Pearson's **r** and the regression coefficient (b) are two distinct measures of association and are calculated in different ways (refer to Procedure 1 of the Appendix). Pearson's **r** is a measure of association and does not distinguish which of the variables is affecting the other. It only shows how strong the association is between the variables, disregarding which direction the arrow of influence is pointing. On the other hand, in calculating the regression coefficient, the researcher must specify the causal direction; that is, the researcher must choose which variable is the independent and which is the dependent variable. As specified much earlier in the chapter, this choice is made by the researcher based on her reading of the relationship. It does not come from the application of statistical techniques; it comes from the imagination and thought of the researcher.

Third, an interesting and valuable extension of Pearson's **r** provides the researcher with another very straight forward interpretation of the explanatory power of the regression equation. R^2 (commonly called the coefficient of determination)[5] records the proportion of the variation in the dependent variable ''explained'' or ''accounted for'' by the independent variable (Lewis-Beck, 1980). Recall that we began our discussion of bivariate relationships by saying that the researcher was investigating the effect of one variable on another. Essentially one of the things

the researcher wants to know is how much of the change (variance) in the dependent variable is ''caused'' by the change in the independent variable. The coefficient of determination provides us with the answer. Since **r** varies between ± 1, R^2 will also vary between ± 1. However, the interpretation of R^2 becomes one of percent—in practical terms, R^2 gives us the percent of variation in the dependent variable explained by the independent variable. If $R^2 = 0$, then the independent variable explains none of the variance in the dependent variable. If $R^2 = 1$, then the independent variable explains all of the variance in the dependent variable. Of course, in the real world, we will seldom get an R^2 equal to any of these perfect values. Instead we get values between 0 and 1.

From our example, Pearson's $\mathbf{r} = -.4700$. Squaring this gives us an $R^2 = .22$ (rounded). The interpretation would then be that the percent federal land in Oregon counties (the independent variable) explains 22 percent of the variance (change) in the percent vote for Clinton (the dependent variable). In the end, essentially what R^2 is telling the researcher is how closely the relationship comes to being a linear relationship (commonly referred to as ''the goodness of fit'' of the regression equation). In practical terms, the coefficient of determination (R^2) indicates to the researcher how much the change in the dependent variable is due to the change in the independent variable.

VI. A RESEARCH EXAMPLE WITH COMPUTER OUTPUT

Let's take a closer look at the research example presented above using an analysis of the computer output to estimate the bivariate relationship between two variables in the context of linear correlation and simple regression. The hypothesis we began our investigation with was: as the percent of federal land in Oregon counties increases, the vote for President Clinton in 1992 decreases. Using SPSS for Windows (SPSS Inc.), the computer output for correlation and simple regression were derived as presented in Tables 2 and 3.[6]

The researcher would first turn to the correlation table (see Table 2) and observe that there is a moderately strong negative linear association between percentage of federal land and the vote for President Clinton in Oregon counties in 1992. The researcher comes to this conclusion because Pearson's correlation coefficient equals $-.4796$. The negative sign indicates that as the percent of federal land in each county increases, the percent vote for Clinton decreases. The value of $-.4796$ places this correlation coefficient about midway between 0 and -1 (see Figure 7) on the standardized continuum indicating strength of a measure (remember, 0 indicates no

TABLE 2 Pearson's Correlation Coefficient

	Percent vote for Clinton	Percent federal land
Percent vote for Clinton	1.0000	−.4796
	(36)	(36)
	P = .	P = .003
Percent federal land	−.4796	1.0000
	(36)	(36)
	P = .003	P = .

Sources: SPSS For Windows 6.0, 1993; U.S. Department of Interior, Bureau of Land Management, Oregon State Office, Portland, Oregon, 1994; *The Election Data Book*, Bernan Press, Lanham, Maryland, 1992.

TABLE 3 Bivariate Regression of Percent Vote For Clinton (Dependent Variable) with Percent Federal Land (Independent Variable)

Multiple R	.47956
R Square	.22998
Adjusted R square	.20733
Standard error	5.96402
F	10.15480
F significance	.00310

Variables in the Equation

Variable	b	Se b	Beta	t	Sig t
Percent federal land	−.139386	.043740	−.479564	−3.187	.0031
(Constant)	41.63130	1.855276		22.439	.0000

Sources: SPSS For Windows 6.0, 1993; U.S. Department of Interior, Bureau of Land Management, Oregon State Office, Portland, Oregon, 1994; *The Election Data Book*, Bernan Press, Lanham, Maryland, 1992.

association and −1 indicates a perfect negative association—the strongest negative association a researcher could obtain).

Furthermore, this association is statistically significant. The researcher comes to this conclusion because the value for significance (p) equals .003, which is less than .05—the most commonly used level of statistical significance in public administration. Of course, this value (.003) would also be considered statistically significant if the researcher had chosen a significance level of .01 or .10, the other two commonly used significance levels.

That the value for our correlation coefficient is statistically significant is an extremely important concept. It means that we can be confident that the value we obtained from our sample is representative of the value in the population. In other words, since our obtained significance value is so small (less than .05, our chosen level of significance), we are reasonably certain that the association we observed is not just due to chance, but exists in the ''real world.'' What is also important to remember is that if our value for the correlation coefficient did not reach statistical significance, we could not be confident that the value we obtained was not simply due to chance and hence, we could not be confident that the association truly exists in our population of study. If that were the case, it would be very risky for the researcher to make claims about the strength and direction of the relationship. Simply put, good researchers would not use correlation coefficients that do not reach statistical significance.

In our example, we did find a statistically significant association between our variable of interest, allowing us to be confident that there exists a moderately strong negative association between percent federal land and vote for Clinton in Oregon counties. The researcher then would turn to analyses of the regression equation. From Table 3, the researcher first notes the Adjusted

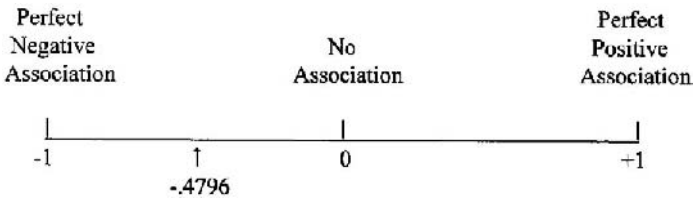


FIGURE 7 Correlation continuum for percent vote for Clinton and percent federal land.

R square (R^2) of .20733. In this case, the researcher chooses Adjusted R^2 (instead of R^2) because it provides a better measure of ''goodness of fit''—Adjusted R^2 adjusts the value of R^2 to take into account the fact that a regression model always fits the particular data on which it was developed better than it fits the population data (Norusis, 1991). In essence, Adjusted R^2 provides the researcher with a more accurate measurement of explained variance.[7]

The interpretation of Adjusted R^2 is fairly straight forward. In this instance, an Adjusted $R^2 = .20733$ means that approximately 21 percent of the variation in our dependent variable (percent vote for Clinton) can be explained by the independent variable (percent federal land). This is important because it not only tells us how much of the variance in vote for President Clinton can be explained by amount of federal land that exists in a county (percent federal land), but it also tells us that almost 70 percent of the variance is left unexplained, meaning that there are other factors besides percent federal land that are affecting the vote for Clinton. For the researcher, this means other variables must be added to the mix for a fuller explanation. This would be accomplished through a multiple regression equation (the topic of the next chapter).

Still, what the researcher is really interested in is the regression coefficient (b). As noted earlier, the first thing the researcher would do is check the significance of the regression coefficient (Sig t). In this case Sig t = .0031, which is less than the researcher's chosen level of significance of .05; hence, the value for **b** (−.139386) is statistically significant and the researcher can feel confident that this value exists in the ''real world'' and is representative of the studied population. The interpretation of the regression coefficient is also straight forward. For **b** = −.139386, it means that for a one percent increase in federal land in a county, the percent vote for Clinton decreases on average about .14 percent.

In the final analysis, using linear correlation and simple regression, the researcher would be quite confident that a relationship exists in Oregon counties between the percent federal land in the county and the vote for Clinton in 1992. It appears that the association is negative (signs of both the correlation coefficient and the regression coefficient were negative), moderately strong (r = −.4796), and on average, for a one percent increase in federal land, the percent vote for Clinton decreases by .14 percent.

Note that the simple regression equation would be the same (other than rounding error) as the one developed earlier from the hand calculations:

percent vote for Clinton = 41.63 − .14 (percent federal land),

with the value of the intercept (a = 41.63) obtained from Table 3 (the value of the ''constant'') and being equal to the predicted percent vote for Clinton when percent federal land equals zero.

F Significance and Beta (the standardized regression coefficient), while unimportant in bivariate analysis become extremely important in multivariate analysis. F significance represents the statistical significance of the entire model (all the independent variables taken together) and Beta is used to select the variable with the strongest impact on the dependent variable, controlling for all the independent variables in the model. In both instances, since bivariate regression only involves one independent variable, the interpretation of these two statistics becomes a moot point. The standard error of **b** (Se b) and the **t** value are measures used to determine the statistical significance of **b** as illustrated in Procedure 2 of the Appendix.

VII. ASSUMPTIONS AND RESIDUALS

In order for the researcher to accurately infer that the measures estimated through bivariate regression are representative of the population values, certain assumptions must be met (Lewis-Beck, 1980; Norusis, 1991). Among these are that the relationship is linear, the dependent variable is normally distributed for each value of the independent variable, the variance in the dependent variable is constant for all values of the independent variable, and that all observations

(cases) in the study are selected independently. It is important to note that these assumptions apply to multivariate regression analysis as well as bivariate regression analysis.

The way that researchers check whether these assumptions are being met is through analysis of the error terms, commonly referred to as residuals. Residuals are simply the difference between the observed and predicted values of the dependent variable. For instance, our prediction model for the relationship between percent vote for Clinton and percent federal land is

$$\text{percent vote for Clinton} = 41.63 - .14\,(\text{percent federal land}).$$

The predicted value of percent vote for Clinton for a county with 60 percent federal land would be

$$\begin{aligned}\text{percent vote for Clinton} &= 41.63 - .14(60)\\ &= 41.63 - 8.4\\ &= 33.23.\end{aligned}$$

The actual observed value of percent vote for Clinton (from our data set—see Table 1) for a county with 60 percent federal land is 28.47. Hence, the observed value minus the predicted value would be $28.47 - 33.23 = -4.76$. The value of -4.76 would be called the residual. By analyzing the scatter plots of the residuals, the researcher can check to see if each of the bivariate regression assumptions are being met.

The easiest way to check for linearity is to inspect the scatter diagram of the dependent variable plotted against the independent variable. In fact, we started the investigation of the relationship between percent vote for Clinton and percent federal land by doing exactly that. The results of that scatter plot can be viewed in Figure 5. Initial inspection of this scatter plot indicated that a linear regression model would be appropriate, as the points seem to cluster around a negative sloping straight line.

A second way to check for linearity is to inspect the plots of the residuals against the predicted values. For our example, these plots are represented in Figure 8. If a non-linear relationship existed, one would expect to see some type of non-linear pattern among the residuals. Since the scatter diagram of our residuals appears to show a random pattern of plots along a horizontal continuum, we can be reasonably certain that our relationship is linear.

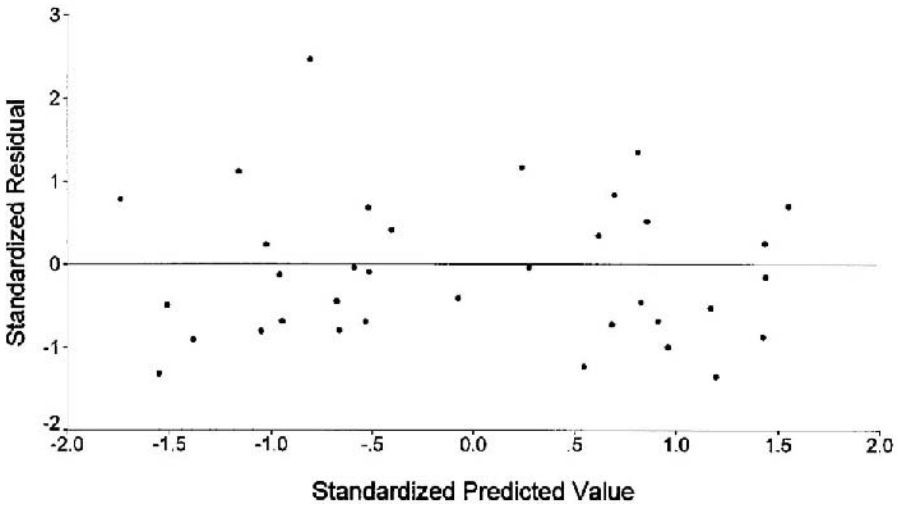


FIGURE 8 Scatterplot of residuals for simple regression.

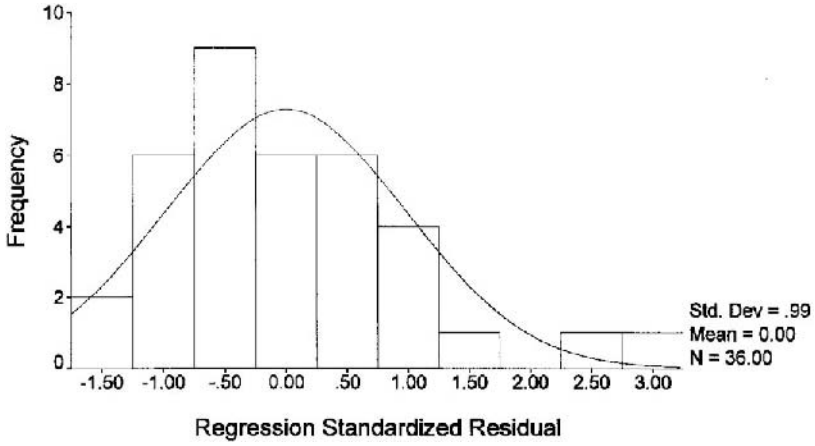


FIGURE 9 Histogram of dependent variable.

To check the normality assumption, the researcher would use the computer to plot a histogram depicting the distribution of residuals. (A histogram is a pictorial display of frequencies commonly used to interpret frequency distributions.) The histogram of standardized residuals for our example is presented in Figure 9. If the dependent variable is normally distributed for each value of the independent variable, then our distribution of residuals should approach normality. Inspection of our histogram shows that for our regression equation, we could be fairly confident that we meet the normality assumption.

The researcher should also check to insure that the variance in the dependent variable is constant for all values of the independent variable. In statistical terms, this is known as homoskedesticity. To insure that we have homoskedasticity for our example, we would inspect the same residual plots that we used to check for linearity (Figure 8). If homoskedesticity was present we would expect to find a balanced scatter plot with an equal distribution of points above and below the zero line. If heteroskedasticity (non-constant variance) existed, we would expect to find a pattern of increasing or decreasing values (Figure 10) for the residuals across

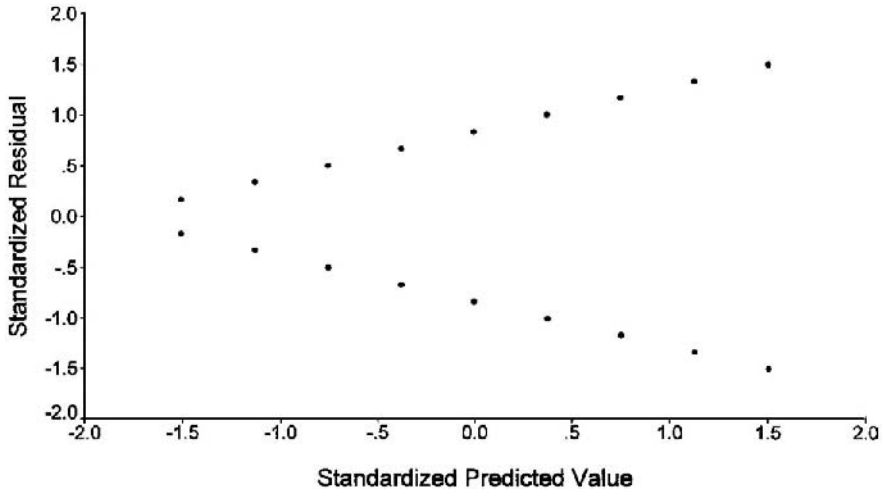


FIGURE 10 Scatterplot of residuals show heteroskedasticity.

Case #	-3.0 0.0 3.0	CLINTON	*PRED	*RESID
147	. *. .	31.81	34.4669	-2.6608
148	. . * .	47.37	39.2617	8.1132
149	. . * .	39.03	34.9687	4.0573
150	. . * .	45.80	41.6302	4.1704
151	. .* .	42.77	41.2550	1.5138
152	. .* .	40.70	38.6345	2.0681
153	. * .	34.49	34.7456	-.2524
154	. .* .	34.76	33.3518	1.4084
155	. . * .	35.73	31.0519	4.6806
156	. *. .	30.83	34.9268	-4.1004
157	. * . .	28.47	33.2681	-4.7935
159	. *. .	28.86	31.7907	-2.9354
160	. . * .	39.61	32.9197	6.6926
161	. .* .	37.80	35.3450	2.4598
162	. *. .	36.59	39.3036	-2.7137
163	. * .	32.80	33.5609	-.7619
164	. *. .	29.77	34.5087	-4.7339
165	. * . .	26.80	32.1949	-5.3932
166	. . * .	48.78	34.0487	14.7344
167	. . * .	44.41	37.4218	6.9892
168	. *. .	34.00	36.4043	-2.4012
169	. * . .	23.81	31.6652	-7.8539
170	. * .	37.28	37.5334	-.2493
171	. * . .	33.79	39.7217	-5.9278
172	. . *.	55.34	37.9655	17.3768
173	. *. .	37.29	40.4047	-3.1165
174	. * . .	32.44	40.4883	-8.0511
175	. . * .	43.89	38.8854	5.0017
176	. *. .	34.55	38.8436	-4.2918
177	. * .	34.43	34.9826	-.5504
178	. *. .	29.53	33.6027	-4.0740
179	. . * .	42.50	39.4011	3.0941
180	. * .	40.39	41.2689	-.8801
181	. * . .	31.05	38.3976	-7.3510
182	. *. .	35.50	39.5684	-4.0720
Case #O	-3.0 0.0 3.0	CLINTON	*PRED	*RESID

FIGURE 11 Casewise plot of standardized residual for simple regression.

the horizontal axis (predicted values). For our example (Figure 8), no such pattern exists, hence we could be confident that we had homoskedasticity.

The final assumption that researchers need to check involves autocorrelation; that is, that all observations (cases) selected for our study are independent from each other. One way to check for independence is to complete a case plot of the residuals in sequence. For our example, the case plot of residuals is presented in Figure 11. If autocorrelation were present, we would find some kind of a pattern in the residuals instead of a random appearing sequence. For instance, one pattern might be that as the sequence of cases increased, the residuals become increasingly more positive (or negative). An example of a case plot that would indicate the presence of autocorrelation is displayed in Figure 12. Inspection of our case plots in Figure 11 (where the sequence is county in alphabetical order) shows no pattern and hence, we could be reasonably certain that autocorrelation is not present.

Case #	-3.0 0.0 3.0 O:.....:.....:O	Y6	*PRED	*RESID
1	. .* .	1.5	1.0292	.4708
2	. .* .	2.5	1.6010	.8990
3	. . * .	3.5	2.1727	1.3273
4	. . * .	4.5	2.7444	1.7556
5	. . * .	5.5	3.3161	2.1839
6	. . * .	6.0	3.8879	2.1121
7	. *. .	3.5	4.4596	-.9596
8	. * . .	3.0	5.0313	-2.0313
9	. * . .	2.5	5.6030	-3.1030
10	. * . .	2.0	6.1748	-4.1748
11	. * . .	1.5	6.7465	-5.2465
12	. * .	7.0	7.3182	-.3182
13	. * .	8.0	7.8899	.1101
14	. .* .	9.0	8.4616	.5384
15	. .* .	10.0	9.0334	.9666
16	. . * .	11.0	9.6051	1.3949
17	. . * .	12.0	10.1768	1.8232
18	. . * .	13.0	10.7485	2.2515
Case #	-3.0 0.0 3.0 O:.....:.....:O	Y6	*PRED	*RESID

FIGURE 12 Casewise plot of standardized residual for simple regression.

Two other points should be made about autocorrelation. First, it is more apt to be a concern when completing trend (or time series) analysis. Often, when observing cases over time (especially when the observations are being made of the same unit at different times), there is good reason to anticipate a problem with autocorrelation. When using cross-sectional data (over one particular point in time), the problems with autocorrelation are frequently minimal. Second, the advent of computer analyses allows for using more sophisticated ways to check for autocorrelation. The Durbin-Watson statistic is one such measure commonly used to test for autocorrelation (Wonnacott and Wonnacott, 1984). The general rule is that if autocorrelation is present, the value for the Durbin-Watson statistic will be close to 0 or 4 (Welch and Comer, 1988). For our core example, the Durbin-Watson statistic equals 2.03, so we can feel confident that autocorrelation is not a serious problem with our data.[8]

For the researcher, the question surely arises of what to do if these assumptions are not met. There are no easy answers to this question. Because of the complexity that may exist in your research, it may be quite likely that you may have to consult with someone who has special expertise in statistical methods. However, there do exist guidelines for the researcher to follow.

If the relationship between the variables is not linear, then you cannot force linearity into your model specification. Nevertheless, there is the possibility that you can change your model from its nonlinear nature into a linear distribution by transforming your original equation to one that approximates linearity. These transformations can be completed on either the independent or dependent variables. Common transformations include the log-linear, reciprocal, and root transformations (Norusis, 1991). Such transformations are relatively easy to make with the use of a computer, but it must be remembered that this transformation process changes the interpretation of your correlation and regression coefficients. In reality, the researcher should go where the data and theory lead and not attempt to fashion a linear relationship from one that is clearly not linear.

The calculus and the central limit theorem prescribe that if the sample size is large enough (in general, sample sizes are considered small if they are less than 30), then the distributions that are required to estimate our coefficients approach normality, regardless of the actual distribu-

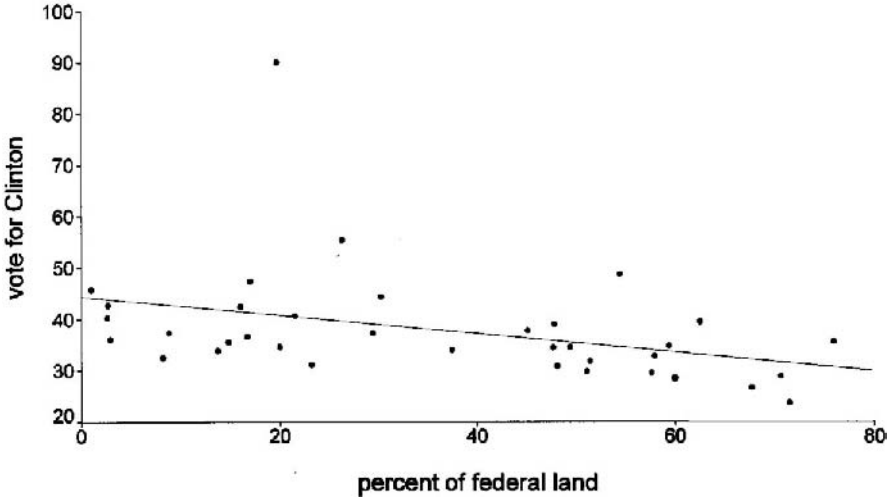


FIGURE 13 Scatter diagram indicating existence of an outlier.

tion in the population (Norusis, 1991). Essentially, what this means to the researcher is that as long as the researcher has an adequate sample size, the assumption of normality becomes a moot point.

If the homoskedasticity assumption is violated, the recommended solution is a weighted least squares procedure (Lewis-Beck, 1980). Again, the advent of the computer allows for a fairly straight forward use of this procedure. However, the interpretation of the results becomes much more complex and the researcher truly should consult with someone familiar with these more sophisticated techniques.

As mentioned earlier, autocorrelation appears more frequently with time-series analysis than with cross-sectional analysis. If the researcher finds autocorrelation or is using time series analysis, the researcher should again consult with someone familiar with the use of these more sophisticated techniques.[9]

One special problem that the researcher may have to deal with is the existence of an outlier(s). An outlier is a case (observation) with an extremely large residual (usually greater than 3 standard deviations from the mean). For instance, Figure 13 is identical to our research example scatter plot except that one county was changed (for the purpose of this discussion) from its actual percent vote for Clinton (44%) to a ''made-up'' percent (90%). If the researcher would encounter a scatter diagram such as this, it would be apparent that one of the cases has an extremely large residual which does not ''fit'' the linear model as well as the other observations. When a researcher encounters such an outlier, the very first thing the researcher should do is inspect the original data to insure that the outlier is not the result of a coding error. It is not uncommon for coding errors to occur when creating a data file. Once the researcher has confirmed that this is indeed a legitimate observation, then the researcher must decide upon a course of action (Lewis-Beck, 1980).

One possibility is to leave the outlier in the equation (if it does not seriously weaken the explanation—surely a judgement call on the part of the researcher) and provide a detailed explanation of why this case varies substantially from the others. On the other hand, the researcher could exclude the outlier and treat the regression model as if this particular observation was never made. This, however, is a risky option for any researcher unless the researcher has strong theoretical and substantive reasons for removing the observation. It is never a good idea to manipulate data to fit your statistical model.

Outliers = 2. *: Selected M: Missing

Case #	-5. O:.....:	-2. 2. :.....:	5. :O	CLINTON	*PRED	*RESID
166	.	.. *	.	48.78	34.0487	14.7344
172	.	.. *	.	55.34	37.9655	17.3768

2 Outliers found.

FIGURE 14 Casewise plot of standardized residual for outliers.

A better approach would be to report two models—one with the outlier and one without it, providing a detailed explanation of why this is appropriate (and necessary). The researcher should also keep in mind that if a large number of outliers exist, the relationship may not be truly linear. In this case, a re-evaluation of the model may be appropriate to determine if a transformation is possible or if the model fits better into a nonlinear format.

A final option is for the researcher to gather more observations. This could lead to a better fit of the data and actually change the linear equation in such a way that eliminates the outlier. On the other hand, in most instances it is not practical for a researcher to go back and add to her sample. In the end, the researcher must adjust for outliers based not on the statistical consequences, but on the theoretical and substantive framework of the research question.

It should be noted that the researcher does not have to rely on a visual inspection of the scatter plot to determine if outliers exist. Most computer statistical packages allow the researcher to simply have the computer calculate and print which cases are outliers. For our original example, no outliers exist outside of 3 standard deviations (the most commonly used standard deviation default). But if we were to look for outliers using a 2 standard deviation limit, we would find two outliers. Referring to Figure 14, the reader observes that the computer output indicates exactly which cases (by number) are outliers.

VIII. THE CORRELATION MATRIX

The most common use of linear correlation and simple regression in public administration research is to establish the existence of relationships between the dependent variable and the independent variable at the beginning of a research project. While this inevitably requires multivariate analysis, the bivariate nature of linear correlation and simple regression provide the initial clue to which variables are important. Let's look at an example of how this process would work.

Normally, a researcher decides on a dependent variable and selects several independent variables that are suspected of having an important influence. For each independent variable the researcher would follow the process delineated in the earlier sections. She would plot the dependent variable against each of the independent variables, compute simple regression and correlation coefficients, and use the residuals to check the assumptions. The next step in moving toward a multivariate analysis would be for the researcher to create a correlation matrix depicting the correlation coefficients of each of the independent variables with the dependent variable and with each other.

For an example illustrating how correlation analysis might work, let's turn our attention to a researcher who was interested in why members of Congress from certain states were active in the acid rain policy debate at the national level while others were not (see Alm, 1993). The researcher surmised that members of Congress from states with a stronger commitment to

environmental protection, greater emissions of sulfur dioxide and nitrogen oxide (the leading causes of acid rain pollution), and higher levels of precipitation acidity (acid rain) would all be more active in developing a national policy on acid rain pollution. For this particular study, the unit of analysis was the 50 American states and the dependent variable (ACTIVITY) was operationalized as the number of times a state representative introduced a bill in the United States Congress or testified at a congressional committee hearing. This measure was standardized by changing states' totals to percentages.

The three independent variables mentioned above were operationalized as follows: a states commitment to environmental protection (PROTECTION) was measured by a widely used, comprehensive scale based on 23 environmental indicators, including such considerations as a state's spending on environmental control, priority given to environmental protection, implementation of the endangered species and wetlands acts, comprehensive land-use planning, and historic preservation; emissions of sulfur dioxide and nitrogen oxide (EMISSIONS) were measured as the total emissions (in tons) of each state; and a state's precipitation acidity (pH RATING) was measured in terms of its pH rating (pH is a widely used measure of acidity; acidity increases with decreasing pH; pH 7 is neutral). The three hypothesis generated from these variables were:

1. As a state's concern for the environment increases (as indicated by a larger Duerksen rating), its activity rating will increase.
2. As a state's emissions increase, its activity rating will increase.
3. As the pH measure of a state's precipitation decreases (meaning a higher level of precipitation acidity), the activity rating will increase.

The correlation matrix for this study is presented in Table 4. Before evaluating the actual values for the correlation coefficients, several points should be made. First, note that the dependent variable appears in the first row and the first column. This was intentionally done so that the researcher can quickly glance at the matrix and delineate the correlation of each of the independent variables with the dependent variable. Inspecting Table 4 shows that the first row and the first column are identical. Second, there is a diagonal of 1.0 values that go from the upper left entry to the lower right entry. This diagonal divides the matrix into two symmetrical parts allowing the researcher to focus on just one side of the matrix. (The values of 1.0 make

TABLE 4 Correlation Matrix

	Activity	Protection	Emissions	pH Rating
Activity	1.0000	.5161	.1965	−.3635
	(50)	(50)	(50)	(50)
	P = .	P = .000	P = .171	P = .009
Protection	.5161	1.0000	.0037	−.1044
	(50)	(50)	(50)	(50)
	P = .000	P = .	P = .979	P = .471
Emissions	.1965	.0037	1.0000	−.2696
	(50)	(50)	(50)	(50)
	P = .171	P = .979	P = .	P = .058
pH Rating	−.3635	−.1044	−.2696	1.0000
	(50)	(50)	(50)	(50)
	P = .009	P = .471	P = .058	P = .

Sources: SPSS For Windows 6.0, 1993; L.R. Alm, ''Regional Influences and Environmental Policymaking: A Study of Acid Rain,'' *Policy Studies Journal 21*:638–650 (1993).

sense because each variable is perfectly correlated with itself.) Third, each entry is made up of three values: the top entry is the value for Pearson's correlation coefficient (r), the middle value is the sample size (n), and the bottom value is the statistical significance level (p).

The first thing a researcher would do would be to inspect the first row (or column) to quickly gain a feel for how each of the independent variables is related to the dependent variable. The researcher initially checks to see which of the correlations are statistically significant. This is accomplished by choosing an acceptable level of significance (e.g., .05) and then observing which of the correlation coefficients have an observed significance level below this value. In this instance, the researcher would note that two of the correlation coefficients are statistically significant at the .05 level—PROTECTION ($p = .000$) and pH RATING ($p = .009$). Since the correlation for EMISSIONS ($p = .171$) is not statistically significant, the researcher could not be confident that this association was not due to chance and hence would not be able to accurately describe its strength. On the other hand, the researcher can be confident in the correlation measures of PROTECTION and pH RATING and can evaluate their strengths (and directions). For PROTECTION, the association is in the direction hypothesized (as a state's commitment to environmental protection increases, so does its activity level) and is fairly strong at .5161. The association for pH RATING is also in the hypothesized direction (as pH rating decreases, the activity rating increases) and is moderately strong at $-.3635$.

Because these two independent variables have both a statistically significant and substantial correlation to the dependent variable, they would be considered strong candidates to include in a multivariate regression equation. The same cannot be said for EMISSIONS. However, it should be mentioned that just because EMISSIONS does not reach statistical significance as a bivariate correlation does not mean that it cannot (and should not) be carried forward into a multivariate analysis. First, it could be carried forward as a control variable. Second, and more important, relationships often change when going from a bivariate to a multivariate analysis and this cannot be discovered unless that relationship is tested under those circumstances. In this study, the emissions variable was carried forward and actually reached statistical significance in a multiple regression analysis (Alm, 1993).

The second major use of the correlation matrix is as a first-line indicator of multicollinearity, a problem whose presence makes interpretation of the regression coefficients highly suspect. Multicollinearity occurs in multivariate analysis when independent variables are highly correlated to each other. The general rule for inspecting a correlation matrix for the presence of multicollinearity is to check each of the correlations (independent variable against independent variable) for high values (usually larger that .7). Inspecting Table 4 shows that only one association between the independent variables is statistically significant at the .10 level (emissions with pH rating, $p = .058$) and since its value (.2696) is much less than .7, multicollinearity does not appear to be a problem.[10] (Remember, the negative sign is used only for determining direction and is not a measure of strength.)

There exist many other tests for multicollinearity that the researcher may use. If multicollinearity does exist, the researcher must then decide how to deal with it. In any case, both the more sophisticated tests and the ''causes'' for multicollinearity are beyond the scope of this chapter and will be covered in the chapter on multivariate regression.

IX. PARTIAL CORRELATION

So far, this chapter has concerned itself with bivariate relationships; that is, with the effect that one variable is having on another variable without controlling for the effect of a third variable. However, just as there are techniques and procedures for controlling for the effect of a third

TABLE 5 Partial Correlation Coefficients

	Percent vote for Clinton	Percent federal land	Percent with college education
Zero order partial correlation coefficients			
Percent Clinton for Clinton	1.0000	−.4796	.5186
	(0)	(34)	(34)
	P = .	P = .003	P = .001
Percent federal land	−.4796	1.0000	−.2262
	(34)	(0)	(34)
	P = .003	P = .	P = .185
Percent with college education	.5186	−.2262	1.0000
	(34)	(34)	(0)
	P = .001	P = .185	P = .
Partial correlation coefficient controlling for percent of people with a college education			
Percent Clinton for Clinton	1.0000	−.4349	
	(0)	(33)	
	P = .	P = .009	
Percent federal land	−.4349	1.0000	
	(33)	(0)	
	P = .009	P = .	

Sources: SPSS For Windows 6.0, 1993; L.R. Alm, ''Regional Influences and Environmental Policymaking: A Study of Acid Rain,'' *Policy Studies Journal, 21*: 638–650 (1993).

variable at the nominal or ordinal levels of measurement (see Chapter 12, Statistics for Nominal and Ordinal Data), there exists a procedure for controlling for the effect of a third variable at the interval level. This procedure is called partial correlation.[11]

For example, let's look at the bivariate relationship modeled earlier in this chapter—that percent federal land in Oregon counties was negatively associated with percent vote for Clinton (**r** = −.4796). This value for Pearson's correlation coefficient was strictly bivariate in nature with no consideration given to any variables other than the two in the original equation. Suppose, however, that the researcher suspected that another variable (percent of people with a college education) was partially responsible for these results. In other words, the researcher wanted to observe the influence (affect) of percent federal land on the percent vote for Clinton while controlling for the influence (affect) of percent of people with a college education. Using linear correlation analysis (as described in previous sections of this chapter) we would derive the bivariate correlations displayed in the top portion of Table 5. Note that bivariate correlations (with no controls) are also referred to as ''zero order partial correlation coefficients.''

Table 5 lists the bivariate (zero order partial) correlation coefficients for all three pairs of variables, including the Pearson's **r** for our initial relationship of interest (percent federal land and percent vote for Clinton). From the table, we observe that the bivariate correlation between percent of people with a college education and percent vote for Clinton is statistically significant (p = .001) and equals .5186. In addition, we observe that the bivariate correlation between percent of people with a college education and percent federal land equals −.2262 and is not statistically significant (p = .185). This would mean that the percent of people with a college education had a statistically significant and moderately strong association with the percent vote for Clinton and a statistically insignificant relationship with percent federal land.

However, what the researcher is truly interested in is the effect that percent federal land is having on percent vote for Clinton while controlling for (accounting for) the effects of percent of people with a college education. To investigate this relationship, the researcher completes a partial correlation procedure resulting in the computer output displayed in the bottom half of Table 5. [These procedures are now easily completed with the help of computer statistical programs (in this case SPSS for Windows). If the reader is interested in the hand calculation of partial correlation, refer to Procedure 4 of the Appendix.] The results show that the linear correlation between percent federal land and percent vote for Clinton, while controlling for percent of people with a college education, is statistically significant ($p = .009$) and equals $-.4349$. This value represents a slight reduction from the original bivariate correlation ($r = -.4796$).

Because the reduction in **r** was slight, the researcher would conclude that the percent of people with a college education has little effect on the relationship between amount of federal land and the vote for Clinton. In fact, the explanatory value of the bivariate model was only reduced by a few percent (from $R^2 = .23$ to $R^2 = .19$). On the other hand, if it happens that the partial correlation value is substantially less than the original value, then the researcher might conclude that the original relationship was spurious; i.e., that the original relationship disappears when controlling for a third variable. For instance, in the above example if the bivariate correlation had been reduced substantially (say, from $-.4796$ to $-.04$) when controlling for percent of people with a college education, then the researcher would regard percent of people with a college education as being the more important influence on the vote for Clinton (as compared to percent federal land).

X. RESEARCH EXAMPLES

A. Linear Correlation

The most common use of linear correlation and simple regression in public administration research is as a first step leading to multivariate regression analysis. Linear correlation is especially helpful in establishing the existence of bivariate relationships, including the strength and direction of these relationships. In this context several published public administration research reports will be showcased.

Edward T. Jennings, Jr. recently completed a research project centered on the basic question of ''What effects do federal legislation, grants, and administrative activities have on state and local level employment and training programs'' (Jennings, Jr., 1994)? He focused on the effort of states and localities to produce successful employment and training programs, which he labeled as ''performance.'' His study used a survey of state officials, official state documents, and Bureau of Census data to measure and operationalize the dependent variable (Performance) and what he believed (based on prior research) to be the most important factors affecting state performance. The independent variables selected included active encouragement by the governor (Gubernatorial Leadership), the use of communication and decision-making approaches (Communication), the use of planning coordination (Planning), the use of operational coordination (Operations), the development of tools for use at the local level (SDAs—Service Delivery Areas), the total number of coordination tools (Total Coordination), and administrative organization (Administrative Structure).

A partial listing of Jennings' correlation matrix is displayed in Table 6. From these bivariate correlations (Pearson's **r**), Jennings concluded that gubernatorial leadership, the use of communication and decision-making approaches to coordination, the use of operational tools of coordination, and the total number of coordination tools used are all related to successful employment and training programs as measured by performance. The reader will note from Table

TABLE 6 Correlations between Performance and Gubernatorial Leadership, Communication, Planning, Operations, Support for SDA Coordination, Total Coordination, and Administrative Structure

	r
Gubernatorial leadership	.28*
Communication	.29*
Planning	.11
Operations	.19*
Service delivery area (SDAs)	.04
Total coordination	.29*
Administrative structure	.03

* Significant at .05 level or better.

Source: E.T. Jennings, Jr., ''Building Bridges in the Intergovernmental Arena: Coordinating Employment and Training Programs in the American States,'' *Public Administration Review*, 54: 59 (1994). Reprinted with permission from *Public Administration Review* © by the American Society for Public Administration (ASPA), 1120 G. Street NW, Suite 700, Washington DC 20005. All rights reserved.

6 that all of these associations reached statistical significance, which was defined by the author at the .05 level. Inspection of the actual values for Pearson's **r** reveals that while we can be confident that these associations do exist in our population of study, they would be considered weak, positive associations. The largest value reached is .29 and the smallest is .19. Remember that Pearson's **r** for positive association varies from 0 to 1 with 0 indicating no association and 1 indicating the strongest association.

Jennings also concluded that the use of planning, the development of tools for use at the local level (SDAs), and administrative structure were unrelated to performance. The reader will note that none of these associations reached statistical significance. In the end, Jennings used multivariate regression analysis to reinforce some of his initial findings—that gubernatorial leadership and overall coordination make a difference in performance at the state and local levels.

David H. Folz also uses linear correlation to explore the relationship between participation in recycling and public education strategies and incentives (Folz, 1991). His study uses survey responses from 264 recycling coordinators to provide empirical evidence about what program strategies work best for municipalities to maximize and sustain citizen participation in solid waste recycling programs. As in Jennings' work, Folz's final conclusions are based on the output from a multivariate regression analysis which was preceded by a linear correlation analysis of what were deemed important factors related to municipal recycling.

Table 7 shows the results of one of Folz's linear correlation outputs. From this output and using a .05 significance level, Folz concluded that communities with higher levels of participation in recycling used pamphlets, brochures, and bumper stickers ($r = .16$, $p = .00$), neighborhood or community information meetings ($r = .14$, $p = .01$), and paid newspaper advertisements ($r = .10$, $p = .04$). It should be noted that while these relationships are all statistically significant ($p < .05$), the values for Pearson's **r** represent weak, positive associations. On the other hand, the output shows that there also existed statistically significant, but weak, negative associations

TABLE 7 Correlations Between Participation in Recycling and Public Education Strategies and Incentives

	r	Significance	N
Strategy			
Pamphlets, brochures, or bumper stickers	.16	.00	241
Neighborhood or community information meetings	.14	.01	241
Paid newspaper advertisements	.10	.04	240
Speeches by officials to schools or local groups about recycling	.09	.07	241
Paid radio commercials	−.13	.01	240
Billboard advertisements	−.10	.04	241
Incentive			
Official public recognition of recycling efforts	.18	.00	239

Source: D.H. Folz, ''Recycling Program Design, Management, and Participation: A National Survey of Municipal Experience,'' *Public Administration Review*, 51: 229 (1991). Reprinted with permission from *Public Administration Review* © by the American Society for Public Administration (ASPA), 1120 G. Street NW, Suite 700, Washington DC 20005. All rights reserved.

between high levels of participation and paid radio commercials ($r = -.13$, $p = .01$) and billboard advertisements ($r = -.10$, $p = .04$). These variables all fall into what Folz delineated as his ''strategy'' factors. Under ''Incentive'' factors, Folz lists only one factor which he found statistically significant—official recognition of recycling efforts ($r = .18$, $p = .00$).

The only association which Folz includes in his output that is not statistically significant is the one between participation in recycling and speeches by officials to schools or local groups about recycling ($r = .09$, $p = .07$). About this relationship, the author states that ''speeches by officials to schools or local groups about recycling was related in the expected direction but just missed attainment of statistical significance at the .05 level.''

A very important point needs to be made about the author's inclusion of this variable and his statements. The author could have easily chosen a significance level of .10, which is a commonly used and accepted measure of significance in the public administration field. Instead, the author maintained the significance level with which he began his research. At times, it is very tempting for researchers to use a less rigorous significance level (after the statistical output is complete) that produces more statistically significant factors. It is much better to do what Folz did here—he included the variable and noted that while it was not significant, it did move in the expected direction. This accomplishes two things. First, by showing the actual significance obtained (in this case $p = .07$), it allows the reader to decide if that is an appropriate level. Second, it confirms the anticipated direction, although if one accepts the significance level at .05, this direction becomes meaningless.

B. Simple Regression

Another published study illustrates the use of simple regression in public administration research. James D. Ward recently studied the impact of the *City of Richmond v J.A. Croson Co.* court decision on minority set-aside programs in terms of dollars spent before and after the decision was handed down (Ward, 1994). His findings suggest that the percentage of African

TABLE 8 Regression Analysis of Dollars Spent versus Percent African Americans in City

Year	1988	1990
Pearson's r	.45*	.43*
Multiple R	.45331	.43124
R^2	.20549	.18597
Adjusted R^2	.14437	.12783

*Significance less than .10.

Source: J.D. Ward, ''Response to Croson,'' *Public Administration Review*, 54: 485 (1994). Reprinted with permission from *Public Administration Review* © by the American Society for Public Administration (ASPA), 1120 G. Street NW, Suite 700, Washington DC 20005. All rights reserved.

American (Black) population is the only variable (in his linear equation) that comes close to explaining the variance in dollars spent. Ward uses a bivariate (simple) regression to support his finding. The results of his bivariate regression are shown in Table 8.

Based on analysis of these results, the author states that the bivariate R-square value between ''percent Blacks in city'' and the dependent variable revealed that 20.5 percent of the variance in dollars spent on MBEs [Minority Business Enterprises] in 1988 and 18.5 percent of the variance in 1990 could be explained by knowing the percent Blacks in the city, regardless of region (Ward, 1994). Essentially, Ward is making the argument (based on his simple regression analysis) that the one factor that he could find that explains spending on MBEs is the percentage of Black population living in the relevant cities.

It is also important to note that in this case, the author chose not to use/interpret the unstandardized regression coefficient as a means of analysis, but instead focuses upon the explanatory power (r and R^2) of the percent Blacks. The author also chooses to highlight R^2 instead of adjusted R^2 while offering the caveat that the model's low adjusted R-square value diminishes its explanatory power. In this case, the author presents the reader with both values, allowing the reader to make his/her own determination.

Finally, note that in this instance the author chooses the .10 level of significance, as opposed to the authors of the two previous examples, who chose .05. These different choices just highlight the fact that the researcher gets to choose the significance level. It is quite common and acceptable in public administration research to use any (or all) of the three most common levels of significance (.10, .05, .01).

C. Partial Correlation

Dennis Daley's work exploring the relationship between types of administrative responsibility (defined as accountability, competence, fairness, and responsiveness) and methods of bureaucratic control (defined as executive control, pluralism, professionalism, and representative bureaucracy) makes use of partial correlation (Daley, 1985). Daley hypothesized that there exists specific relationships between accountability and executive control, competence and professionalism, fairness and representative bureaucracy, and responsiveness and pluralism. In setting up his research project, Daley conducted a five state survey of administrators, executives, and legislators where the respondents were asked to rank their preferences among the four types of

TABLE 9 The Linkages Between Administrative Responsibility and Bureaucratic Control Methods: Partial Correlation Coefficients

	Executive control	Pluralism	Professionalism	Representative bureaucracy
Accountability				
Administrators	.18***	−.08	.04	−.14**
Executives	.34**	−.22*	−.10	−.10
Legislators	.01	.00	−.05	.03
Competence				
Administrators	−.02	−.10**	.21***	−.07
Executives	−.07	−.06	.21*	−.07
Legislators	.13	.14*	.17*	−.11
Fairness				
Administrators	−.13**	.10*	−.13**	.16***
Executives	−.10	.03	−.03	.11
Legislators	.05	.10	−.11	−.05
Responsiveness				
Administrators	−.05	.08	−.10*	.06
Executives	−.20	.25*	−.07	.07
Legislators	−.17**	.03	.01	.12

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

Source: D. Daley, ''Administrative Responsibility and Control of the Bureaucracy: The Dog That Did Not Bite,'' *State and Local Government Review*, 17: 195–199 (1985). Reprinted by permission of the author and the Carl Vinson Institute of Government, University of Georgia.

administrative responsibility and the four methods for controlling the bureaucracy. As a first step in the analysis, a partial correlation analysis was completed with the results listed in Table 9.

From the results, the author first concludes that the resultant partial correlations are only marginally different from the uncontrolled, zero-order relationships. (The author did not present the original zero-order measures.) To the reader this would indicate that no spurious relationships exist and that the original bivariate relationships could be viewed as appropriate. However, the author uses the partial correlations as a basis for analysis because they provide an association value while controlling for the effects of the other variables. For instance, the partial correlation coefficient for administrators between accountability and executive control equals .18 and is statistically significant to the .001 level. The reader would interpret this to mean that there was a statistically significant weak positive association between accountability and executive control (for administrators) while controlling for the effects of the other three types of bureaucratic control methods (pluralism, professionalism, and representative bureaucracy).

Based on the partial correlation analysis (see Table 9), Daley concluded that there was some evidence supporting the relationships between accountability and executive control, and between competence and professionalism. He also concluded that the projected relationships linking responsiveness to pluralism and fairness to representative bureaucracy were not confirmed. Inspection of Table 9 indicates how Daley came to these conclusions. First, the partial correlation coefficients for the two supported relationships are substantially larger (although still weak) than the partial correlation coefficients for the unsupported relationships. Second, the supported relationships generally show statistically significant associations, while the unsupported relationships do not. In his final analysis, Daley concludes that the results of partial correlation analysis show that the relationships are not overwhelming because even the relationships which reached statistical significance were very weak.

TABLE 10 Zero Order Partial Correlation Coefficients between Selected Independent Variables and three Components of Legislative Professionalism

Components of legislative professionalism	Service sector	Manufacturing sector	Per capita income	Bureaucracy	Budget size
Staff/support					
1953	.90	.74	.34	.76	.82
1961	.89	.78	.32	.86	.90
1971	.93	.78	.44	.92	.96
1981	.93	.83	.38	.88	.95
Compensation					
1953	.79	.81	.48	.65	.73
1961	.86	.86	.51	.71	.78
1971	.71	.68	.59	.69	.69
1981	.69	.70	.48	.68	.70
Session length					
1953	.32	.37	.27	.30	.37
1961	.27	.30	.13	.25	.28
1971	.48	.53	.27	.48	.48
1981	.59	.59	.20	.58	.62

Source: G.F. Moncrief, ''Dimensions of the Concept of Professionalism in State Legislatures: A Research Note,'' *State and Local Government Review*, 17: 195–199 (1985). Reprinted by permission of the author and the Carl Vinson Institute of Government, University of Georgia.

Another approach for using partial correlation analysis in support of public administration research is illustrated by the work of Gary Moncrief (1988). Moncrief collected data for all 50 states at four different times to see if certain state characteristics (size of the service and manufacturing sectors, per capita income, size of the bureaucracy, and size of the budget) influenced different components of legislative professionalism (staff and support, compensation, and session length).[12] He first calculated the bivariate correlations (also called the zero order partial correlation coefficients) for the three components of legislative professionalism (staff support, compensation, and session length) against the selected independent variables (service sector, manufacturing sector, per capita income, bureaucracy, and budget size). The results are shown in Table 10. Note that Moncrief did not include any information about the statistical significance of these correlations. The reason for this is that the 50 states were treated as an entire population and not as a sample, hence the use of significance was not appropriate. (See Note 4.)

From the results of the bivariate correlations, Moncrief concluded that because the correlations between the independent and dependent variables are moderately high for many pairs and very high for some pairs, there exists initial corroboration of his hypothesis that socio-economic factors and governmental complexity are strongly associated with legislative professionalism. Yet Moncrief felt that it would be useful to observe the relationship of each of the independent variables with the components of legislative professionalism while controlling for the other socio-economic independent variables. The results are shown in Table 11.

Inspection of the partial correlation coefficients led Moncrief to conclude that the original associations were weakened considerably when controlling for the effects of all the independent variables in the model. This is indicated by the substantial lower values of the partial correlation coefficients as compared to the original bivariate correlations. More importantly, however, was that three of the independent variables retained important associations even with the relatively stringent controls. The author felt that the most interesting feature was the ''staying power'' of

TABLE 11 Partial Correlation Coefficients between Selected Independent Variables and three Components of Legislative Professionalism

Components of legislative professionalism	Service sector	Manufacturing sector	Per capita income	Bureaucracy	Budget size
Staff/support					
1953	.82	−.12	−.14	.13	.25
1961	.70	−.30	−.22	.42	.50
1971	.46	−.48	.04	.01	.59
1981	.49	−.15	.10	.10	.61
Compensation					
1953	.28	.38	.19	−.18	.17
1961	.30	.37	.22	−.19	−.04
1971	−.06	−.01	.44	−.02	.06
1981	.06	.29	.38	.20	.07
Session length					
1953	−.10	.11	.15	−.11	.24
1961	.01	.15	.00	−.05	.16
1971	−.04	.17	.08	−.07	.12
1981	.20	.23	−.04	.09	.21

Source: G.F. Moncrief, ''Dimensions of the Concept of Professionalism in State Legislatures: A Research Note,'' *State and Local Government Review*, 17: 195–199 (1985). Reprinted by permission of the author and the Carl Vinson Institute of Government, University of Georgia.

the service sector variable as a correlate with staff/support. The author could make this statement because the partial correlation coefficient representing the association between staff/support and service sector is substantial for all four years studied. Another way to look at this would be to note that the reduction from the original bivariate correlations to the partial correlations were not as substantial for this association as compared to the others.

Moncrief noted two other results of his partial correlation analysis: that budget size has a strong independent effect on staff/support and that per capita income exhibits a stable independent impact on compensation. The rational for these conclusions was based on the same logic posited above. In the end, Moncrief concluded that certain independent variables (e.g. the size of the service sector, budget size, and per capita income) have important independent relationships with specific measures of professionalism.

XI. CONCLUSION

When public administration researchers seek to determine the relationship between two variables that are measured at the interval or ratio level of measurement, they turn to linear correlation and simple regression analysis to provide an initial indication of that relationship. Specifically, linear correlation and simple regression analyses help describe the exact nature of the relationship between two variables, allowing the researcher to predict the value of one variable based on the value of the other. Furthermore, these two statistical procedures allow the researcher to estimate both the direction and strength of the association between the variables. In the end, linear correlation and simple regression provide an indication that two variables may be causally connected, and in so doing, also provide the foundation from which to begin multivariate analysis.

APPENDIX

PROCEDURE 1: Calculation of Regression and Correlation Coefficients

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
51.40	31.81	15.59	−4.83	−75.30	243.05	23.33
17.00	47.37	−18.81	10.73	−201.83	353.82	115.13
47.80	39.03	11.99	2.39	28.66	143.76	5.71
.01	45.80	−35.80	9.16	−327.93	1281.64	83.91
2.70	42.77	−33.11	6.13	−202.96	1096.27	37.58
21.50	40.70	−14.31	4.06	−58.10	199.09	16.48
49.40	34.49	13.59	−2.15	−29.22	184.69	4.62
59.40	34.76	23.59	−1.88	−44.35	556.49	3.53
75.90	35.73	40.09	−.91	−36.48	1607.21	.83
48.10	30.83	12.29	−5.81	−71.40	151.04	33.76
2.90	36.03	−32.91	−.61	20.08	1083.07	.37
60.00	28.47	24.19	−8.17	−197.63	585.16	66.75
70.60	28.86	34.79	−7.78	−270.67	1210.34	60.53
62.50	39.61	26.69	2.97	79.27	712.36	8.82
45.10	37.80	9.29	1.16	10.78	86.30	1.35
16.70	36.59	−19.11	−.05	.96	365.19	.00
57.90	32.80	22.09	−3.84	−84.83	487.97	14.75
51.10	29.77	15.29	−6.87	−105.04	233.78	47.20
67.70	26.80	31.89	−9.84	−313.80	1016.97	96.83
54.40	48.78	18.59	12.14	225.68	345.59	147.38
30.20	44.41	−5.61	7.77	−43.59	31.47	60.37
37.50	34.00	1.69	−2.64	−4.46	2.86	6.97
71.50	23.81	35.69	−12.83	−457.90	1273.78	164.61
29.40	37.28	−6.41	.64	−4.10	41.09	.41
13.70	33.79	−22.11	−2.85	63.01	488.85	8.12
26.30	55.34	−9.51	18.70	−177.84	90.44	349.69
8.80	37.29	−27.01	.65	−17.56	729.54	.42
8.20	32.44	−27.61	−4.20	115.96	762.31	17.64
19.70	43.89	−16.11	7.25	−116.80	259.53	52.56
20.00	34.55	−15.81	−2.09	33.04	249.96	4.37
47.70	34.43	11.89	−2.21	26.28	141.37	4.88
57.60	29.53	21.79	−7.11	−154.93	474.80	50.55
16.00	42.50	−19.81	5.86	−116.09	392.44	34.34
2.60	40.39	−33.21	3.75	−124.54	1102.90	14.06
23.20	31.05	−12.61	−5.59	70.49	159.01	31.25
14.80	35.50	−21.01	−1.14	23.95	441.42	1.30
1289.31	1319.00	.15*	−.04*	−2539.19	18585.56	1570.40

$$\bar{x} = \frac{\sum x}{n} = \frac{1289.31}{36} = 35.81 \quad \bar{y} = \frac{\sum y}{n} = \frac{1319}{36} = 36.64$$

$$b = \frac{\sum (x - \bar{x})\ (y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{-2539.19}{18585.56} = -.1366$$

$$a = \bar{y} - b\bar{x} = 36.64 - (-.1366)\ (35.81) = 36.64 + 4.89 = 41.53$$

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{-2539.19}{\sqrt{(18585.56)\ (1570.40)}} = \frac{-2539.19}{5402.48} = -.4700$$

* Due to rounding errors, numbers are not exact. In these two cases, actual values should equal zero.

Source: W.F. Matlack, *Statistics for Public Managers*, F.E. Peacock Publishers, Itasca, Illinois, 1993, pp. 217–228.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

PROCEDURE 2 Hypothesis Test for the Regression Coefficient

x	y	y_{pred}	$y - y_{pred}$	$(y - y_{pred})^2$
51.40	31.81	34.51	−2.70	7.29
17.00	47.37	39.21	8.16	66.59
47.80	39.03	35.00	4.03	16.24
.01	45.80	41.43	4.27	18.23
2.70	42.77	41.16	1.61	2.59
21.50	40.70	38.59	2.11	4.45
49.40	34.49	34.78	−.29	.08
59.40	34.76	33.42	1.34	1.80
75.90	35.73	31.16	4.57	20.88
48.10	30.83	34.96	−4.13	17.06
2.90	36.03	41.13	−5.10	26.01
60.00	28.47	33.33	−4.86	23.62
70.60	28.86	31.89	−3.03	9.18
62.50	39.61	32.99	6.62	43.82
45.10	37.80	35.37	2.43	5.90
16.70	36.59	39.25	−2.66	7.08
57.90	32.80	33.62	−.82	.67
51.10	29.77	34.55	−4.78	22.85
67.70	26.80	32.28	−5.48	30.03
54.40	48.78	34.10	14.68	215.50
30.20	44.41	37.40	7.01	49.14
37.50	34.00	36.41	−2.41	5.81
71.50	23.81	31.76	−7.95	63.20
29.40	37.28	37.51	−.23	.05
13.70	33.79	39.66	−5.87	34.46
26.30	55.34	37.94	17.40	302.76
8.80	37.29	40.33	−3.04	9.24
8.20	32.44	40.41	−7.97	63.52
19.70	43.89	38.84	5.05	25.50
20.00	34.55	38.80	−4.25	18.06
47.70	34.43	35.01	−.58	.34
57.60	29.53	33.66	−4.13	17.06
16.00	42.50	39.34	3.16	9.99
2.60	40.39	41.17	−.78	.61
23.20	31.05	38.36	−7.31	53.44
14.80	35.50	39.51	−4.01	16.08
			.06*	1209.13

From Procedure 1: $\sum(x - \bar{x})^2 = 18585.56$; $a = 41.53$; $b = -.1366$

$$y_{pred} = a + bX = 41.53 - .1366(x)$$

$$s_{yx} = \sqrt{\frac{\sum(y - y_{pred})^2}{n - 2}} = \sqrt{\frac{1209.13}{34}} = \sqrt{35.56} = 5.96$$

$$s_b = \frac{s_{yx}}{\sqrt{\sum(x - \bar{x})^2}} = \frac{5.96}{\sqrt{18585.56}} = \frac{5.96}{136.33} = .04$$

$$t_{calc} = \frac{b - \beta}{s_b} = \frac{-.1366 - 0}{.04} = -3.42$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

PROCEDURE 2 Continued

From t-table for two-tail significance equal to .05 and degrees of freedom equal to 34, read t_{table} equals 2.034. Since t_{calc} is greater than t_{table}, reject the null hypothesis that the regression coefficient equals zero and accept the hypothesis that for this set of data, the regression coefficient equals −.1366. Note: Remember that you may disregard the sign for the calculated value of t in making this interpretation as it does not affect magnitude; it only refers to direction (positive or negative).

* Due to rounding errors, numbers are not exact. In this case, actual value should equal zero.

Source: W.F. Matlack, *Statistics For Public Managers*, F.E. Peacock Publishers, Itasca, Illinois, 1993, pp. 219–221.

PROCEDURE 3 Hypothesis Test for the Correlation Coefficient

From Procedure 1: $r = -.4700$

$$t_{calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-.4700\sqrt{36-2}}{\sqrt{1-(-.4700)^2}} = \frac{-2.74}{.88} = -3.11$$

From t-table for two-tail significance equal to .05 and degrees of freedom equal to 34, read t_{table} equals 2.034. Since t_{calc} is greater than t_{table}, reject the null hypothesis that the correlation coefficient equals zero and accept the hypothesis that for this set of data, the correlation coefficient equals −.4700.
Note: Remember that you may disregard the sign for the calculated value of t in making this interpretation as it does not affect magnitude; it only refers to direction (positive or negative).

Source: W.F. Matlack, *Statistics For Public Managers*, F.E. Peacock Publishers, Itasca, Illinois, 1993, p. 227.

PROCEDURE 4 Calculation of Partial Correlation Coefficient

For this example, the variables are defined as follows:

Dependent Variable	=	X_1	percent vote for Clinton
Independent Variable	=	X_2	percent federal land
Control Variable	=	X_3	percent of people with a college education

The bivariate correlation coefficients delineating the associations between each of the above variables are obtained using procedures established earlier in this chapter (See Procedure 1 and Table 5) and displayed below:

$r_{12} = -.4796$
$r_{13} = -.5186$
$r_{23} = -.2262$

Using these values, we can then calculate $r_{12.3}$ as follows:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{-.4796 - (.5186)(-.2262)}{\sqrt{1-(.5186)^2}\sqrt{1-(-.2262)^2}} = \frac{-.4796 + .1173}{(.8550)(.9741)} = -.4350$$

Source: R.L. Cole, *Introduction to Political Science and Policy Research*, St. Martin's Press, New York, 1996, p. 253.

NOTES

1. Remember that there exist several levels of measurement. Nominal refers to the classification of cases into groups. Ordinal classifies cases into groups and allows ordering of those groups. Interval not only allows classification and ordering but indicates how far each case is from each other case. Ratio is an extension of the interval level (it establishes a true zero value whereas an interval level allows for an arbitrary zero value) and as such is often treated identically to an interval level measure. For a more comprehensive discussion of levels of measurement see K. Hoover and T. Donovan, *The Elements of Social Science Thinking*, Sixth Edition, St. Martin's Press, New York, 1995, pp. 92–98.
2. If the slope of the line was negative (**b** is less than zero), then the relationship would be represented by the equation $\mathbf{Y = a - bX}$.
3. For an explanation of how the method of least squares works, see S.K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, Second Edition, Radius Press, New York, 1991, pp. 164–165.
4. Statistical significance is founded in sampling theory. The idea is for the researcher to make empirically supported generalizations about a population (the entire set of relevant units of analysis) based on a sample (subset) from that population. Accuracy of these generalizations is a function of the definition of the population, the sample design, and the size of the sample. If the entire set of data is available to the researcher, there would be no need to base results on a sample, hence statistical significance (and sampling theory) would not be appropriate to use. For a full discussion of these concepts, see C. Frankfort-Nachmias and D. Nachmias, *Research Methods in the Social Sciences*, Fifth Edition, St. Martin's Press, New York, 1996, pp. 177–202.
5. R^2 is the *multiple* coefficient of determination that indicates the degree of variation in the dependent variable explained by *all* the independent variables in the model. Since a simple (bivariate) model has only one independent variable, R^2 should actually be symbolized by r^2 (or the square of the Pearson's correlation coefficient). For purposes of this chapter, we will use R^2 to represent the coefficient of determination for all regression models, including bivariate regression.
6. The reader will observe that the values from the computer output are slightly different than the values from the hand calculations. These differences are simply due to the rounding-off procedures inherent in calculating the values by hand.
7. The equation for Adjusted R^2 is: Adjusted $R^2 = R^2 - [p(1 - R^2)/(N - p - 1)]$, where *p* is the number of independent variables in the equation and *n* is the sample size. One can see from the equation that Adjusted R^2 takes into consideration both sample size and the number of independent variables in the equation in an attempt to correct R^2 to more closely reflect the goodness of fit of the model in the population. For a more extensive discussion of Adjusted R^2, see M.J. Norusis, *SPSS For Windows Base System User's Guide: Release 6.0*, SPSS Inc., Chicago, 1993, p. 318.
8. The Durbin-Watson test has some drawbacks. There exists a large indeterminate range (depending on the number of variables and the sample size) that you cannot accept or reject the null hypothesis of no autocorrelation. Further, the Durbin-Watson statistic tests only for first-order autocorrelation and should not be used if there are lagged values of the dependent variable used as independent variables. Instead, if exploring time series data, researchers may now use more sophisticated autocorrelation, partial correlation, and cross-correlation functions to test for the presence of autocorrelation. For a more extensive discussion of these techniques, see M.J. Norusis, *SPSS For Windows Base System User's Guide: Release 6.0*, SPSS Inc., Chicago, 1993, pp. 615–624.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

9. Three common autocorrelations for transformed time series values are the natural log transformation (using the natural logarithm—base e—of the series), the difference transformation (calculating the difference between successive values in the series), and the seasonally difference transformation (calculating the difference between series values a constant span apart). For more information on these transformations, see M.J. Norusis, *SPSS For Windows Base System User's Guide: Release 6.0*, SPSS Inc., Chicago, 1993, pp. 620–621.
10. For multiple regression, a better approach of assessing multicollinearity is to regress each independent variable against all the other independent variables. For an expanded discussion of this technique, see M.S. Lewis-Beck, *Applied Regression: An Introduction*, Sage Publications, Beverly Hills, California, 1980, p. 60.
11. The procedure described here and illustrated in Procedure 4 of the Appendix is based on controlling for only one variable. For a discussion of partial correlation controlling for more than one variable at a time, see R.L. Cole, *Introduction to Political Science and Policy Research*, St. Martin's Press, New York, 1996, p. 253–255.
12. While not discussed here, Moncrief also controlled for population in his model.

REFERENCES

Alm, L.R. (1993). ''Regional Influences and Environmental Policymaking: A Study of Acid Rain,'' *Policy Studies Journal*, *21*: 638–650.

Alm, L.R. and S.L. Witt (1995). ''Environmental Policy in the Intermountain West, The Rural-Urban Linkage,'' *State and Local Government Review*, *27*: 127–136.

Alm, L.R. and S.L. Witt (1997). ''The Rural-Urban Linkage to Environmental Policy Making in the American West: A Focus on Idaho,'' *The Social Science Journal*, *34*: 271–284.

Daley, D. (1985). ''Administrative Responsibility and Control of the Bureaucracy: The Dog That Did Not Bite,'' *State and Local Government Review*, *17*: 195–199.

Farley, C. (1995). ''The West: Sorry No Vacancies,'' *Time*, August 7, 1995, pp. 34–35.

Folz, D.H. (1991). ''Recycling Program Design, Management, and Participation: A National Survey of Municipal Experience,'' *Public Administration Review*, *51*: 227–231.

Hays, S. (1991). ''The New Environmental West,'' *Journal of Policy History*, *3*: 223–248.

Hoover, K. and T. Donovan (1995). *The Elements of Social Science Thinking*, Sixth Edition, St. Martin's Press, New York, pp. 92–111.

Jennings, Jr., E.T. (1994). ''Building Bridges in the Intergovernmental Arena: Coordinating Employment and Training Programs in the American States,'' *Public Administration Review*, *54*: 52–60.

Lewis-Beck, M.S. (1980). *Applied Regression: An Introduction*, Sage Publications, Beverly Hills, California, pp. 13–42.

Moncrief, G.F. (1988). ''Dimensions of the Concept of Professionalism in State Legislatures: A Research Note,'' *State and Local Government Review*, *20*: 128–132.

Norusis, M.N. (1991). *SPSS/PC+ Studentware Plus*, Prentice Hall, Englewood Cliffs, New Jersey, pp. 197–400.

SPSS for Windows is a copyright © of SPSS Inc., 444 Michigan Avenue, Chicago, IL 60611.

Ward, J.D. (1994). ''Responses to Croson,'' *Public Administration Review*, *54*: 483–485.

Welch, S. and J. Comer (1988). *Quantitative Methods For Public Administration*, Brooks/Cole Publishing Co., Pacific Grove, California, pp. 168–296.

Wonnacott, T.H. and R.J. Wonnacott (1984). *Introductory Statistics for Business and Economics*, Third Edition, John Wiley & Sons, New York, pp. 624–695.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

14

Cross-Sectional, Longitudinal, and Time-Series Data: Uses and Limitations

Lynn Burbridge
Rutgers University, Newark, New Jersey

I. INTRODUCTION

In any research endeavor, having a clear sense of one's goals in undertaking a project always comes first. The choice of an appropriate database follows, as dictated by these research goals. While it is tempting to allow the database to drive the study, such an approach is invariably flawed. Generally speaking, no one database can give all the information that is sought; therefore, having a clear set of research objectives is essential in sorting through different database options so that the one that is *most suited to the key research objectives* can be identified.

This chapter discusses the issues involved in using three types of data: cross-sectional data, time-series data, and longitudinal data. Most students in public administration or in any of the social sciences learn about quantitative methods using cross-sectional data. Conceptual and methodological issues in working with these data are easier to present and to understand. Therefore, in describing the issues involved in using different types of data, cross-sectional data are used as a starting point and a frame of reference for the discussion of time-series and longitudinal databases.

The outline of this chapter is to discuss each type of database in turn, identifying their characteristics, their uses, and methodological issues specific to each. The discussion is developed keeping in mind the first paragraph, that research objectives ultimately must dictate which database is used. Of the three types of databases, no one approach is good or bad. One may be better than the other for specific purposes. In many instances, however, it could be argued that the best of all worlds is to approach a subject from all three points of view, since all three have something to offer. It is hoped that at the end of this chapter, readers will appreciate the many advantages of these databases as well as their limitations.

II. CROSS-SECTIONAL DATA

A. Description

Cross-sectional data are data taken at a point in time. They are distinguished from time-series and longitudinal data in terms of their relationship to time. Time-series data usually consist of discrete indicators that are collected for a relatively long period of time in repeated, relatively short, intervals of time. Longitudinal data usually follow a cohort of individuals or other entities

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

through time, with ''the cohort'' usually being persons or entities born at a specific time or entering the data at a particular time. Thus, cross-sectional data represent a ''snapshot'' of one point in time, while the other databases allow for *time-dependent* analyses.

The differences between cross-sectional data and the other forms of data are often more apparent than real, however. One of the most used cross-sectional databases in the United States—the United States Census of the Population—can be turned into a time series database. Burbridge (1994) in her study of patterns of employment in government, the third sector (a proxy for the nonprofit sector), and the private sector followed trends in employment from 1950 to 1990, using census records for 1950, 1960, 1970, 1980, and 1990. The analysis also looked at cohort changes: those who were 25–34 in the 1950 Census, were 35–44 in the 1960 Census, were 45–54 in the 1970 Census and so on. Cohort differences can be identified by tracking ten-year age groupings, as they age, through different census years.

But using the U.S. Census as a time-series or longitudinal database is inherently limited. Since it is only collected every ten years, large chunks of history are lost to the analysis. Between 1960 and 1970 major events occurred that may have affected employment patterns, such as the Civil Rights Act of 1964, the effects of which are not captured until 1970. If the adjustment to the new legislation is slow, this may not be a problem. If changes began occurring immediately the analysis can only capture a snapshot of changes as of 1970. If changes continue to occur after 1970, they will only be captured in the snapshot of 1980. Thus, one's ability to describe or understand the historical processes that have produced any changes are seriously compromised.

In addition, the number of variables that could be used to look across time were limited. When Burbridge wanted to look at trends by occupation and sector, inconsistencies were found in how occupations were defined. The census regularly changes occupational classifications in order to keep up with the many occupational changes in the labor market. Being that the Census was primarily designed to present a point in time cross-sectional view of the U.S., maintaining old occupational categories for the sake of examining trends was less important than the accurate description of contemporary occupations.

Similarly, while a cohort analysis can be constructed from census data, it will only allow examination of inter-cohort differences. Intra-cohort differences would require more detailed information on characteristics, by cohort. Thus, while the Burbridge study found evidence that younger, white women in 1990 were relying less on government employment than younger, black women, the reasons for this are hard to discern from these data.

Why, then, use these data? One of the key objectives of the Burbridge study was to examine trends since World War II and differences in employment by race or ethnicity and sex. The U.S. Census of Population had the advantage of being available on computer tapes going back as far as 1940 and of being large enough to capture differences for ten race-by-sex categories (American Indian, Asian, black, Hispanic, and white men and women). It was the only database that could do this. Thus, the primary objectives of the study dictated the choice of the database, in spite of its limitations.

Why, then, does the U.S. Census only collect decennial snapshots? Cost is, of course, a primary issue. Collecting population data at more frequent intervals would be expensive. Even if the money were available, however, it would be difficult to get the population to cooperate more often. Methodological problems associated with attrition from the database is a key issue plaguing longitudinal databases that often require a long-term commitment from respondents in the data. Since one of the goals of the U.S. Census is to capture an accurate count of the entire population, a high response rate is critical. More frequent census counts could make this less possible. In spite of this the U.S. Census is frequently criticized for under counting certain populations.

What the Census loses in terms of its usefulness in time-dependent analyses it gains in the richness of the cross-sectional data. Not only does the census provide detailed data on minority

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

populations that may not be captured in smaller databases, it allows analysis of states, cities and even neighborhoods that would be difficult with other data. Indeed, any analysis focusing on a small sub-group in the population is probably better served by the census than by most other databases. Similarly, other cross-sectional databases sacrifice depth in time for breadth in a given point in time. If the goal of the data collection is to obtain a lot of information about what is happening in a particular point in time—and there are limits on what can be spent—than a cross-sectional database is the appropriate choice.

B. Uses of Cross-Sectional Data

From the previous discussion it appears that cross-sectional data are useful for two reasons: they often provide great breadth in data and they allow for comparisons on several dimensions. Taking three issues that are important to Americans—employment and earnings, education and crime—there are cross-sectional databases that present considerable data on these issues.

Data from the U. S. Census and the Current Population Survey are frequently used to examine differences in employment and earnings. Numerous studies have examined differences by race or ethnicity (Farley and Allen, 1989) or by gender (Sokoloff, 1992).

These data have also been used to examine within race differences, such as differences between immigrant and non-immigrant Hispanics or Asians (Sandefur and Tienda, 1988; U.S. Commission on Civil Rights, 1992). Inter-industry differences are frequently examined as well as regional variations in employment and earnings (Vroman, 1987).

Every seven years the National Assessment of Educational Progress tests American students on their verbal and mathematical skills and information are produced on differences by race or ethnicity, by gender, and by region (Mullins, 1993). The Federal Bureau of Investigation provides Uniform Crime Reports that have been used to examine demographic and regional differences in crime (U.S. Department of Justice, 1993).

In addition, international agencies are interested in cross-country differences for these and other measures (United Nations Development Program, 1995). As with the U.S., these data have appeared in agency publications, as well as being used by researchers in their own studies.

It should be noted, however, that in spite of the fact that these are cross-sectional databases, rarely are data presented or analyzed for one year alone. It is difficult to find a publication or article that does not show or examine these data for multiple years. The desire to look at change over time is almost irresistible to researchers and analysts.

Further, for those interested in program and policy evaluation, some kind of time series is essential. It is not possible to assess the effectiveness of a program or policy without looking over time. So while there is a discussion of program and policy evaluation with respect time-series and longitudinal data, none is presented here.

C. Methodological Issues

One of the first things a student learns in college methods courses, once he or she gets beyond probability theory, is how to construct a regression analysis using ordinary least squares (OLS). OLS is particularly useful when examining behavioral relationships in cross-sectional data. As shall be discussed shortly, issues of auto-correlation arise when examining data across time, making OLS a less useful vehicle for analyzing data. But for cross-sectional data, OLS works well under a certain set of conditions: the dependent variable is continuous, the explanatory variables are exogenous, the relationship between the dependent variable and independent variables is linear, the disturbances are uncorrelated with a 0 mean, and the variance is constant.

These are fairly restrictive assumptions even for many cross-sectional studies. If there is heteroscedasticity in the data or if a qualitative dependent variable (dichotomous or polytono-

mous) is being used, the variance will not be constant (Aldrich, 1989, Harvey, 1981). Further, in the case of a qualitative dependent variable, what is being estimated is a probability that is constrained to take a value between 0 and 1; it is unlikely that the relationship between the dependent and independent variables will be linear.

Under these circumstances, estimates must be obtained by maximum likelihood estimation, a more general method than OLS. Maximum likelihood estimation is an iterative process that does what it sounds like it does, it attempts to find the most likely parameters to explain the data (Hagenaars, 1990). As well as being the method of choice with cross-sectional data that violate the conditions for OLS, most of the methods used with time-series and longitudinal data are based on this method. Thus, maximum likelihood estimation is increasingly the estimation of choice, not just because of an increasing use of time-dependent data, but also because it has become computationally easier and less expensive with advances made in computer technology (Harvey, 1981). Programming has also become easier, as many software packages incorporate procedures using maximum likelihood for specific kinds of analysis, such as logit or probit procedures for use with qualitative dependent variables. So even with cross-sectional data, researchers have been more willing to relax the assumptions they make about their data and to use this methodology.

Nevertheless, it would be fair to say that analysis of cross-sectional data is more likely to involve OLS, which is conceptually and computationally easier than the methods used for other kinds of data. Further, as suggested earlier, attrition bias is less of an issue for cross-sectional data since these data do not require the same commitment from subjects as longitudinal data. This does not mean that a related issue, selection bias, can not be a problem if the cross-sectional data are not judged to be representative of the sample of people or entities being researched. But it tends to loom larger as an issue with other data.

In spite of the advantages of cross-sectional data in these respects, it is problematic in other ways. The increasing interest in other forms of data, in part, reflect difficulties in the interpretation of cross-sectional data. One frequently expressed concern is that with cross-sectional data, age effects are frequently confused with cohort effects (Hagenaars, 1990). For example, Schaie and Willis' (1991) longitudinal study of adult personality demonstrated that findings from cross-sectional analyses exaggerated the extent to which people become more inflexible as they age. What cross-sectional studies interpreted as an age effect was actually a cohort effect. More recent cohorts have scored on personality tests as more flexible, but as they age these cohorts will be more flexible than older cohorts have been as they aged. The cross-sectional analyses suggested a personality change to greater inflexibility with age that was not found to the same extent in longitudinal data.

Nevertheless, it should also be noted that the problem of confounding age effects, cohort effects, and (historical) period effects also plagues analyses with other forms of data (Hagenaars, 1990). Thus, this issue will appear again in the following discussion of other kinds of data.

III. TIME-SERIES DATA

A. Description

McCleary and Hay (1980) describe time-series as

> a set of N time-ordered observations of a process. Each observation should be an interval level measurement of the process and the time separating successive observations should be constant . . . [T]ime series is a discrete data set . . . [which] may be a measure of some underlying continuous process (pp. 21).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Time series data can be collected on any number of processes. The U.S. Government collects many data that fit McCleary and Hay's definition: The Bureau of Labor Statistics (1996) collects monthly data on employment and unemployment rates that are used to assess the strengths and weaknesses in the labor market; the U. S. Department of Commerce (1995) collects monthly data on business activity (e.g., value of output in manufacturing industries) that are used to assess economic growth and stability.

Obviously, in any given month, these data can be used as cross-sectional data, showing again the thin line between different kinds of data. What makes them more amenable to time series analysis—in comparison to the U.S. Census discussed above—is that these data are collected monthly. Thus, while there may be only two observations of census data over an eleven-year period (e.g. 1980 and 1990), there are 132 observations of employment and business activity data over the same period. Time-series analysis is particularly suited to these kind of data.

Basically, time-series data can be used to examine historical trends and patterns. But within many historical processes are peaks and troughs: unemployment goes up and down, business activity goes through booms and busts. If we cut into a set of data at any one point in time, the data may reflect a peak period or a trough period without really being a good indicator of the average level of employment or business activity, or without identifying important time-dependent patterns in the data.

For example, if we conducted a cross-sectional analysis with one month of unemployment data, we would find that youth have higher unemployment rates than adults. But this is only a part of the story. If we examined unemployment data for two years, we would notice that youth unemployment rises in the summer—when young people are out of school and looking for work—and then falls again in September when they return to school, indicating a seasonal pattern to youth unemployment. If we examined these data over ten years, we would find that youth unemployment is very sensitive to the business cycle, rising in periods of slow economic growth. If we examined these data over 30 or 40 years we would find a long term, secular increase in youth unemployment—particularly for black youth.

But it is not only the length of the data, but the frequency with which data is collected that is of importance. Using U.S. Census data, for example, we would be able to find the long-term increase in youth unemployment, but seasonal or business-cycle changes would allude us. What permits us to examine these issues is the availability of monthly data. Thus, the frequency with which the data are collected is just as important for our analysis as how long the data are collected. The more data points that are available, the better we are able to understand underlying historical processes.

For a researcher deciding whether or not to use time-series data, the first question he or she should ask is whether the issue he or she is examining is essentially a dynamic process. Does the understanding of the phenomenon require analyzing change over time?

In deciding whether change should be observed using time-series data as opposed to longitudinal data, Markus (1979) suggests that time-series are observations usually collected on a single entity such as a country, a corporation or so on, while longitudinal data are collected on multiple entities (e.g. a cohort of people) followed over time. Further, since longitudinal data is collected on multiple entities, it is only feasible to collect data at relatively long intervals of time (every two, three, four, or more years). Thus, longitudinal data are not appropriate for analyses requiring more frequent data collection.

Perhaps these points can be best illustrated by indicating that time-series is really the only kind of data that can be collected on one individual, that then can be analyzed by quantitative, statistical methods. McCleary and Hay (1980) give the example of a schizophrenic patient from whom a time series of 120 daily perceptual speed scores were collected for many days, the goal ultimately being to test the effect of a medication on these scores. The time-series data thus

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

collected can be used to assess any pattern to these scores and whether the medication precipitated a change in the pattern. While cross-sectional or longitudinal data can be collected on one individual, the result will be more of a qualitative case study, since insufficient data points will be available to conduct a statistical analysis.

B. Uses of Time Series Data

Time-series data are used for many different purposes and focus on a number of issues. For purposes of exposition, the kinds of time-series analyses are sorted into three categories: analyses of trends and forecasting, causal analyses, and program and policy analyses.

1. Trends and Forecasting

Harvey (1981) and McCleary and Hay (1980), note that unlike other forms of statistical analyses that examine behavioral relationships between dependent and independent variables, one form of time-series analysis (univariate time-series analysis) can analyze a variable only in terms of its past. Some of the trend and forecasting activities undertaken with time series involve only the one time-series variable under concern, and an attempt to determine patterns over time in this variable. Patterns may be seasonal, cyclical, monotonic increases or decreases, or ''random walks.'' As was indicated in the discussion of youth unemployment rates, multiple patterns may be observed: seasonal, cyclical, as well as long-term trends.

Once these patterns are determined, it is then possible to forecast the future, by projecting past trends into the future. The accuracy of forecasts obviously depends on the extent to which the past repeats itself in the future. Since history is more than a repetition of the past, analysts often go beyond simple univariate analyses and introduce other structural parameters into their trend and forecasting equations. Nevertheless, even a univariate analysis can be an important first step in understanding patterns in the data.

The kinds of issues that have been explored with time-series data are extensive. We have mentioned time-series analyses that can be conducted on labor market variables, business activity variables, or on an individual's brain patterns. Analyses have been conducted on manufacturing processes, agricultural output, crime rates, and even on sunspots (Harvey, 1981; McCleary and Hay, 1980; Pankratz, 1983; Pandit and Wu, 1983). One of the most remunerative abilities of our time is being able to forecast stock prices, although many fortunes have been lost in this pursuit as well.

As mentioned earlier, an examination of trends is often the first step in an analysis. Most analysts want to be able to understand underlying causal processes as well.

2. Causal Models

Causal models using time-series data have taken many different forms. Clark and Summers (1981) investigated the extent to which different demographic groups are affected by the business cycle. They analyzed this with a model that introduced time as an *independent* variable in the analysis as well as the lagged unemployment rate of prime age males. They found that the employment of youth and minorities is particularly sensitive to cyclical variation in employment. Clark and Freeman (1980) examined how changes in the relative price of labor affects the demand for labor using a model that assumes a lagged response to relative price changes. Their study suggests that adjustment to price changes occur over time, with a smaller change over the short run and a larger change in the long run. They also found that the elasticity found was highly sensitive to assumptions underlying the model used.

Analyses of time-series data have involved a more complex set of issues. For example,

Mishler, Hoskin, and Fitzgerald (1989) examined support for the Conservative and Labour parties over time in England. In spite of the view that the English population has become more conservative, their trend analysis found cyclical swings in attitudes toward both parties. When examining support for the Conservative Party in recent years, they found that it was dictated by concerns about specific issues: unemployment, inflation, and labor strife. They concluded that while the English public was more likely to vote on issues and less likely to vote along class lines, there was no evidence that they were more likely to support the Conservative Party in and of itself.

Probably the most complex time-series model is one involving interdependent equations. In this kind of model one process is hypothesized as influencing another process. The equations for this model form a system of simultaneous equations, which are solved as a system rather than individually (Harvey, 1981).

These causal time-series models have not been without their critics. Issac and Griffin (1989) have criticized them for being ''ahistorical,'' particularly those that have examined labor history over relatively long periods. They argue that many of these studies have assumed continuous, underlying historical processes, using models that do not take into account breaks or discontinuities in history. Few studies, they argue, incorporate periodicity: the idea that there are separable periods in history within which processes may be different. Often researchers begin their analysis by slicing into a particular time in history, without any theoretical justification for choosing that particular time. In fact Issac and Griffin's main critique seems to be that these time-series analyses are a-theoretic, that they are driven by generic statistical models rather than by any theoretical grasp of historical process. In this they are making an important point. While advances in time-series methods and in access to historical data have allowed these empirical forays into history, they do not replace having a solid foundation in theory and historical analysis.

3. Policy and Program Evaluation

Generally, when evaluating a program or policy, researchers have preferred to do so using randomized experiments where a treatment group is exposed to the program or policy and a control group is not (Rossi and Freeman, 1993). When a randomized experiment is not possible researchers will attempt quasi-experimental designs using a matched comparison group and longitudinal data to control for pre-program differences. (This approach will be discussed in greater detail shortly.)

But there are many cases when it is not possible to find a control or comparison group. Some programs or policies provide full coverage: basically everyone eligible for the program is covered by it. For example, The Family Leave Act covers everyone targeted by the act. It is not possible to find an ''untreated'' control group. In some cases there are community-wide initiatives, such as Enterprise Zones or Empowerment Zones, that may affect everyone in a given community. While it may be possible to find a comparison community, to see if the ''treatment'' community benefits from the program relative to the ''comparison'' community, matching communities is very difficult. No one community is exactly like another and there are any number of variables that may precipitate a change, other than the program itself. Disentangling the program effect then becomes very difficult.

Interrupted time-series has been proposed as the appropriate way for evaluating a full-coverage program or community-wide initiatives (Rossi and Freeman, 1993; Connell, Kubisch, Schorr, and Weiss, 1995). The basic premise behind interrupted time series is that the introduction of a program or policy will produce a clear break in the time-series trend for a certain variable or set of variables affected by the program or policy. What is needed for this kind of analysis to work is enough pre-program data to establish a pre-program trend, the exact time

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

of the introduction of the policy or program, and some reasoned assumptions about how long it will take for the policy or program to affect the long term trend.

This method has been used to evaluate a variety of different policies. Rossi and Freeman (1993) describe one study that analyzed the introduction of gun control laws in Massachusetts. A time-series analysis found a statistically significant decrease in armed robberies and assaults following the introduction of the law, although no decrease in the homicide rate. They describe another study to analyze the impact of the introduction of compulsory breathalyzer tests on traffic accidents in England. The time series analysis shows a significant drop in traffic accidents, particularly on weekends.

Davidson and Houston (1981) examine the impact of Richard Nixon's wage and price controls on inflation and wage changes. They found that while the controls were effective in their early stage, inflation increased after 1972, resulting in a small overall effect. Bailey and Peterson (1989) examine the controversial issue of the impact of the death penalty on the homicide rate. An examination of cross-sectional data indicates no relationship; in fact homicide rates were lower in states without the death penalty. A time-series analysis by Stack (1987) found a relationship, however. When Bailey and Peterson revised Stack's model, however, they found a relationship that was too small and too inconsistent to conclude that the death penalty served as a deterrent to homicide.

In terms of evaluating community-wide initiatives, the use of interrupted time series is still in its formative stages. Bloom (1996) proposes a strategy that would use interrupted time series *and* a comparison community. Time series on a variety of variables (e.g. unemployment, welfare receipt) would be developed for both communities. But since there are other things that may result in a change other than a community-based program, such as welfare reform, some kind of change in the time series is expected even without the program. Thus, the change for the treatment community would be compared to the change for the comparison community. If it appears that different processes have been changing the time series for the treatment community, this would be ascribed to the program. Bloom (1996) cautions that for this to work there would have to be relatively large treatment effects to produce statistically significant results.

C. Methodological Issues

The central element of time-series analysis, as mentioned earlier, is that time series are dynamic. Many time-series models may resemble models used in cross-sectional analyses, with the simple addition of a lagged variable. For example, the standard regression model using OLS takes the form:

$$Y_t = bX_t + u_t \tag{1}$$

In this model, X is an exogenous independent variable and u is the disturbance term. It can become a time-series model with the introduction of a lagged independent variable:

$$Y_t = bX_t + bX_{t-1} + u_t \tag{2}$$

Or a lagged version of the dependent variable can be introduced:

$$Y_t = bX_t + aY_{t-1} + u_t \tag{3}$$

Or time (dates) can be introduced as an independent variable:

$$Y_t = bX_t + \mathrm{aT} + u_t \tag{4}$$

Equations 2–4 are dynamic in the sense that time enters into the models in one way or the other. In Equation 2 the dependent variable in time t is hypothesized as being determined

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

by X in time t and X in time $t - 1$, suggesting a continuing impact from the prior period. In Equation 3 the dependent variable in time t is partly determined by the level of the same variable in the prior period, and in part by an exogenous variable X. In Equation 4 the dependent variable is determined by an exogenous variable X, but also is subject to an independent trend as indicated by T. While these are relatively simple models, they can be expanded to include more lagged variables.

In theory all of these Equations 1–4 can be estimated using OLS, if it is assumed that the disturbances are uncorrelated. But this may not be a reasonable assumption with this kind of data. Further, the independent variables may be correlated with one another or with the disturbance term. For these reasons, OLS is generally not recommended for these kinds of models (Harvey, 1991). There are numerous techniques available for working with these kinds of data, depending on what constraints one wants to impose on the coefficients, the assumptions made about the disturbance term, and assumptions made about the lag structure in the model. Since it is not always immediately clear which model is most appropriate for the data, however, choosing a model is often an iterative process, relying on a variety of tests to check the usefulness of different approaches (Harvey, 1981).

As indicated earlier, most of these techniques rely on maximum likelihood estimation. Numerous software packages are available with a variety of techniques for estimating time-series data. The literature on time-series analysis is also booming. An interested reader can find as many as 2, 3, or more articles on this subject in each issue of the *Review of Business and Economics Statistics*.

Other than issues of estimation, issues also arise with the data being used. For example, if a time-dependent process occurs very quickly, but data are only available bi-annually, there is not a technique sophisticated enough to catch an effect. The impression will be given that the process occurs simultaneously. Similarly, missing values in the data can also be problematic if the process occurs quickly.

Unlike longitudinal data, however, attrition is rarely a problem. Since time-series data focuses on one entity, data should be available as long as the country, firm, person is in existence and as long as data are being collected. Some data, such as country-level data, may actually become more reliable over time as data collection processes improve over time.

Finally, the usefulness of time-series analysis ultimately depends on how well models are developed and how results are interpreted. Earlier it was pointed out that with cross-sectional data, it was easy to confuse cohort effects with age effects. With time-series data it is easy to confuse period effects with cohort effects. Taking the example of youth unemployment, we may find that the rise in youth unemployment after a certain period to be the result of some structural changes occurring at that time. But it may also be due to the entrance of a new cohort of youth that is larger (e.g. the baby boom generation) or that has a greater desire for jobs to buy the consumption goods associated with an emerging ''youth culture'' (Steinberg and Greenberger, 1986).

IV. LONGITUDINAL DATA

A. Characteristics

Longitudinal databases, often referred to as panel studies, are relatively new. The first panel studies were conducted in the 1940s, with Lazarsfeld and his associates laying out some of the key issues in using these data at that time (Hagenaars, 1990). But the number of panel studies have proliferated particularly since the 1960s. While most longitudinal databases are of individuals, usually age cohorts, they are not limited to this. The Panel Study of Income Dynamics

(PSID), one of the longest continual longitudinal databases, was originally based on 5000 households and continues to follow both people and households through time. Delacroix and Swaminathan (1991) conduct an analysis on a longitudinal study of wineries.

In many ways, following individuals is easier, however. When the PSID was implemented in 1968, the expectation was that families would be more constant than they have proven to be (Duncan, 1983; Becketti et al., 1988). But in the almost 20 years since the start of the PSID many families have split apart, because of divorce, death or children starting their own homes. New families have formed through marriage and have increased through birth.

Thus, there are new sample members who have entered the data through birth and marriage who have posed a problem in terms of how to weight them in the data (Duncan, 1989). While weights were originally applied to all households in order to make the sample representative of the entire U.S. population, the entry of new members into the sample, often into new households, required a rethinking of the weighting scheme. Newborns were assigned the weight of their household and those who came into the data by way of marriage were assigned the weight of their spouse.

In addition, attrition from the data raised concerns of sample selection bias. Although response rates have been about 96 percent since 1969, attrition accumulates over time, so that 40 percent of those in the original sample are no longer in current waves of the PSID (Duncan, 1989). In spite of these changes, comparisons of the PSID with the Current Population Survey, a nationally representative sample of Americans, indicates relatively few biases (Becketti et al., 1988; Duncan, 1989). Fortunately, the country has changed as much as has the PSID.

In one respect the country has changed more than the PSID, because of immigration. Another drawback of panel studies that hope to be representative, therefore, is that they generally do not have a mechanism for adding new people to the panel, to correspond to the new people being added to the U.S. population. While the changes in family structure found in the PSID were at least consistent with changes in the population, a focus on individuals runs the risk of only being representative of the population at the time the sample is drawn. It may not be representative of the population when the study ends.

This has not come to be a major issue for panel studies, however, for two reasons. First, most longitudinal studies are of short duration. Relatively few last a lifetime within which marked changes can take place. Second, since longitudinal studies are expensive, few of them are large enough to obtain appreciable numbers of immigrants to make an analysis of them feasible. Most of the major panel studies attempt to over-sample blacks and Hispanics, but the total number of persons of immigrant status are probably small. As indicated at the beginning of this paper, small sub-samples are found most easily in large cross-sectional databases.

Nevertheless, the changes in the PSID has raised the issue as to whether it may not be better to make individuals the unit of analysis, rather than families (Duncan, 1983; Becketti et al., 1988). Analyses of families could still be possible but they would be the families of the individuals constituting the core of the sample. By and large, most individuals remain the same person over time, making them a more stable source of study.

B. Uses of Longitudinal Data

Longitudinal data have been used widely throughout the social sciences and in the medical and biological sciences. Although it was difficult to come up with an exhaustive classification of the uses of panel data, studies are divided into three categories: (1) short, two-period studies; (2) life-course studies; and (3) program evaluations.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1. Two-Period Studies

In spite of the advantages of using longitudinal data to look at long-term life changes, there are actually many small studies that just may examine samples over two periods. Many of these studies have fairly limited goals, but deal with topical subjects. Work, job satisfaction and stress is one of the subjects that has received a lot of attention from these relatively small panel studies. For example, Blegen and Mueller (1987) followed a sample of 370 nurses at five hospitals.

They were interviewed twice, eight months apart. Their model examined causal effects in Time 2 for variables measured in Time 1, while controlling for satisfaction in Time 1. They found that nurses were most satisfied who felt they had opportunities for promotion, who were not in over-routinized jobs, who were older, worked a day shift and were neither over nor under worked.

Other studies have examined adjustment to work on the part of newcomers to a job. Fisher (1985) followed a sample of 366 newly graduated nurses in their first full-time jobs, over the course of six months. She found that social support had strong effects in reducing ''unmet-expectations'' stress. Bauer and Green (1994) followed 193 new Ph.D. students entering doctoral programs and found that those who had previous research experience and were more involved in their doctoral programs had less role conflict and were more productive. Fedor, Rensvold, and Adams (1992) in a study of 137 helicopter pilot trainees found that those with a low tolerance for ambiguity were more likely to seek high levels of external feedback, while those with high self-esteem were reluctant to seek feedback.

There have also been numerous studies of the effect of employment on youth. Using a sample of high school students in Orange County, Steinberg and co-workers (1982) and Greenberger and Steinberg (1986) find several negative consequences of teenage employment. They find greater cynicism and a higher tolerance of unethical behavior among working youth. They find that working youth have less time with their peers, less time for extracurricular activity, and less family closeness. Steinberg and Dornbusch (1991) using data from six schools in California and three in Wisconsin also find more psychological and behavioral dysfunction, such as drug and alcohol use and delinquency among students who work—especially those working more than 20 hours per week.

2. Life-Course Studies

While the above analyses are of great interest, it is the life course studies that have exploited the primary benefit of panel data which, is the ability to follow individuals over a long enough period of time to investigate how issues and events early in life may affect later development. Many of these studies have focused on youth development.

Of particular interest in some of the youth-oriented studies is in identifying those young people that are ''at risk'' of poor outcomes in early adulthood. In a review of many of these studies, Dryfoos (1990) found that while definitions of ''risky'' behavior varied somewhat, most analyses found that indicators of being ''at risk'' can be identified relatively early in a child's life, suggesting the need for early intervention. The results from the New York Longitudinal Study (Chess and Thomas, 1984) found that while most children ''at risk'' early in their lives did exhibit behavioral problems later in life, they were also able to identify high risk, ''stress-resistant'' youth that were able to beat the odds. The reasons why they were able to perform well in spite of being at risk was not determined.

Other studies have focused on specific issues of youth development. Jacobsen et al. (1994) in a study of Icelandic children found that secure attachment to parents was associated with higher cognitive performance in childhood and adolescence. Kovacs et al. (1993) examined the

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

precursors to suicide among youth. They rejected the idea that suicidal behavior was ''normal'' for adolescents at a certain age, but was almost always the result of diagnosable psychiatric disturbances.

Adult development has also been a source of longitudinal research. Mentioned earlier was the study by Schaie and Willis (1991) which found large generational differences in adult personality, but little evidence of large, age-related changes in personality. This result was confirmed in Schaie's (1983) edited volume of other studies focusing on adult development. Somewhat comforting is the finding that cognitive decline does not occur much before the age of 75 among healthy adults.

One study that is of particular interest is a meta-analysis of studies of drinking behavior of young adults conducted by Fillmore et al. (1993). This analysis not only examined the effect of individual-level variables on drinking but the effect of cohort-level variables as well, to assess the influence of peers on drinking behavior. They found that while individual characteristics were important, group-level behavior had an affect, particularly on women. This analysis was able to exploit one of the advantages of panel data, which is the ability to examine the individual in the context of his or her cohort.

Many of the studies of using longitudinal analysis focus on event histories. Event-history analysis is a methodological approach (discussed in greater detail shortly), that analyzes shifts between different states; for example, from employment to unemployment, from marriage to divorce. These analyses focus on concrete, measurable events over the course of a life.

Mayer and Carroll (1990) using data from Germany, examine the extent to which job changes result in changes in class status. They found that only a small number of job changes result in changes of class status and that women are much less likely to achieve social mobility through job changes than men. Another German study by Hujer and Schneider (1990) examined the transition from unemployment to employment and found that the tightness of the labor market had a significant effect on the ease with which unemployed people found jobs. They found little evidence that unemployment compensation encouraged people to stay unemployed.

Event-history analysis has been used by several researchers to explore the dynamics of welfare receipt. Bane and Ellwood (1983), Pavetti (1993), and Plotnick (1983) have examined transitions into and out of welfare. They found that most people who enter the welfare caseload exit very quickly (within two years). But that returns to welfare were also fairly common. They also found that most women leave welfare because of a job (rather than because of marriage) and that there were few racial differences in leaving welfare for work.

3. Program Evaluation

Since program evaluations usually follow a treatment and control group through time (at least from pre-program to post-program), longitudinal data are used for most intensive program evaluations. With a randomized design, the main issue is maintaining contact with subjects in the evaluation, particularly the control group who have few incentives to stay in the study since they are not receiving the treatment. Attrition from the control group has been an issue in some evaluations (Rossi and Freeman, 1993) Nevertheless, random assignment of program applicants into treatment and control groups insures that there is no selection bias in the selection of program participants. If people in the program choose to be in the program or are selected by program operators, it is possible that those who are most likely to benefit from the program select into the program or are selected in the program. Thus, any program effect may reflect characteristics of the participants, rather than characteristics of the program itself. Since self-selection may occur on the basis of ''unmeasured'' variables (e.g. motivation), differences between the treatment and control group may be difficult to control for statistically.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

For these reasons, randomized designs are preferred. With a randomized design, it is only necessary to compare mean outcome variables for treatment and control groups members (or mean gains in pre-to-post outcomes) to obtain the effect of a program, since in all other respects, treatment and control group members should be the same (given sufficiently large sample sizes).

But randomized designs are not always feasible. For ethical reasons, some people object to a design that involves denying services to people. Randomized designs are often expensive to undertake. Further, even with random assignment, control group members may sometimes seek services from a similar program that may be available to them, thus contaminating the design.

For these reasons, some have called for better methods in statistically controlling for selection bias, that will allow for comparisons of nonrandomized treatment and comparison groups. For example, comparison groups could be drawn from waiting lists. Or a matched comparison group could be drawn from a secondary database such as the Current Population Survey. Employing a matched comparison group has been controversial, however. Fraker and Maynard (1984) and LaLonde and Maynard (1987) re-analyzed data from an evaluation of the Supported Work Program—an evaluation that was conducted using random assignment—and compared the results from an analysis of the original study to results obtained from a matched comparison group. They found that the results using the matched comparison group did not perform as well or as consistently as results from using a randomly assigned control group.

Heckman et al. (1989) re-analyzed the data and took issue with these results. They developed a model selection process that would allow researchers to choose the best model for adjusting for selection bias. The model selection process would be used on pre-program earnings data for treatment and control groups. The upshot of their analysis is that it is possible to use a matched comparison group in program evaluations and to control for selection bias, but that a longitudinal database with several years of pre-program data was necessary in order to do so. While some have had difficulty in accepting Heckman and Hotz's results and conclusions, they do suggest that longitudinal databases can be very useful in developing matched comparison groups for program evaluation.

3. Methodological Issues

Obviously, selection bias is an important methodological issue for panel data. Assuming that the original sample is a representative sample, selection bias can arise because of biases in the attrition from a longitudinal data base. It can also arise when longitudinal data are used to form matched comparison groups for program evaluations. There is a fairly extensive literature on selection bias with panel data. Solutions range from using fixed-effects estimators (differencing the data over two periods) to using fairly sophisticated models and statistical tests. A review of some of these methods can be found in Heckman and Robb (1985) and Verbeek and Nijman (1992).

There are also pro-active measures that can be undertaken to limit attrition bias. These involve encouraging and maintaining a strong commitment to the survey on the part of survey participants (Chess and Thomas, 1984). Getting subjects to buy into the goals of the project and actually paying for interviews are among methods that have been used.

In other ways, longitudinal data are easier to work with than other data, or at least they have a great deal of versatility. Cross-sectional studies using OLS can be undertaken with longitudinal data if one only wants to focus on data in one time period. Studies using lagged values can be undertaken, using methods similar to those used with time-series data. Or one can undertake event-history analysis, which was designed specifically for use with longitudinal data.

The earliest versions of event-history analysis can be found in the life tables used by demographers (Allison, 1984). Bio-medical researchers have also been traditionally interested

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

in survival rates, while engineers focused on failure time for machinery. In the 1970's event-history analysis increasingly came to be used by social scientists as a way of exploring ''social dynamics'' (Tuma and Hannan, 1984).

One of the central concepts in event-history analysis is the ''hazard rate,'' which measures the probability that a certain event will occur to an individual, given that the individual is at risk. Risk is defined very broadly; for example, unmarried adults are at risk of getting married, while married people are at risk of getting a divorce. The construction of an event-history not only involves working with discrete and measurable ''events,'' but being able to discretely define the group that is at risk; for example, a twelve-year-old may not be at risk of marriage, but a fourteen-year-old may well be.

The hazard rate in period one is the number of people who have an event (e.g. get married) as a percentage of those at risk of getting married (e.g. those over 14 years old). In period two the hazard rate is the number of people who get married as a percentage of those at risk. But in period two those who got married in period one are subtracted from those who were at risk in period one, thus changing the denominator for calculating the hazard rate in period two.

Many event history analyses use the hazard rate for their dependent variables. For example, the study by Mayer and Carroll (1990), discussed earlier, examined the hazard rate for achieving upward mobility with a job change. Some researchers may focus on the survival rate—rather than the hazard rate—which is the probability of not having an event prior to time t (Yamaguchi, 1991). Various researchers have also focused on repeating events, multiple events, and competing events, increasing the complexity of the analyses undertaken (Alison, 1984).

Either way what is essentially being used as a dependent variable is a probability. Thus, the same issues arise for a hazard rate or a survival rate as with a qualitative dependent variable: the dependent variable is constrained to take values in between 0 and 1, while the independent variables can be any real number. The problems involved with this can be solved by taking a logit transformation of the dependent variable and estimating with maximum likelihood techniques (Alison, 1984).

There are several statistical programs that are available for event-history analysis, but as always it is necessary to specify the type of model one wants to work with. There are a variety of approaches to event-history analysis. First, there are discrete time methods and continuous time methods, both with their advantages and disadvantages (Alison, 1984). Also the shape of the hazard function can take many different forms; for example, the hazard of being arrested for a crime decreases with age, while the hazard of retiring increases with age. The hazard rate for dying is U-shaped: it is high for the very young and for the very old (Alison, 1984). Thus, the distribution of the hazard is an important consideration in choosing a model. One model, Cox's proportional odds model is very popular, since it is not necessary to specify the effects of time. It can also be used to construct stratified models for categorical covariates (e.g. stratified by race or sex) (Yamaguchi, 1991). But the proportional-odds model is not appropriate for all situations.

In using event-history analyses, researchers confront two other major issues: how to handle sample censoring and how to handle time-varying independent variables. Sample censoring is very difficult to avoid. Taking again the example of marriage, unless the data cover an entire cohort from the time they are born to the time they all die, there will be some marriages or divorces that will occur outside of the time frame of the study. If the study begins when a cohort is 20 years old, there will be some people who are already married (left censored). If the study ends when the cohort is 40 years old there will be many people who are still married (right-censored). Thus, most studies will encounter people in the middle of an interval between events. Since the researchers may not know the starting and ending points for these intervals, some adjustments must be made for them.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Left-censoring—starting the study with people already in the middle of an interval—is generally considered harder to adjust for (Alison, 1984). If right censoring occurs for a variety of reasons (death, migration out of the risk sample, attrition from the sample), and seems to be independent of the events under question, then it is considered random. If right-censoring is associated with an event—for example, those who are divorced are more likely to drop out of the study—then right censoring is more problematic (Alison, 1984).

Time-varying independent variables are those that can change in the course of the study (e.g. earnings or health status). If the time-varying independent variable is measured at different intervals than the dependent variable, it is difficult to capture their effect. Discrete-time methods allow for time-varying independent variables since a separate observation is created for each time unit for each person at risk. A value for the independent variable can be assigned to each person-time unit. Data using discrete time can become very large and unwieldy, however.

In spite of the difficulties with event-history analysis it has proven to be a very useful vehicle for analyzing events that occur over time. One of the great advantages of panel data, is its amenability to this form of analysis.

Nevertheless, like cross-sectional analysis and time-series analysis, longitudinal analysis also has its flaws in dealing with time. As noted earlier, in cross-sectional analysis it is possible to confuse cohort effects as age effects and with time-series analysis it is possible to confuse cohort effects as period effects. With longitudinal data it is possible to confuse period effects as age effects (Hagenaars, 1990). Since the focus is overwhelmingly on the cohort, it is possible to miss the effects of events that are unique to that cohort. For example, the ''baby boom'' cohort was deeply affected by the Civil Rights Movement and the Vietnam War. While youth rebellion is common to all cohorts, that for this generation was exacerbated by these period-specific events. Thus, it would be inappropriate to examine the rebellion of a cohort in the baby-boom generation without paying close attention to the events specific to the period in which they achieved maturity.

V. CONCLUSION

Fortunately, there are many researchers examining issues using many different kinds of data. Thus, we have more than one perspective on any given subject. Nevertheless, there are relatively few deliberate attempts to examine issues with different databases, sorting through age, cohort, and period issues, for example. As new approaches have been developed, people have embraced them and focused attention on new methodological techniques. Perhaps the future will show more interest in multi-level research programs.

BIBLIOGRAPHY

Aldrich, J.A. and F.D. Nelson (1989). *Linear Probability, Logit, and Probit Models*, Sage Publications, Newbury Park.

Allison, P.D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*, Sage Publications, Newbury Park.

Bailey, W.C. and R.D. Peterson (1989) ''Murder and Capital Punishment: An Monthly Time-series Analysis of Execution Publicity,'' *American Sociological Review*, *54*: 722.

Bane, M.J. and D.T. Ellwood (1983). *The Dynamics of Dependence: The Routes to Self-Sufficiency*, Urban Systems Research and Engineering, Cambridge, MA.

Bauer, T.N. and S.G. Green (1994). ''Effect of Newcomer Involvement in Work-related Activities: a Longitudinal Study of Socialization,'' *Journal of Applied Psychology*, *79*: 211.
Becketti, S., W. Gould, L. Lillard, and F. Welch (1988). ''The Panel Study of Income Dynamics After Fourteen Years: an Evaluation,'' *Journal of Labor Economics*, *6*: 472.
Blegen, M. and C.W. Mueller (1987). ''Nurses Job Satisfaction: a Longitudinal Analysis,'' *Research in Nursing and Health*, *10*: 227.
Bloom, H.S. (1996). ''Building a Convincing Test of a Public Housing Employment Program Using Nonexperimental Methods: Planning for the Jobs-Plus Demonstration,'' Presented at the Association for Public Policy and Management Annual Meetings, Pittsburgh, November. 1996.
Burbridge, L.C. (1994). *Government, For-Profit and Third Sector Employment: Differences by Race and Sex, 1950–1990*, Center for Research on Women, Wellesley College, Wellesley, MA.
Chess, S. and A. Thomas, (1984). *Origins and Evolution of Behavior Disorders From Infancy to Early Adult Life*, Brunner/Mazel Publishers, New York.
Clark, K.B. and R.B. Freeman (1980) ''How Elastic is the Demand for Labor?'' *Review of Economics and Statistics*, 509.
Clark, K.B. and L.H. Summers (1981) ''Demographic Differences in Cyclical Employment Variation,'' *Journal of Human Resources*, *16*: 61.
Connell, J.P., A.C. Kubisch, L.B. Schorr, and C.H. Weiss (1995.) *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, The Aspen Institute, Washington, D.C.
Davidson, L.S. and D. Houston (1981) ''A Reexamination of the Nixon Wage-price Controls: an Application of Time Series Methods, *Journal of Economics and Business*, *33*: 246.
Delacroix, J. and A. Swaminathan (1991). ''Cosmetic, Speculative and Adaptive Organizational Change in the Wine Industry: a Longitudinal Study, *Administrative Science Quarterly*, *36*: 631.
Dryfoos, J.G. (1990). *Adolescents At Risk*, Oxford University Press, New York.
Duncan, G.J. (1983). ''The Implications of Changing Family Composition for the Dynamic Analysis of Family Economic Well-being,'' in *Panel Data on Incomes*, A.B. Atkinson and F.A. Cowell, (eds.), London School of Economics, London.
Duncan, G.J. and D.H. Hill (1989). ''Assessing the Quality of Household Panel Data: the Case of the Panel Study of Income Dynamics,'' *Journal of Business and Economic Statistics*, *7*: 441.
Farley, R. and W.R. Allen (1989). *The Color Line and the Quality of Life in America*, Oxford University Press, New York.
Fedor, D.B., R.B. Rensvold, and S.M. Adams (1992). ''An Investigation of Factors Expected to Affect Feedback Seeking: a Longitudinal Field Study,'' *Personnel Psychology*, *45*: 779.
Fillmore, K.M., B.M. Johnstone, E.V. Leino, and C.R. Ager (1993). ''A Cross-study Contextual Analysis of Effects from Individual-level Drinking and Group-level Drinking Factors: a Meta-analysis of Multiple Longitudinal Studies from the Collaborative Alcohol-related Longitudinal Project,'' *Journal of Studies on Alcohol*, *54*: 37.
Fisher, C.D. (1985). ''Social Support and Adjustment to Work: a Longitudinal Study,'' *Journal of Management*, *11*: 39.
Fraker, T. and R. Maynard (1984). ''An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs,'' Mathematica Policy Research, Princeton, NJ.
Greenberger, E. and L. Steinberg (1986). *When Teenagers Work: the Psychological and Social Costs of Adolescent Employment*, Basic Books, New York.
Hagenaars, J.A. (1990). *Categorical Longitudinal Data: Log-Linear, Panel, Trend, and Cohort Analysis*, Sage Publications, New York.
Harvey, A.C. (1981). *The Econometric Analysis of Time Series*, John Wiley and Sons, New York.
Heckman, J.J. and R. Robb (1985). ''Alternative Identifying Assumptions in Econometric Models of Selection Bias,'' in *Longitudinal Analysis of Labor Market Data*, J.J. Heckman and B. Singer, (eds.), Cambridge University Press, Cambridge.
Heckman, J.J., V.J. Hotz, and M. Dabos (1987). ''Do we Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?,'' *Evaluation Review*, *11*: 395.
Heckman, J.J. and V.J. Hotz (1989). ''Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: the Case of Manpower Training,'' *Journal of the American Statistical Association*, *84*: 862.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Hujer, R. and H. Schneider (1990). ''Unemployment Duration as a Function of Individual Characteristics and Economic Trends,'' in *Event History Analysis in Life Course Research*, K.U. Mayer and N.B. Tuma, (eds.), University of Wisconsin Press, Madison.

Issac, L.W. and L.J. Griffin (1989). ''Ahistoricism in Time-series Analyses of Historical Process: Critique, Redirection, and Illustrations from U.S. Labor History,'' *American Sociological Review*, *54*: 873.

Jacobsen, T., W. Edelstein, and V. Hofmann (1991). ''A Longitudinal Study of the Relation Between Representations of Attachment in Childhood and Cognitive Functioning in Childhood and Adolescence, *Developmental Psychology*, *30*: 112.

Kovacs, M., D. Goldston, and C. Gatsonis, (1993). ''Suicidal Behaviors and Childhood-onset Depressive Disorders: a Longitudinal Investigation, *J Am. Acad. Child Adolesc. Psychiatry*, *32*: 8.

LaLonde, R. and R. Maynard (1987). ''How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment,'' *Evaluation Review*, *11*: 428.

Markus, G.B. (1979). *Analyzing Panel Data*, Sage Publications, Newbury Park.

Mayer, K.U. and G.R. Carroll (1990). ''Jobs and Classes: Structural Constraints on Career Mobility,'' *Event History Analysis in Life Course Research*, K.U. Mayer and N.B. Tuma, (eds.), University of Wisconsin Press, Madison.

McCleary, R. and R.A. Hay, Jr. (1982). *Applied Time Series Analysis for the Social Sciences*, Sage Publications, Newbury Park.

Mishler, W., M. Hoskin, and R. Fitzgerald (1989). ''British Parties in the Balance: a Time-series Analysis of Long-term Trends in Labour and Conservative Support, *British Journal of Political Science*, *19*: 211.

Mullis, I. (1993). *NAEP 1992 Reading Report Card for the Nation and the States: Data from the National and Trial State Assessments*, Government Printing Office, Washington, D.C.

Pandit, S.M. and S.M. Wu (1983). *Time Series and System Analysis With Applications*, John Wiley and Sons, New York.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, John Wiley and Sons, New York.

Pavetti, L. (1993). *Dynamics of Welfare and Work: Exploring the Process by which Young Women Work their Way off Welfare*, PhD. Dissertation, John F. Kennedy School of Government, Harvard University, Cambridge.

Plotnick R. (1983). ''Turnover in the AFDC Population: an Event-history Analysis,'' *Journal of Human Resources*, *18*: 65.

Rossi, P.H. and H.E. Freeman (1993). *Evaluation: A Systematic Approach*, Sage Publications, Newbury Park.

Sandefur G.D. and M. Tienda (1988). ''Introduction: Social Policy and the Minority Experience,'' in *Divided Opportunities: Minorities, Poverty and Social Policy*, G.D. Sandefur and M. Tienda, (eds.), Plenum Press, New York.

Schaie K.W. and S.L. Willis (1991). ''Adult Personality and Psychomotor Performance: Cross-sectional and Longitudinal Analysis, *Journal of Gerontology*, *46*: 275.

Schaie, K.W. (ed.) (1983). *Longitudinal Studies of Adult Psychological Development*, The Guilford Press, New York.

Sokoloff, N.J. (1982). *Black Women and White Women in the Professions*, Routledge, New York.

Steinberg L. and S.M. Dornbsch (1991). ''Negative Correlates of Part-time Employment During Adolescence: Replication and Elaboration,'' *Developmental Psychology*, *27*: 304.

Steinberg, L., E. Greenberger, L. Garduque, M. Ruggiero, and A. Vaux, (1982). ''Effects of Working on Adolescent Development,'' *Developmental Psychology*, *18*: 385.

Tuma N.B. and M.T. Hannan (1984). *Social Dynamics*, Academic Press, New York.

United Nations Development Program (1995). *Human Development Report, 1995*, Oxford University Press, New York.

United States Bureau of Labor Statistics (1996). *Employment and Earnings*, U.S. Department of Labor, Washington, D.C.

United States Bureau of Economic Analysis (1995). *Survey of Current Business*, U.S. Department of Commerce, Washington, D.C.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

United States Commission on Civil Rights (1992). *Civil Rights Issues Facing Asian Americans in the 1990s*, U.S. Commission on Civil Rights, Washington, D.C.

United States Federal Bureau of Investigation (1993). *Crimes in the United States: Uniform Crime Report*, U.S. Department of Justice, Washington, D.C.

Verbeek M. and T. Nijman (1992). Testing for selectivity bias in panel data models, *International Economic Review*, *33*: 681.

Vroman, W. (1987). *The Economic Performance of Minorities in National and Urban Labor Markets*, The Urban Institute, Washington, D.C.

Yamaguchi, K. (1991). *Event History Analysis*, Sage Publications, Newbury Park.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

15
Forecasting Methods for Serial Data

Daniel W. Williams
Baruch College, New York, New York

I. INTRODUCTION

An analyst may want to know the value of future members of a data series, for example, if a public policy is created to reduce the number of teen pregnancies, the analyst may want to know how many pregnancies would occur in future years in the absence of this policy. This chapter discusses techniques for forecasting serial data, that is numeric data arranged in a meaningful order or series. Serial data follow in a particular order because some process generates them in this order. Commonly, these data are arranged in chronological order, that is some process generates data across time and the data are recorded beside a time index (such as the dates on which the data are observed). In this chapter, data of this sort are called ''serial data,'' ''time serial data,'' or ''time series.'' The techniques discussed in this chapter are designed for forecasting numeric time series.

A. Simple Techniques

The techniques discussed here are relatively simple to understand and use. While there are more complex techniques, they are frequently constrained by severe criteria that are difficult to meet. For example, to use a regression based time series technique, it is necessary to know the future values of the predictor variables (the x variables in the equation, $\hat{Y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$: Equation 1); often with real world forecasting problems predictor variables are no better known than the series that is to be forecast (Vollmann, Berry, and Whybark, 1993). Research over the past twenty years shows that simpler techniques, such as the ones discussed in this chapter, often provide forecasts that are as accurate as those provided by more complex techniques (R. Ashley, 1983; Ashley, 1988; Makridakis et al., 1982). The main constraint for use of techniques discussed here is that there should be no known reason to expect the data generating process to change significantly over the horizon (the future periods) that is to be forecast.

B. Notation

Forecast literature contains some standard, or nearly standard, notation and some unsettled notation. Almost universally, time varying data are indexed with a time period subscript. Usually, this subscript is symbolized with a lower case t, thus X_t refers to the observation X at time t. There is relatively general agreement that F_t refers to a forecasted value at time period t; X_t

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

refers to the actual data at time period t; and e_t (error or deviation at time t) refers to X_t minus F_t. Order is important. In this chapter e_t refers to $X_t - F_t$ not $F_t - X_t$. Time change is usually shown with such notation as $t - 1$, $t + m$ or $t - L$. These and similar notation refer to relative differences in time. Thus, $t - L$ means relative to an observation at period t look for the observation L time units earlier (L is capitalized to avoid confusion with the number one). It is not hard to find articles that violate any or all of these conventions; however, they will be used here. Other terms are defined as they are introduced.

The symbol $\sum_{i=x}^{\nu}$, is read as ''sum from x to ν;'' that is, add up the items indexed by i beginning with x and ending with ν. This symbol does not stand alone. To follow the instruction there must be a term following it. For example, $\sum_{t=1}^{n} (X_t - F_t)$ instructs the reader to sum a column of numbers each of which is computed by subtracting F_t from X_t. Further, the numbers to include start with the first observation and end with the observation indicated by the superscript n (which usually means the last observation).

Greek characters, such as α, β, ϕ, γ, and others refer to ''parameters.'' Parameters are values that summarize a time series generating process, and are estimated through observation of a sample. Greek letters may also refer to values that function like parameters in producing forecasts, but are not estimated from the data. When forecasting, a sample is found by observing a segment of a data series. Since a portion of the series is in the future, it is unobservable, so, there are only samples; the universe is beyond observation. Again, the forecast literature does not reflect standard usage, so it is easy to find articles making conflicting use of these terms.

II. PREPARING TO FORECAST

Forecasting begins with data. Usually data must be collected and prepared before it is forecast.

A. Time Intervals

Usually data accumulate over a time interval. If an analyst records the number of cars passing through an intersection each day, the cars do not all pass through the intersection at once. The number accumulates all day. The particular time unit used to refer to those data may be the beginning of that time interval, the end of that time interval, the mid-point of the time interval, a time unit that refers to the whole time interval, or some other rationally selected time unit. No particular reference method is ''right''; however, the analyst should know the applications, advantages, and limitations of the method used. Analysts planning to collect data may want to use either the mid-point of the time interval or a unit that refers to the whole time interval to limit confusion about the relationship between the data and the index. However, practical reasons may lead the analyst to prefer a different reference. For example, some Medicaid data accumulates over weeks and checks are dated for a day in the next week. For some forecasting purposes the data are cumulated over months and recorded beside the month and year, a time unit that refers to the whole period. For other purposes, the date of the check is of interest, so the data are recorded beside that date, a time unit that is entirely outside the time period over which the data cumulated. In each case, the index is meaningful considering how the data are used. Here are two more considerations for cumulating data:

1. Use equal or nearly equal time intervals. It would be best to always use equal time intervals, but business practices and the calendar make this hard. Months are not equal in size, and although weeks have the same number of days, they may not have the same number of business days, as one may contain a holiday. Select a standard time

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

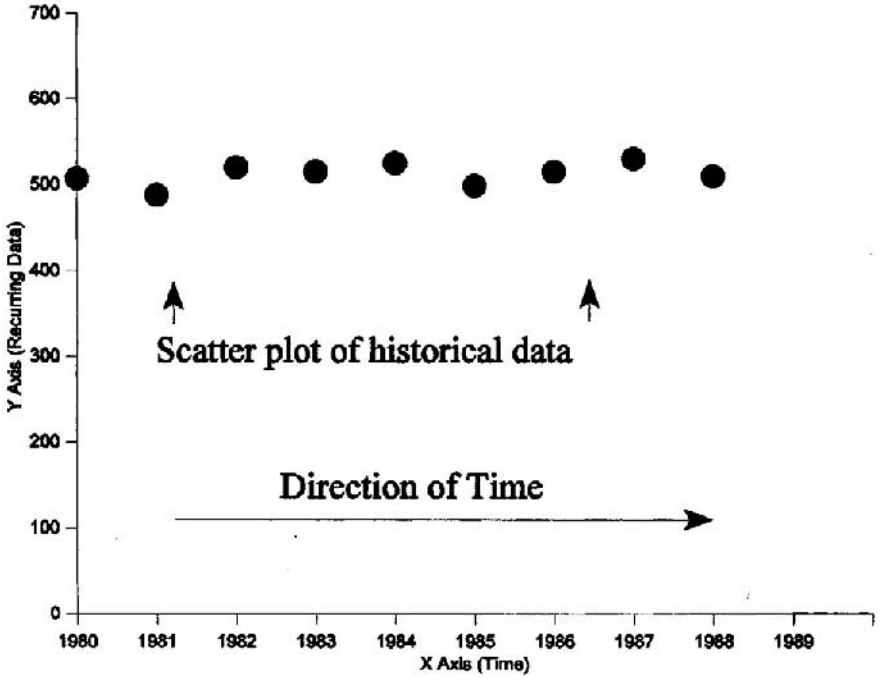


FIGURE 1 Example *XY* graph of time series.

interval (years, quarters, months, weeks, days, hours, minutes, seconds, etc.) and stick with it. Choose the smallest unit of interest. It is far easier to combine data that are too detailed (e.g., add days together to get weeks) than to break aggregated data apart (e.g., find data for days from weekly data).

2. Be consistent. If Fridays' data are held over to the next week, they should always be held over to the next week.

B. Graphing Historical Data

Graph the data to look for recognizable patterns and anomalies. In general practice, forecasters graph data in an *XY* scatter plot or line plot using the *Y* axis as the values of the observed data and the *X* axis as the time index. Older data are to the left and more current data are to the right as shown in Figure 1.

When forecasters show the forecast on the same graph, they often demonstrate the break between historical data and future data by drawing a vertical line at the end of the historical data as shown in Figure 2.

C. Data Editing

Once the data are arranged in a serial order, make sure that they do not contain mistaken entries. After data entry, graph the data and look for outliers, that is values that are unusually large or small. In the Figure 3, the observation marked by a triangle is an outlier.

The most likely explanation of an outlier will be incorrect data entry. Two common sources of data entry error are reversal of numbers (for example, entering 36 as 63) and decimal error

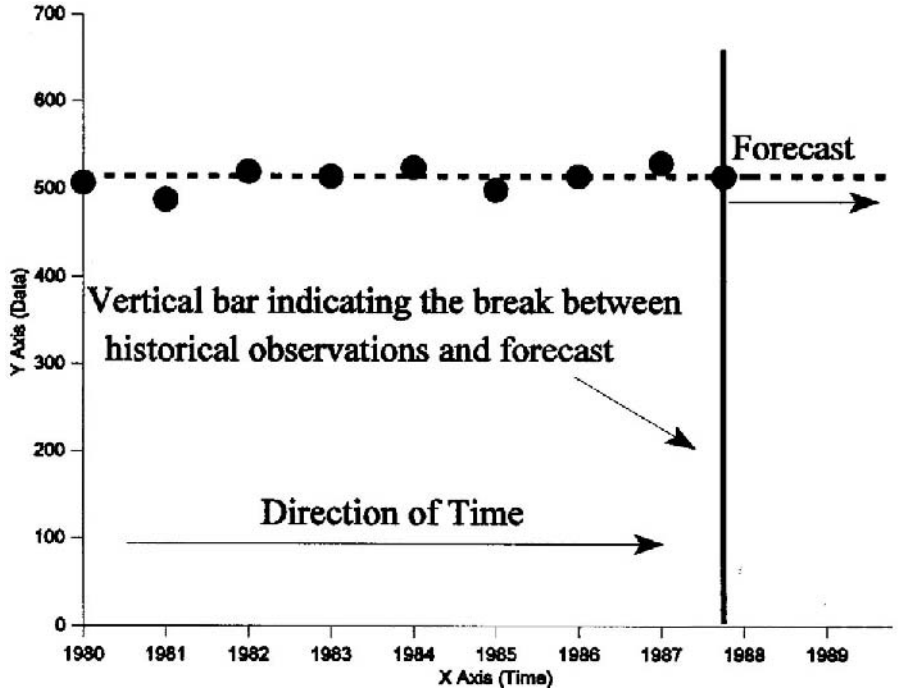


FIGURE 2 Example graph with forecast.

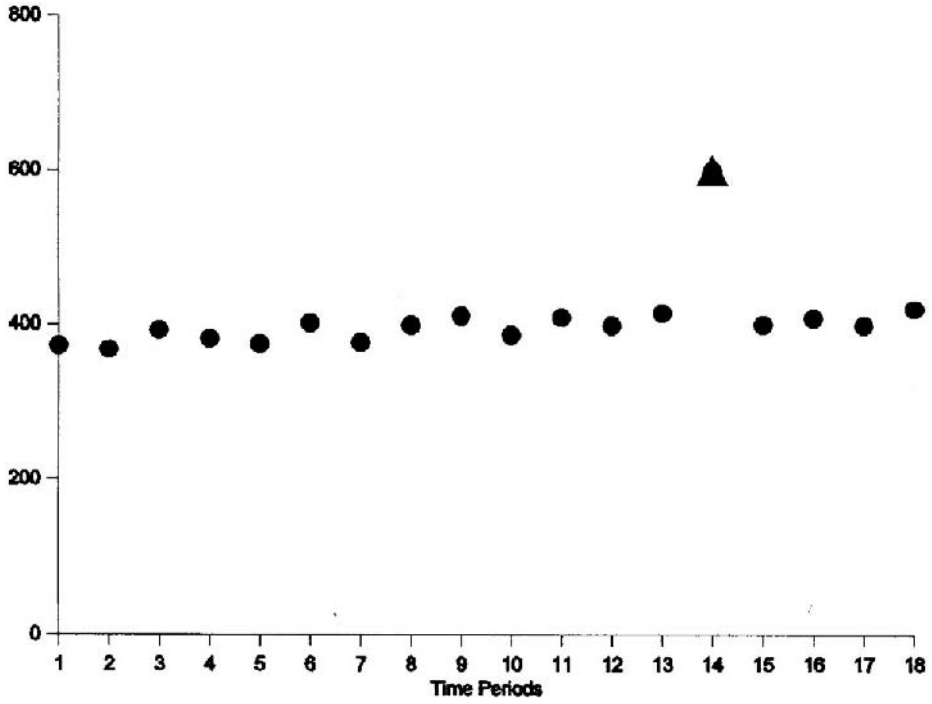


FIGURE 3 Example outlier.

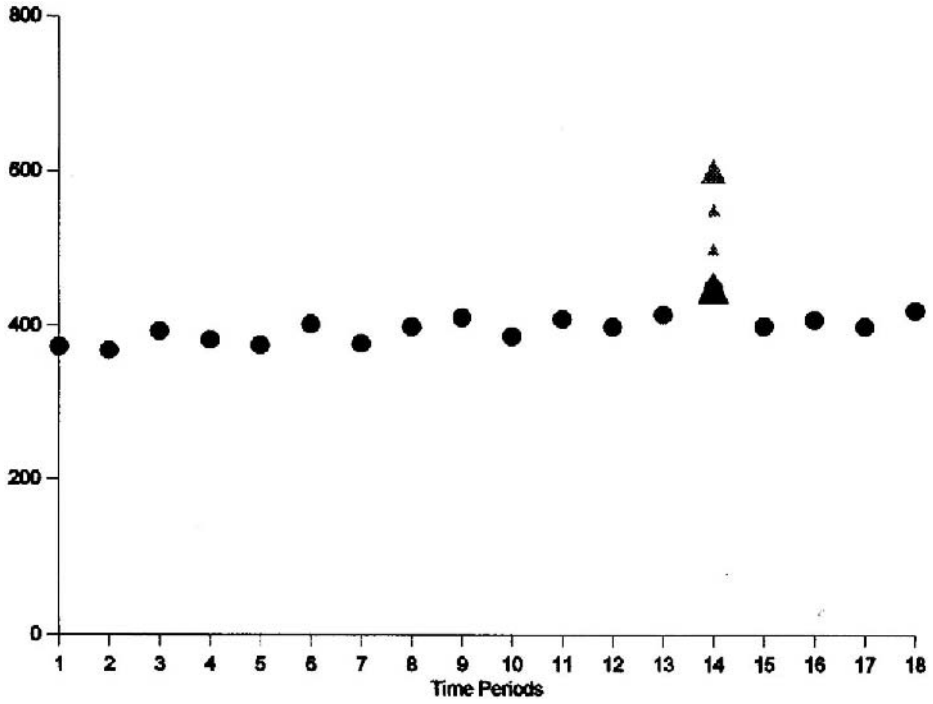


FIGURE 4 Reducing an outlier value.

(for example, entering 36.03 as 360.3). Also, when entering several columns of data, the analyst may copy from the wrong column. When an analyst finds an outlier, he or she should look for a data entry error. However, sometimes the analyst does not have the original data with which to compare, or does not find an error. Does the outlier mean:

1. The data are subject to occasional large disturbances? If so, leave the data as it is.
2. There is an undiscoverable recording error? (Make this decision only if the observation is impossible or nearly so, otherwise see option three.) If so, correct the error with the best information available, by substituting a corrected observation for the erroneous one. There are several candidates for corrected observations. First, the analyst might calculate the average of the two surrounding observations. Second, if the data are seasonal (discussed below), the analyst might calculate the average of the two nearest observations that occur at the same point in the seasonal cycle. For example, if the data are subject to annual seasonality, calculate the average of the observations from the previous year and the next year. These two approaches may also help solve the problem of missing data.
3. In the past, there was an unusual disturbance, but it is unlikely to recur? Or, is it probably a recording error, but it might not be? In this case the analyst may choose to leave the data alone, or he or she might choose to windsorize the observation (Armstrong, 1985). This practice consists of reducing the outlier to the most extreme value that is likely to occur. For example, if nearby observations take values from 360 to 420, and the extreme observation is 600, the analyst might choose to reduce the extreme observation to 450. This adjustment is shown in Figure 4. Be very cautious with the use of windsorizing. Do not repeatedly windsorize the same series.

TABLE 1 Windsorizing Data

Period	Original X	X'	$E = X$-Average	E Squared	Revised X
1	373	373	−22.1	489.2	373
2	368	368	−27.1	735.4	368
3	393	393	−2.1	4.5	393
4	382	382	−13.1	172.1	382
5	375	375	−20.1	404.7	375
6	402	402	6.9	47.4	402
7	377	377	−18.1	328.2	377
8	399	399	3.9	15.1	399
9	411	411	15.9	252.2	411
10	387	387	−8.1	65.9	387
11	409	409	13.9	192.7	409
12	399	399	3.9	15.1	399
13	415	415	19.9	395.3	415
14	600				442.2
15	400	400	4.9	23.8	400
16	408	408	12.9	166.0	408
17	399	399	3.9	15.1	399
18	420	420	24.9	619.1	420
Total*		6717	SSQ	3941.8	
Count		17	DF	16	
$\bar{x}'$		395.1	VAR	246.4	
			σ'	15.7	
			SD * 3	739.1	
			$\bar{x}' + 3\sigma'$	442.2	

* For this and later tables additions and subtractions may not be exact due to rounding.

Suppose that the analyst is not sure what the most extreme likely value is and the data are not following a particularly large trend. Then, an option is to calculate the standard deviation (excluding the outlier) of the immediately surrounding data and place the observation at three standard deviations from the average of those data in the direction of the outlier. If the resulting observation is more extreme than the original outlier, the original value should be retained. This technique will not work, however, with rapidly trending data, or data that are extremely seasonal. In those cases, the data may be windsorized using the judgmental estimate of the most extreme likely value.

In Table 1, observation 14 is windsorized by calculating the average plus three standard deviations using the equation, $O' = \bar{x}' + 3 * \sigma'$: Equation 2), where O' is the windsorized observation, and $\bar{x}'$ and σ' are the mean and standard deviation of the series excluding the extreme observation.

D. Patterns in Data

Good analysts evaluate their data for patterns that show systematic variation, which can be used to simplify the data.

1. Variation along the Time Index

Figures 5 and 6 demonstrate typical patterns that can be found in data. What gives rise to these patterns is that the phenomena measured is strongly related to recording periods. In Figure 5,

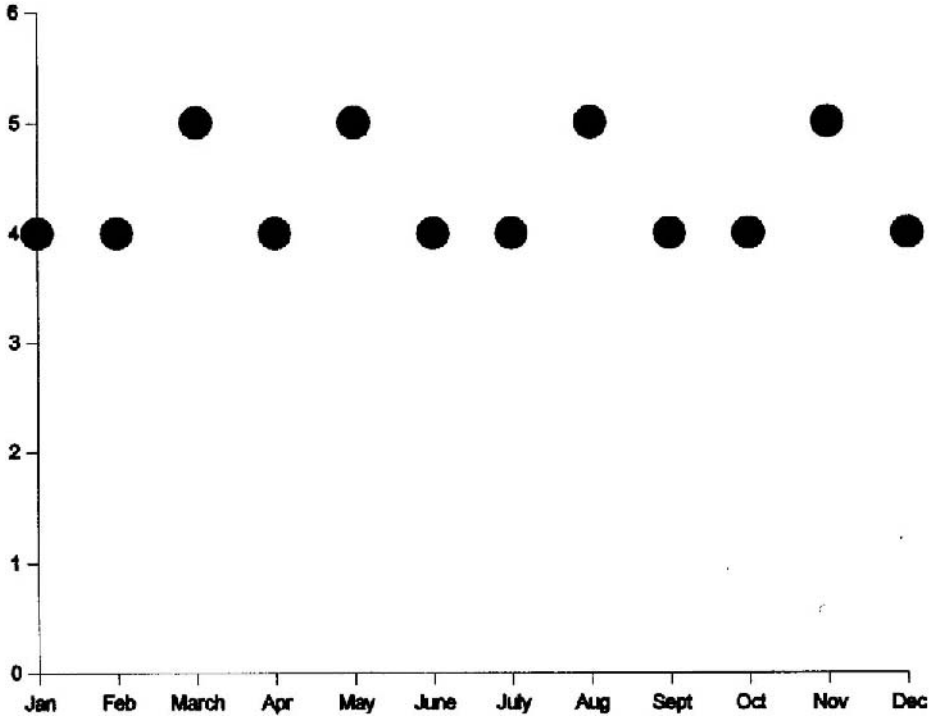


FIGURE 5 Business recording days.

data accumulate over the whole recording period. In longer periods more data accumulate. For example, during the late 1980s and early 1990s, nursing facilities billed Virginia Medicaid for the number of days of service delivered to their patients over monthly billing periods. Longer months contained more days of service than shorter months.

In Figure 6, data accumulate over a week but is recorded on one specific day of the week. With the Virginia Medicaid program this happens when health care providers bill within a few days of delivering the service. The program receives bills all week long every week and takes action on the bills on Fridays. If the forecaster accumulates the data to months, there will be a natural fluctuation because some months contain four Fridays and others contain five.

In preparing to forecast data that exhibit such patterns, forecasters should first account for this completely predictable variation. For the forecasting techniques discussed here, the best way to account for this predictable phenomena is to normalize it, that is divide the data by the factor that causes it to fluctuate. For example, when forecasting the number of days of care delivered in a nursing home from monthly billing data, first divide the data by the number of days in the month over which the data cumulated. Forecast the normalized data series. To complete the forecast, reverse the normalization, that is multiply the forecast by the future month normalizing factor.

There are many other possibilities. For example, if employees are paid weekly and the pay day happens to be the first day of the year, it will also be the last day of the year, that year will contain 53 pay checks, which has the effect of increasing the payroll cost by roughly 2%. By graphing data, analysts can discover patterns, and by examining the process that generates the data, they can determine what a pattern means.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

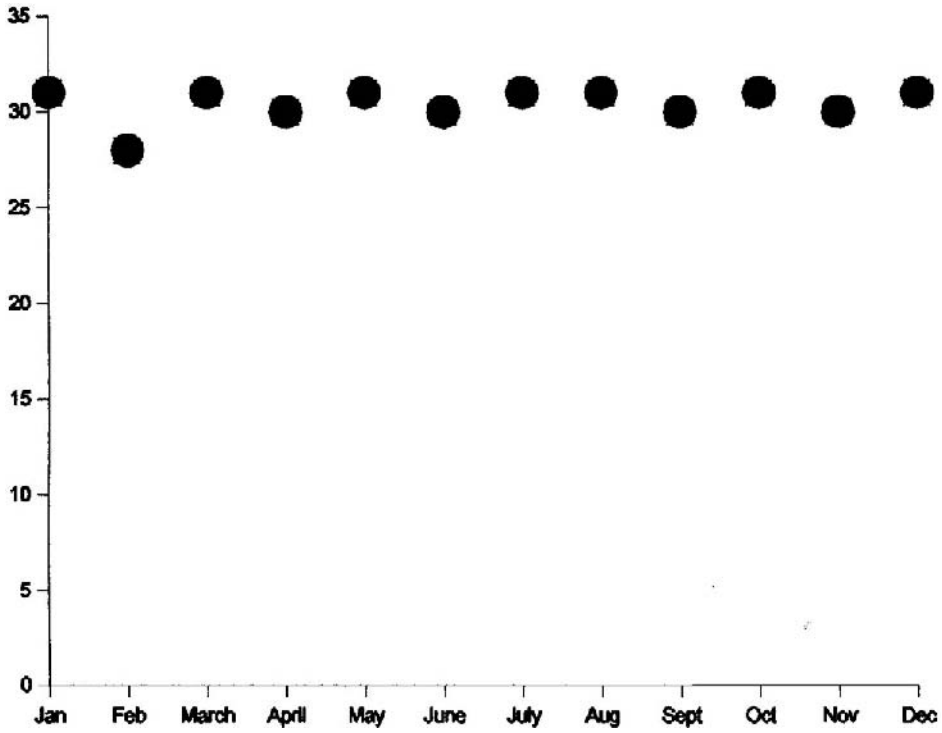


FIGURE 6 Days in the month.

2. Other Complex Phenomena

The procedure of taking sources of variation, such as the days-of-the-month or recording-day-of-the-week, into account is sometimes called decomposition. With decomposition, a complex data series is broken into several component series (Armstrong, 1985). The simpler data series should be easier to forecast, and where relevant, different methods can be used to forecast different component series. In the examples from the last section, the systematic variation (days-of-the-month, etc.) can be known without error, so forecasting it can only introduce error. Once this variation is removed from the data series, the task of the mathematical forecasting model is simpler.

While graphing the data reveals patterns that arise across the time index, it may not help with other complexities. Consider the teen pregnancy issue at the beginning of this chapter. Two components of this series are the number of female teens and the rate at which they become pregnant. It is ineffective to confuse these issues. Predicting the number of teens over the next few years may be relatively easy, since they are already around as pre-teens (assuming no important net migration issues). The forecasting challenge involves pregnancy rate. The best way to find these components is to examine the process that generates the serial data of interest.

Often data can be simplified through such adjustments. Sometimes these adjustments eliminate the need to forecast some of the variation (as with days-in-the-month variation). Other times component forecasts can be obtained from outside sources. Yet other times the main gain through decomposition is the ability to forecast more meaningful homogenous data series.

TABLE 2 Constant Dollars

	Finding the base			Conversion to constant dollars		
Year	Rev R	Tax rate T	Nom $ ND R ÷ T	CPI (82–84) DF source BLS	DF_{1995} ÷ DF Factor DF_1/DF_1	Const $ CD Factor * ND
1980	725	0.050	14,500	82.4	1.850	26,818
1981	750	0.050	15,000	90.9	1.677	25,149
1982	750	0.050	15,000	96.5	1.579	23,689
1983	763	0.050	15,250	99.6	1.530	23,334
1984	775	0.050	15,500	103.9	1.467	22,735
1985	866	0.055	15,750	107.6	1.416	22,308
1986	880	0.055	16,000	109.6	1.391	22,248
1987	908	0.055	16,500	113.6	1.342	22,136
1988	935	0.055	17,000	118.3	1.288	21,900
1989	963	0.055	17,500	124.0	1.229	21,508
1990	1080	0.060	18,000	130.7	1.166	20,989
1991	1110	0.060	18,500	136.2	1.119	20,700
1992	1140	0.060	19,000	140.3	1.086	20,639
1993	1284	0.065	19,750	144.5	1.055	20,830
1994	1300	0.065	20,000	148.2	1.028	20,567
1995	1365	0.065	21,000	152.4	1.000	21,000

3. Constant Dollars

An important form of decomposition for public decision making is the removal of inflation from revenues and expenditures (Ammons, 1991). The impact of inflation can be estimated from indexes known as deflators which, in the United States, are available from the Bureau of Labor Statistics of the Department of Commerce. There are many deflators depending on the sorts of things a government agency usually purchases. Analysts must choose a deflator that relates to the data forecasted. To apply the deflator, use Equation 3:

$$CD_t = ND_t * DF_b/DF_t \tag{3}$$

Where,

ND_t = Nominal dollars, funds expressed in dollars before adjusting for inflation, in year t
CD_t = Constant dollars, funds expressed in dollars after adjusting for inflation, in year t
DF_t = Deflator index for year t
DF_b = Deflator index value for a chosen constant year b

As an example, the analyst may be interested in forecasting sales tax revenue. First, sort out the components of this revenue. If there are no data on total sales within the locality, reason backwards from taxes received to tax base. If there is a constant tax rate, simply divide the tax income by the rate. If there is more than one rate, or if the rate changed during the period of time over which there is data, divide each amount by its related rate. Reconstructing the base is shown in columns 2 through 4 of Table 2 (the revenue data are artificial). Choose an index, for sales tax revenue the analyst might choose the Consumer Price Index (CPI) for all Urban Consumers, and convert nominal dollars to constant dollars using Equation 3 as shown in columns 4 through 7 of Table 2, using the CPI for all Urban Consumers based in 1982–1984 as

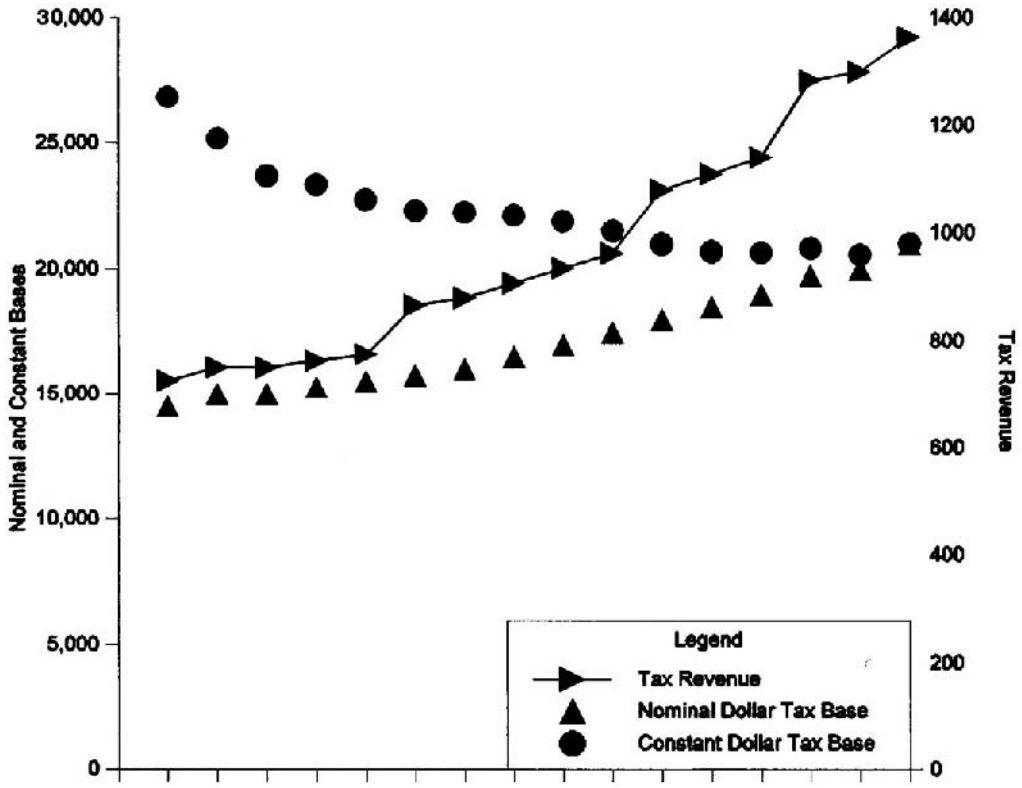


FIGURE 7 Comparing series.

published at http://www.bls.gov in July 1996 (for monthly data, the CPI and other deflators are also available in monthly factors).

Figure 7 demonstrates the effect of these calculations. The tax revenue (indexed against the right *Y* axis) grows faster than the nominal base (left *Y* axis), because the rate has several incremental increases. More significantly, while the nominal base is growing, the constant base (left *Y* axis) is shrinking. Forecasting the tax revenue or the nominal tax base without adjusting for these factors could lead to significant error.

4. Aggregated Data

Sometimes, two or more unrelated data series are added together. Figure 8 shows the total Virginia Medicaid enrollment from 1971 through 1994. With these data we see a massive enrollment climb beginning with 1990. It may be unrealistic to make a forecast that indefinitely projects the same sort of growth. By breaking the data into separate groups for the two major types of enrollment (Figure 9), we are able to observe that Aged and Disabled category contributes only a small amount to the accelerated enrollment growth. The more rapid growth is associated with Families with Children.

Examining the process that generates these data reveals that federal and state policies have prompted rapid growth in the Families with Children categories and that the growth should continue until 1999, but at a decelerating rate of growth. Additional examination shows that other federal policies prompted the more modest enrollment growth in the Aged and Disabled

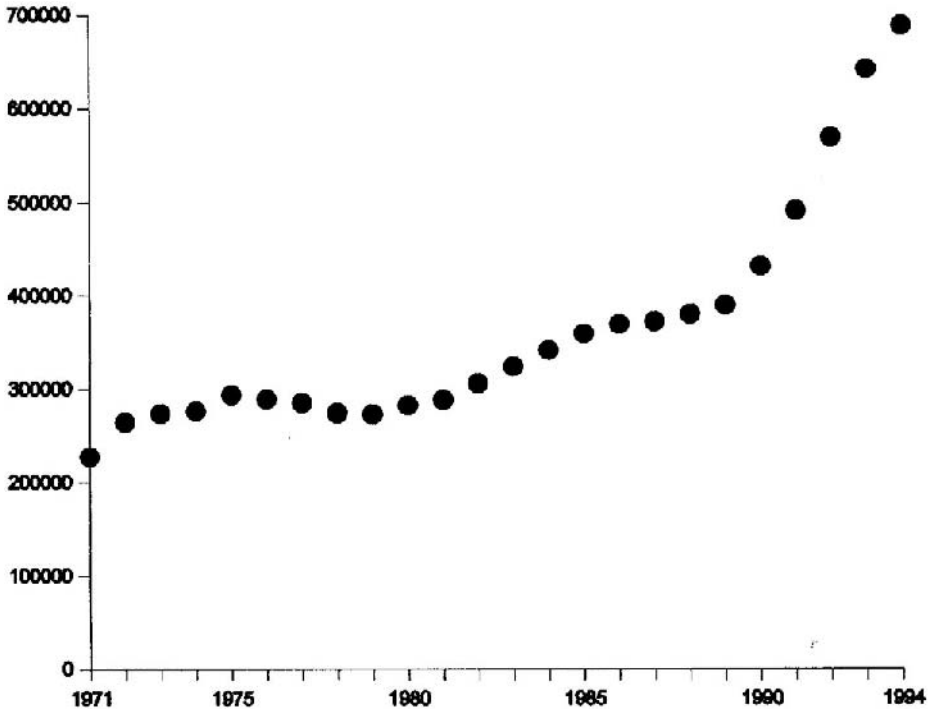


FIGURE 8 Total medicaid enrollment.

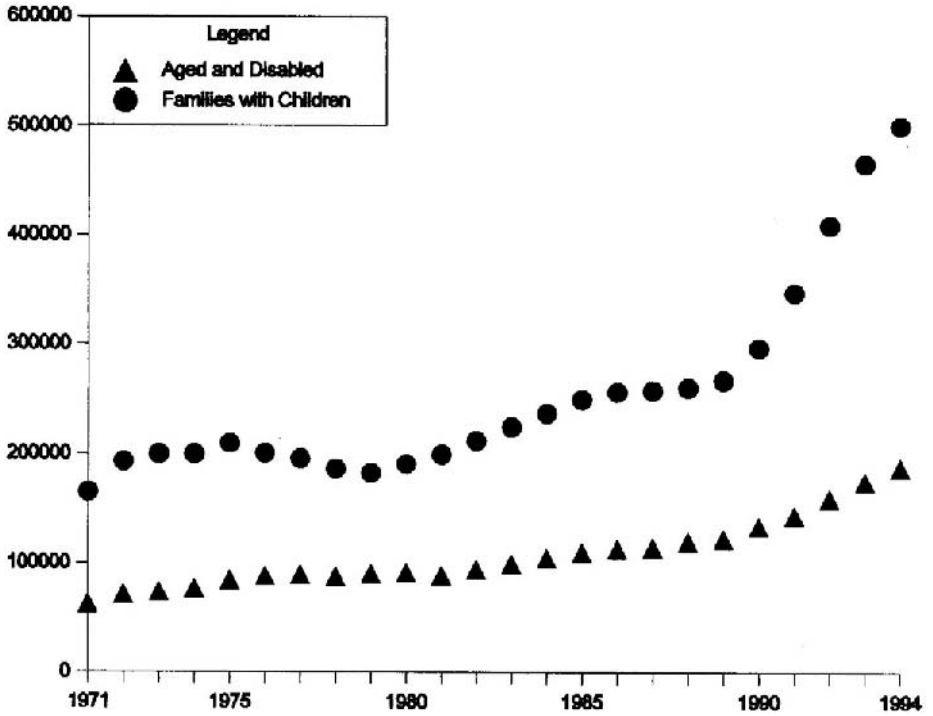


FIGURE 9 Disaggregated data.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

categories, that there might be a reduced rate of growth beginning in 1995, but that once that rate of growth stabilizes, it should continue indefinitely. This information suggests that the two series require different forecasting approaches. The Families with Children series needs an adjustment to prevent it from over-projecting enrollment after 1999, while the Aged and Disabled series does not require such a restriction. Further disaggregation could lead to further insights into these data.

In the examples of this and the previous sections, the decomposition or disaggregation is relatively simple, following a few easy steps; however, when working with real world data, analysts may need to go through a series of steps to decompose their data sufficiently to make forecasts. Practical analysts avoid decomposing their data so far that they have extremely small numbers; it is difficult to forecast a data series that has zero values for some observations or one that has large variation relative to the average observation. As a general guideline consider this, the Medicaid budget in Virginia accounted for roughly $2 billion in 1995, the state contribution to this expenditure accounted for roughly 13% of the state general fund revenue. Virginia disaggregated these data to forecast roughly 30 categories of health care service with each category of service further broken into a minimum of 2 data series and, except for two or three of the most complex categories, a maximum of 10 series.

5. Completeness

Another important consideration when breaking up data is whether the resulting series are complete. Break up of the data series provides the opportunity of discovering information left out of the combined data, but also increases the risk of losing something that is included in the gross data. For example, when forecasting income from licenses, what happens to fines for late applications? Also, if the licensee moves out of the locality, does he or she receive a refund? What source of money pays the refund? When working with financial data, obtain the organization's annual financial reports and reconcile the data sources with these reports. Find out what is missing and assure that it is accounted for. With other data, look for annual reports or other periodic reports with which to reconcile. Imagine how the data could be incomplete and look to see what happens with such data. An excellent forecast of the wrong data can be useless.

When decomposing complex data to make a forecast, decomposition must be reversed to complete the forecast. Combine the data by precisely reversing the steps followed when decomposing them.

III. FORECASTING

Some important components of variation in data series are known as level, trend, cycle, and seasonality (Makridakis, Wheelwright, and McGee, 1983). Each of these components is discussed below. Data are forecast by extrapolating these components.

A. Level

Level refers to the component of the data that determines its location on the *Y* axis. While some data series do not exhibit trend, cycle, or seasonality, all data series exhibit a level. Figure 10 shows a series that appears to vary around an unknown, but approximately constant, level. It does not particularly increase or decrease over time, nor does it show any other distinguishable pattern. Figure 11 shows the same series as a random scatter plot; data that varies randomly

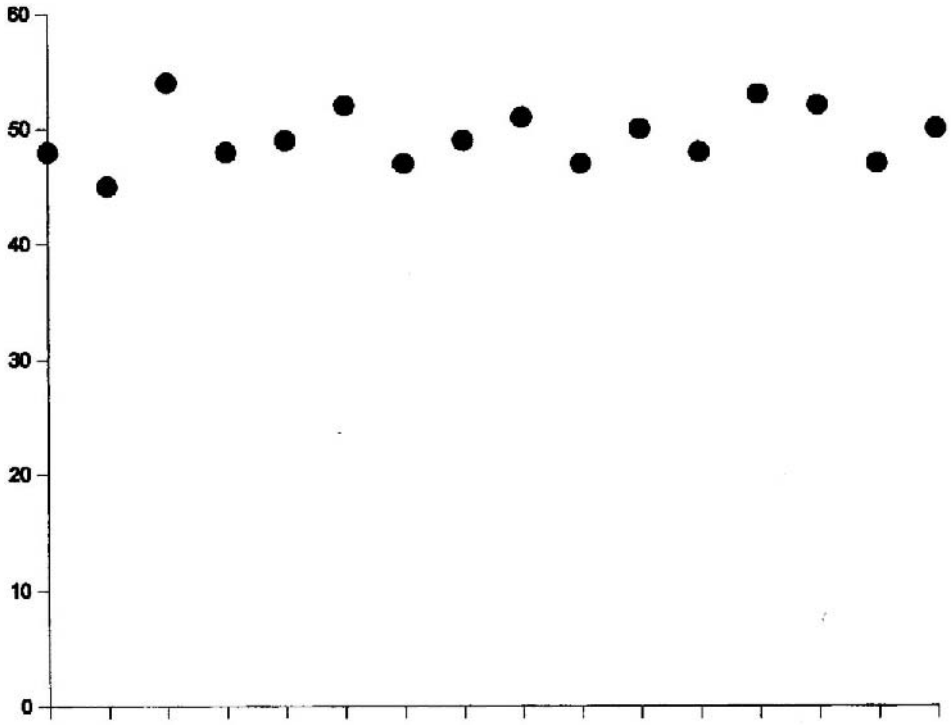


FIGURE 10 Variation around the average.

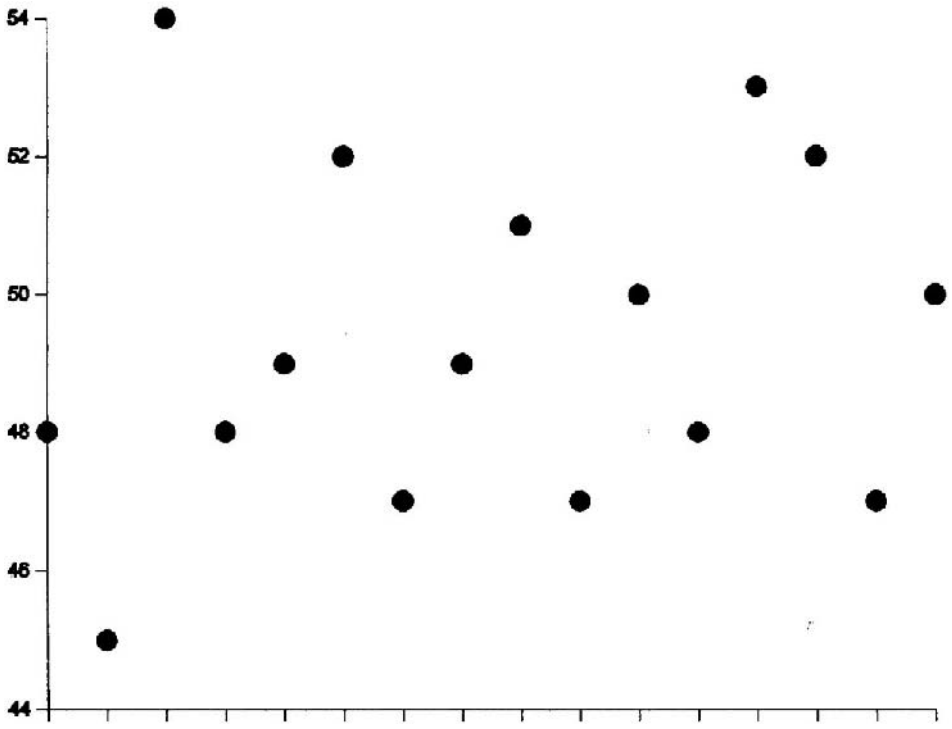


FIGURE 11 Random scatter plot.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 3 Average as Forecast

	Data	Forecast
Period 1	30	
Period 2	36	30
Period 3	31	33
Period 4	34	32.33
Period 5	29	32.75
Total	160	
t	5	
Average	32	
Future periods		32

around a mean will appear as a random scatter plot if the variation is large and the mean is near the *X* axis.

The data may not keep the same level across the whole series. Consequently, we need a way to talk about the level that is proximate to a particular location on a time index. Here I refer to such a level by subscripting the estimate of level with a time index, so S_t refers to level *S* at time *t*. I examine three stages of forecasting techniques for this kind of data (these stages develop the idea of forecasting the level, they are not necessarily the way these techniques developed).

1. Last Observation

A forecast model is an equation or set of equations used to generate a projected value. The simplest forecast model of non-trending data is to assume that the last observation will be repeated indefinitely into the future. Expressed mathematically, this model is $F_t = X_{t-1}$ (Equation 4), when X_t, runs out all future periods are forecast with $F_{t+m} = F_t$ (Equations 5). This assumption is sometimes called Random Walk or the Naive model (Makridakis et al., 1982; Armstrong, 1985). While this approach seldom produces the most accurate forecast, it provides a baseline for evaluating other forecast approaches. Any method used should perform no worse than this approach. If, after repeated use, a method's track record is worse than Last Observation, stop using it.

2. Improving Last Observation

Last Observation is overly influenced by the random component in the data series; while the last observation approaches a number, it misses it because of all the noise (random variation) that impacts the particular observation. Rather than moving to the future from the last observation, it is would be better to forecast from the central tendency of the data series. The arithmetic mean, more commonly known as the average, fulfills this requirement, so $F_{t+m} = S_t = \bar{x}_t = \sum_{i=1}^{t} X_i/t$ (Equation 6), where F_{t+m} is the forecast for the *m*th future period (any future period) at time *t*, *S* is the level at time, $\bar{x}$ is the average at time *t*, X_i is the vector of time ordered observations ending at time *t*, and *t* is the total number of observations at time *t*. Table 3 demonstrates this calculation.

a. Projecting the Past Table 3 forecasts both future and historical periods. The forecast for each historical periods is the average of all periods preceding that period, based on the information that would have been available to forecast the specific period. This forecast of the

TABLE 4 Loss Functions Used to Compare Forecasts

	Data	Forecast (average)	Deviation = e	e^2	Forecast (last obs)	Deviation = e	e^2
Period 1	30						
Period 2	36	30	6	36	30	6	36
Period 3	31	33	−2	4	36	−5	25
Period 4	34	32.33	1.67	2.78	31	3	9
Period 5	29	32.75	−3.75	14.06	34	−5	25
SSQ				56.84			95
t				4			4
MSE				14.21			23.75
RMSE				3.770			4.873

past allows comparison of forecast methods through a ''loss function.'' Loss functions are used to select a forecast model, that is, to choose between different ways to forecast data. A commonly used loss function is the Root Mean Squared Error, RMSE (Armstrong, 1985). Root Mean Squared Error is found by: RMSE $= \sqrt{(\sum_{i=1}^{t}(X_i - F_i)^2/t)}$ (Equation 7). There is no correct value for an RMSE since it depends on the size of the data and the amount of variability in the data; however, between any two values of RMSE, the smaller reflects greater accuracy. Table 4 compares two forecasts of the same data series as shown in Table 3. The first forecast is the average; the second is last observation. The forecasts use only the knowledge the forecaster could have at the time a forecast is made. For period 1 the forecasters had no information, so no forecast is made. Because of the distorting effect, this period is left out of the loss function calculation. For period 2, the forecaster could have known period 1, so the average is the last observation. Beginning with period 3, the two methods produce different forecasts. By forecasting with the average rather than the last observation, the RMSE is reduced from 4.87 to 3.77 or 23%.

3. Better Methods

While the average accounts for random variation, it treats the data as static. It assumes that the level is constant across the whole series. As shown in Figure 12, many series exhibit occasional shifts in the average. Two approaches for handling this condition are the moving average and single exponential smoothing.

a. Moving Average Figure 13 shows a moving average (Makridakis, Wheelwright, and McGee, 1983), which is calculated much like an average except that it is the average of only the most recent observations. As a new observation is included, the oldest observation is discarded. A moving average has some of the advantages of the average, yet it recognizes that the average contains irrelevant data that predate the most recent level shift. The number of periods included in the moving average depends on how frequently the level shifts. Since the average is expected to shift, it would be ineffective to include a large number of time periods; however, as the number of observations is reduced, the forecast becomes more and more susceptible to the random variation of the last few observations. The number of observations included in the moving average is a compromise between these considerations. A moving average is calculated using Equation 8 as demonstrated in Table 5:

$$F_{t+1} = \text{MAVG}_t = (X_t + X_{t-1} + \cdots + X_{t-(L-1)})/L \tag{8}$$

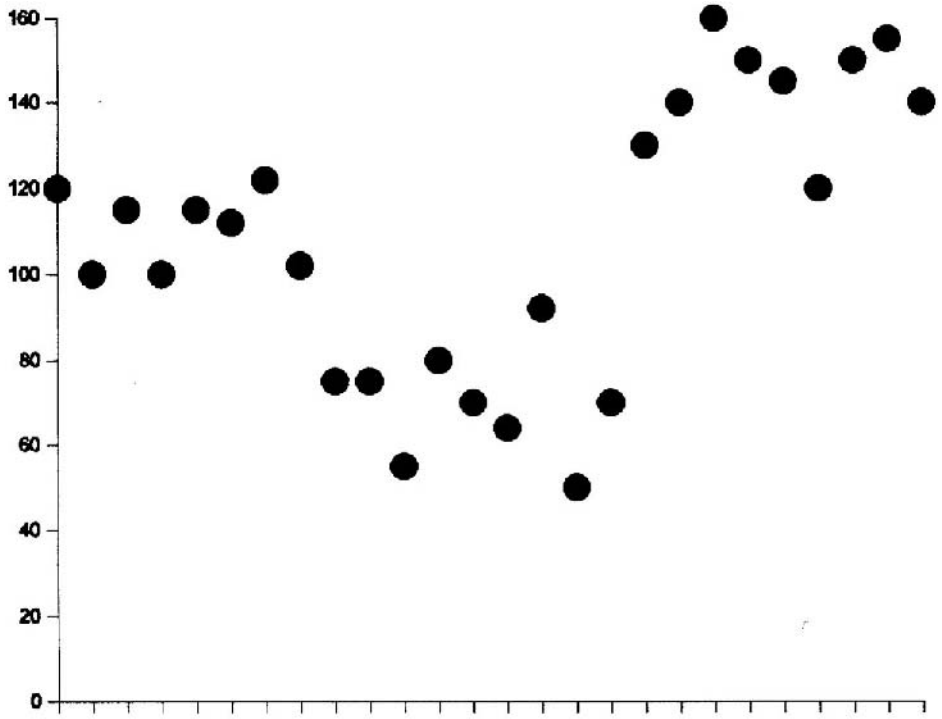


FIGURE 12 Shifting mean.

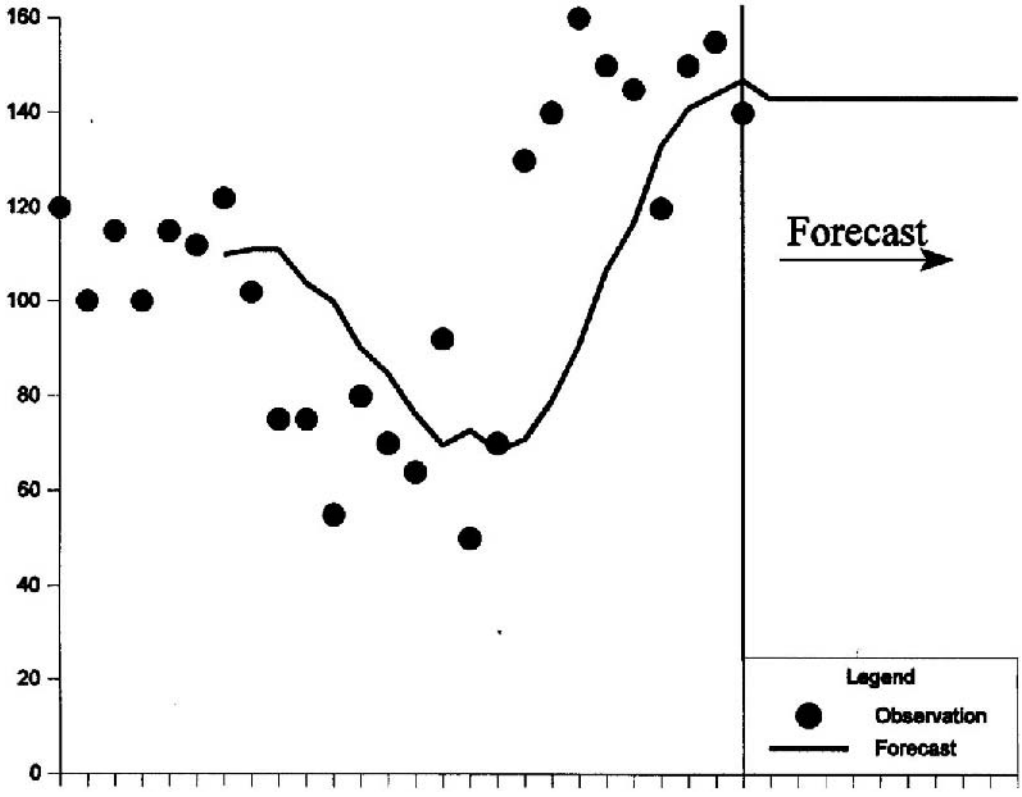


FIGURE 13 Moving average.

TABLE 5 Moving Average

X	Sum 6 obs	L	MA	MA as forecast	Error sq $= e^2$	Average as forecast	Error sq $= e^2$
120	—	—	—	—	—	—	—
100	—	—	—	—	—	—	—
115	—	—	—	—	—	—	—
100	—	—	—	—	—	—	—
115	—	—	—	—	—	—	—
112	662	6	110.3	—	—	—	—
122	664	6	110.7	110.33	136.11	110.33	136.11
102	666	6	111.0	110.67	75.11	112.00	100.00
75	626	6	104.3	111.00	1296.00	110.75	1278.06
75	601	6	100.2	104.33	860.44	106.78	1009.83
55	541	6	90.2	100.17	2040.03	103.60	2361.96
80	509	6	84.8	90.17	103.36	99.18	367.94
70	457	6	76.2	84.83	220.03	97.58	760.84
64	419	6	69.8	76.17	148.03	95.46	989.83
92	436	6	72.7	69.83	491.36	93.21	1.47
50	411	6	68.5	72.67	513.78	93.13	1860.48
70	426	6	71.0	68.50	2.25	90.44	417.69
130	476	6	79.3	71.00	3481.00	89.24	1661.76
140	546	6	91.0	79.33	3680.44	91.50	2352.25
160	642	6	107.0	91.00	4761.00	94.05	4349.06
150	700	6	116.7	107.00	1849.00	97.35	2772.02
145	795	6	132.5	116.67	802.78	99.86	2037.88
120	845	6	140.8	132.50	156.25	101.91	327.28
150	865	6	144.2	140.83	84.03	102.70	2237.70
155	880	6	146.7	144.17	117.36	104.67	2533.44
140	860	6	143.3	146.67	44.44	106.68	1110.22
Sum of squares	—	—	—	—	20862.81	—	28665.84
RMSE	—	—	—	—	32.3	—	37.9
Future periods	—	—	—	143.33	—	107.96	—

where,

F = forecast
X = observed
t = time index
L = length (number of periods in the average)

The forecast, F_t, at any particular period is the moving average as of the last previous period. For all future periods, it is the moving average where the historical data are exhausted. If the data are seasonal, follow the steps for deseasonalizing them discussed below before calculating the moving average. The last two columns show the comparative forecast based on the average. The bottom of the table shows that the six period moving average has a RMSE or 32.3 which is 15% lower than the RMSE of 37.9 for the simple average.

b. Single Exponential Smoothing (SES) Single Exponential Smoothing (SES) uses a parameter, α, to choose how much influence the most recent observation has on the forecast (Makridakis, Wheelwright, and McGee, 1983; Williams, 1987). If α is zero, the forecast is set at an original value (zero, unless the model is initialized). As α increases, more weight is placed

TABLE 6 Single Exponential Smoothing (SES)

	$\alpha = 0.5$		
X	$F = S_{(t-1)} + \alpha e_{(t-1)}$	e	e^2
120		120.0	
100	60.0	40.0	
115	80.0	35.0	
100	97.5	2.5	
115	98.8	16.3	
112	106.9	5.1	
122	109.4	12.6	157.8
102	115.7	−13.7	188.2
75	108.9	−33.9	1146.5
75	91.9	−16.9	286.6
55	83.5	−28.5	810.2
80	69.2	10.8	115.9
70	74.6	−4.6	21.3
64	72.3	−8.3	69.0
92	68.2	23.8	568.6
50	80.1	−30.1	904.6
70	65.0	5.0	24.6
130	67.5	62.5	3903.8
140	98.8	41.2	1700.8
160	119.4	40.6	1650.0
150	139.7	10.3	106.3
145	144.8	0.2	0.0
120	144.9	−24.9	621.1
150	132.5	17.5	307.6
155	141.2	13.8	189.6
140	148.1	−8.1	65.9
Sum of squares			12838.6
RMSE			25.3
Future periods	144.1		

on the recent observations. If α is 1, the forecast is last observation. SES is made with these equations:

$$F_{t+m} = \text{Forecast at time } t \text{ of time } t + m = S_t \tag{9}$$

$$S_t = \text{Level at time } t = F_t + \alpha e_t = S_{t-1} + \alpha e_t \tag{10}$$

$$e_t = \text{Error at time } t = X_t - F_t = X_t - S_{t-1} \tag{11}$$

where,

X_t = The observation at time t
α = Alpha, a level smoothing parameter subject to $0 \leq \alpha \leq 1$
t = A time index
m = The number of periods between an observation period and a forecast period.

Table 6 demonstrates SES using the same data as in Table 5. Over the same observations RMSE for SES is 25.3, 22% less than the 32.3 RMSE for the 6 period moving average and 33% lower

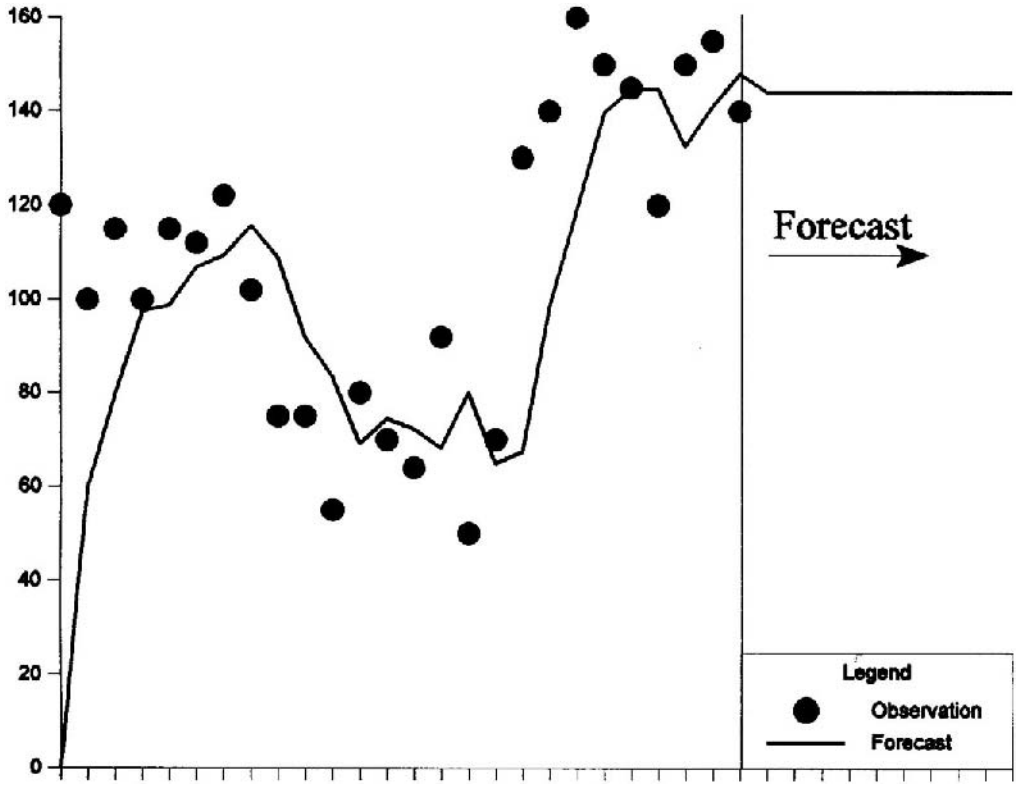


FIGURE 14 Single Exponential smoothing (SES).

than the 37.9 RMSE for the average. The forecast made with these equations is shown in Figure 14.

c. Selecting α The parameter of SES is α, a number which is multiplied times the error (the amount the current forecast missed the mark) to improve the next forecast. Choosing a particular value of a parameter is called fitting a forecast. In Table 6, α is arbitrarily set at 0.5. Common practice requires that α be set between 0 and 1 (formally, inclusive of these limits, but practically, more than zero and less than one). A common method for selecting a specific value for α is used to ''optimize α'' or ''optimize the model.'' This method uses a grid of potential α values such as in the first row of Table 7. The analyst calculates a forecast with each α value; determines the value of the loss function, as shown in the second row of Table 7 (for the same data as calculated in Table 6); and selects the optimal α value. The optimal α value is the one with the best loss function value, which for RMSE is the lowest value. Using the grid search of Table 7, the α value of 0.9 would be selected as optimal.

d. Initializing SES Initializing is selecting the first value of the SES forecast, the value of S_0 (the period before the period of the first available observation). In Table 6 the forecast beside the first observation is zero, resulting in a large error for this period. (Few applied forecasters would actually use such an uninitiated model.) Since $F_1 = S_0$, as given by equation 9, initializing provides a value for this missing forecast of the first period. Initialization is not difficult to accomplish. A likely initial value for SES is the average of the first few observations. Here, we could borrow the average of the first six periods from our earlier consideration of the 6 period moving average and initialize our forecast at 110.3. When I add this value to the

TABLE 7 Grid Search for Parameters for SES

α	0.05	0.1	0.2	0.3	0.5	0.7	0.9
RMSE	62.1	42.4	31.3	28.3	25.3	24.0	23.6

calculations in Table 6, I obtain a new forecast of 144.1 at the end of the series (unchanged to the first decimal point). RMSE drops to 25.4. The error for the first observation drops from $120 - 0 = 120$ to $120 - 110.3 = 9.7$, and the errors for the next several observations also drop considerably; however, these observations are known. The main purpose of the forecast is to project the observations beyond the last known value. As we have seen in the example, initializing may have little impact on the forecast into the future period. This result is consistent with forecasting literature (Makridakis and Hibon, 1991).

Little impact is not the same as none: First, when the forecast series is brief, non-initialization may impact the forecast into the future period. Second, initialization can impact the choice of α values. Table 8 compares the RMSE for the series {50, 55, 59, 48, 63, 57, 52, 54, 58, 51, 55, 47, 52, 64, 61, 50, 55, 47, 52, 59, 60, 50, 45, 57, 50, 55} first uninitialized and then initialized with, 55.3, the average of the first 6 observations. The RMSE penalty for selecting the lowest α value drops from 26.57 to 5.16 when the series is initialized. More importantly, the optimal parameter value changes from $\alpha = 0.3$, with a forecast beyond the last observation of 53.0, to $\alpha = 0.05$ with a forecast beyond the last observation of 54.2, or about 2% higher. This impact is the particular effect of initialization. Uninitialized SES models use observations at the beginning of the series to initialize themselves. When α is set low, SES uses more of the beginning observations to overcome the effects of not being initialized, so the model produces a higher RMSE. This can be seen by comparing the SES model estimate of the first 10 observations of the data in Table 6 using $\alpha = 0.05$ and $\alpha = 0.9$ as shown in Figures 15 and 16. If conditions are such that a low α value is optimal, failure to initialize the SES model may mislead the forecaster into selecting an excessively high α value.

B. Trend-Cycle

Trend refers to a data series' tendency to grow or shrink over time. On an *XY* graph this tendency can be observed as the slope of the line connecting the observations. Cycle refers to the data series' tendency to curve away from either the constant level or trend location but later to come back and curve away in the other direction. Except in its special seasonal variety, we will treat cycle as simply variation in trend. Thus, trend-cycle is the second component of a data series. It also varies over time so I refer to it as B_t. Figures 17 and 18 demonstrate trending data.

1. Last Change

If, from frequent experience with a data series, the forecaster knows that the observations usually grow or shrink, it is clear that the last observation is an unlikely value for the data. More likely, the data will continue to grow or shrink. A simple forecast method of this tendency begins

TABLE 8 Comparing Grid Search, Initialized, and Unintialized

	α	0.05	0.1	0.2	0.3	0.5	0.7	0.9
RMSE	(Uninitialized)	26.57	15.00	7.07	5.72	6.06	6.46	6.81
	(Initialized)	5.16	5.23	5.44	5.67	6.09	6.46	6.81

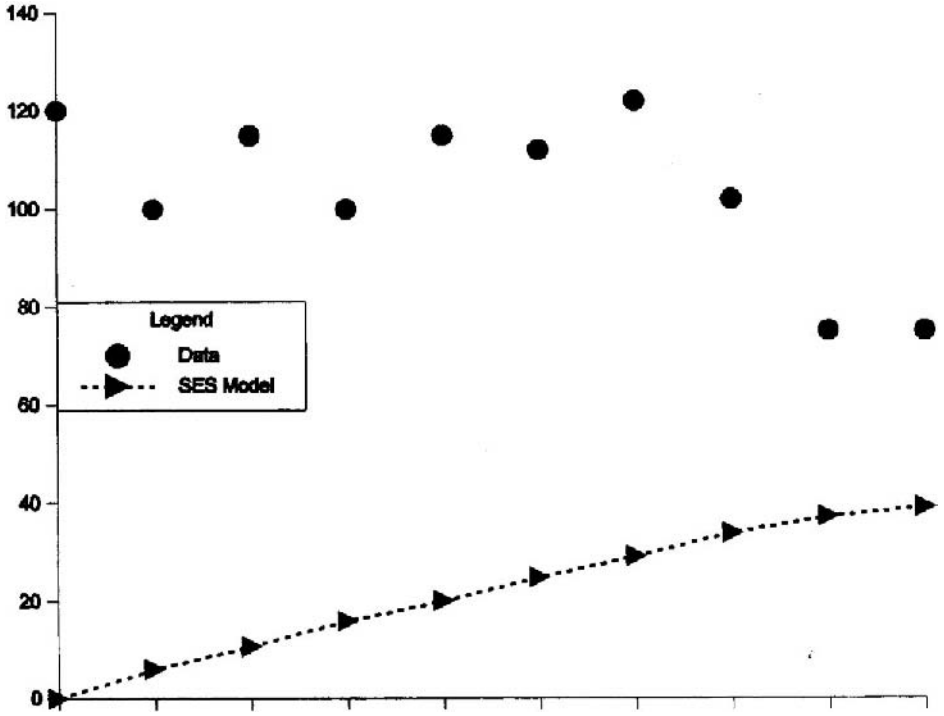


FIGURE 15 Uninitialized with alpha = 0.05.

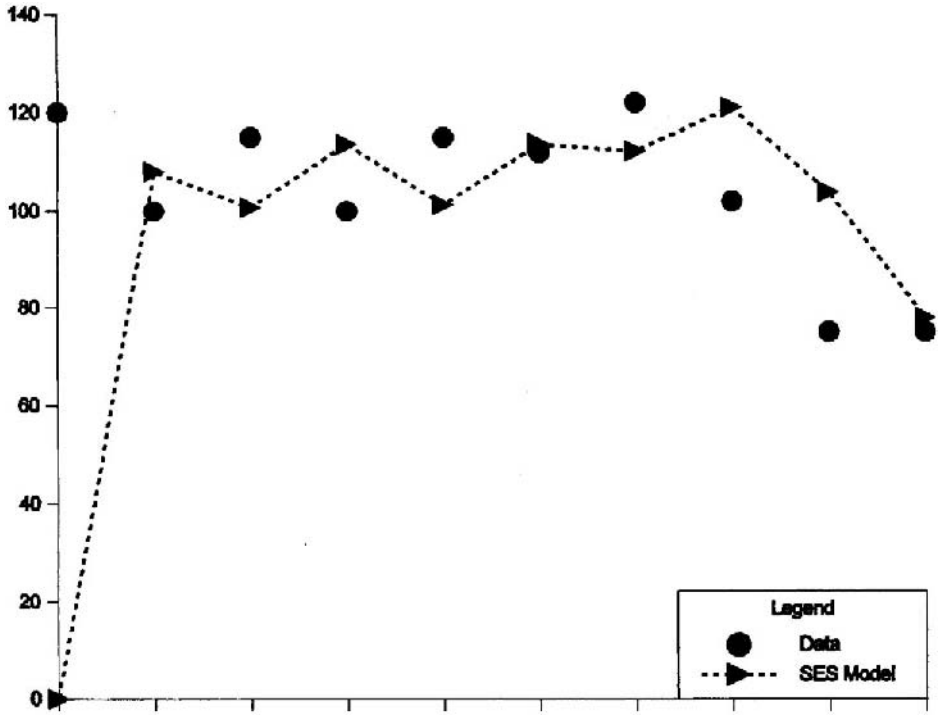


FIGURE 16 Uninitialized with alpha = 0.9.

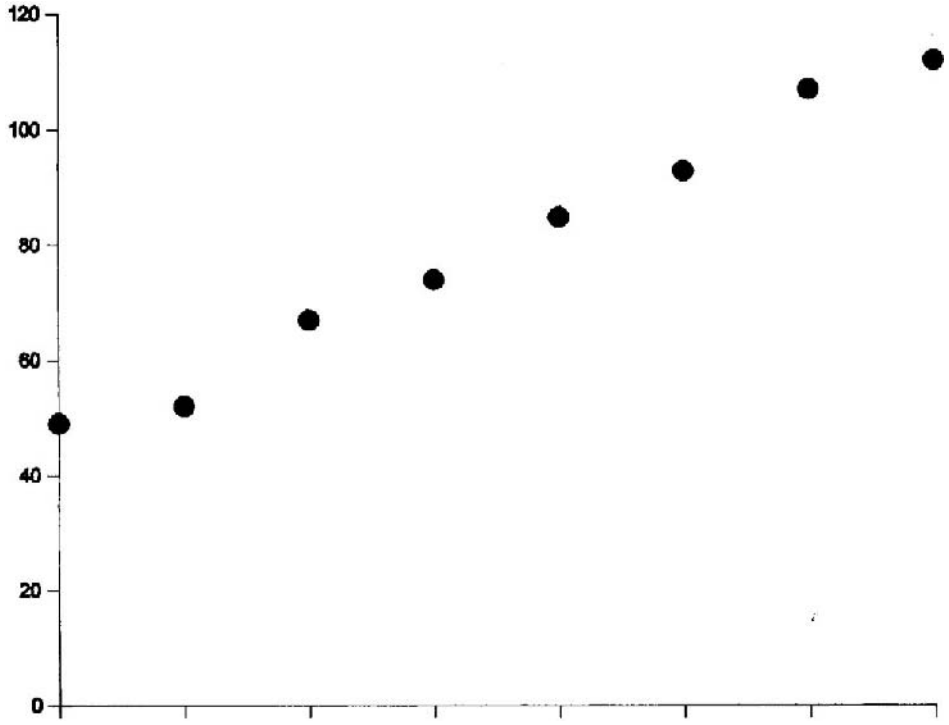


FIGURE 17 Upward sloping trend.

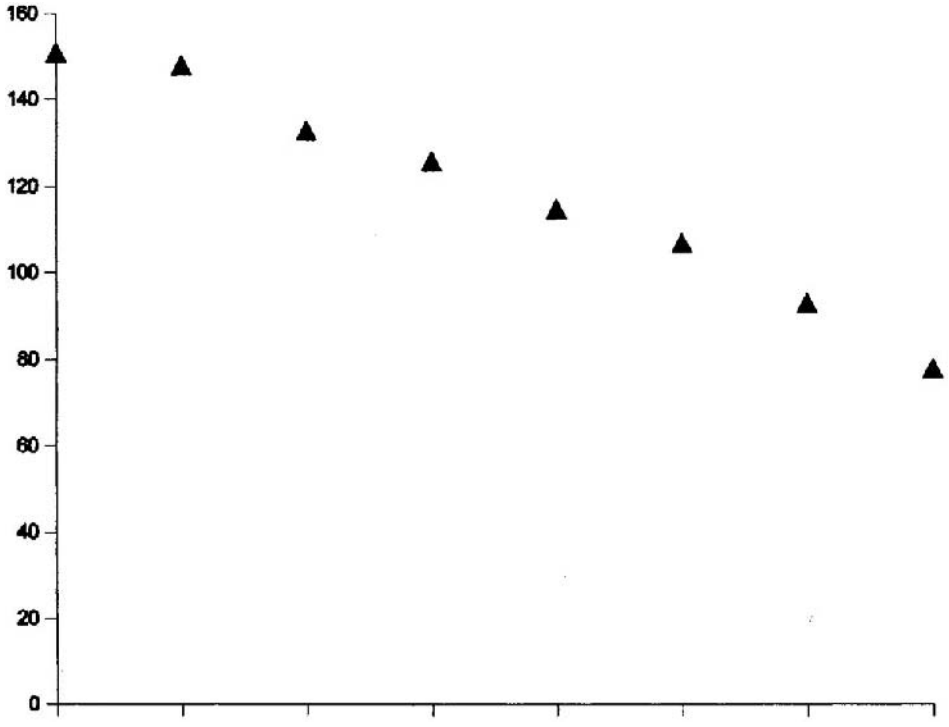


FIGURE 18 Downward sloping trend.

with last observation and adds last change. Last change is calculated as last observation minus observation before last:

$$B = \text{The change from Period 1 to Period 2} = X_2 - X_1 \tag{12}$$

$$F_{t+1} = \text{The forecast (next period)} = X_2 + B \tag{13}$$

$$F_{t+m} = \text{Forecast at time } t \text{ of the mth future period} = X_2 + m * B \tag{14}$$

Where, X_1 = the observed value for the older period and X_2 = the observed values for the more recent period.

Notice Equation 14. With data that do not trend, the forecast is the same for each future period; however, with these data, each future period has its own forecast.

Last change includes two undesirable sources of variation: First, the point for beginning calculation of the future data is last observation, which, as we have seen, is not a good estimate of the time adjusted central tendency of the data. Second, the estimate of change is taken from a single pair of observations, and the change between these two observations is strongly related to the random component of each of these observations. Last change can be improved through the following techniques.

2. *First Differences*

Differences capture the trending component of a data series. They are calculated using Equation 15:

$$D_t = \text{Any difference (period to period change)} = X_t - X_{t-1} \tag{15}$$

These differences, technically ''first differences,'' can be forecast using a moving average or SES. The result is a forecast of the trend component of the series. After forecasting the differences, reconstruct the data series by reversing the differences beginning with the first observation of the data series using these equations.

$$F'_2 = X_1 + F_1 \tag{16}$$

$$F'_3 = F'_2 + F_2,\ F'_4 = F'_3 + F_3, \text{ etc.} \tag{17}$$

Where X_1 is the first observation of the original (pre-differenced) data series, F'_t is the reconstructed data series (Forecast), and F_t is the moving average or SES forecast of the differenced data series. This unrolling process is shown in the two right columns of Table 9 below (after the discussion of trending exponential smoothing).

3. *Trending (Two Parameter) Exponential Smoothing*

A technique that is widely used for forecasting trending data is two parameter exponential smoothing. A common version of this technique is Holt exponential smoothing or Holt's method. The version discussed here allows the two parameters to be calculated independently through a minor modification developed by T. M. Williams. (1987) The forecast is produced through these equations as shown in Table 9:

$$F_{t+m} = \text{Forecast at time } t \text{ of time } t + m = S_t + (B_t * m) \tag{18}$$

$$S_t = \text{Level at time } t = F_t + \alpha e_t \tag{19}$$

$$B_t = \text{Trend at time } t = B_{t-1} + \beta e_t \tag{20}$$

$$e_t = \text{Error at time } t = X_t - F_t \tag{21}$$

TABLE 9 Holt Exponential Smoothing and SES of First Differences

		Holt exponential smoothing					SES of first differences			
				α 0.9	β 0.05	RMSE	SES	0.9		RMSE
Month	Data	F_1	e	S	B	5.3	D	F	F'	5.1
Jan	65.6		65.6	59.0	3.3					
Feb	66.1	62.3	3.8	65.7	3.5		0.5		65.6	
Mar	68.5	69.2	−0.7	68.6	3.4		2.4	0.5	66.1	
Apr	80.6	72.0	8.6	79.7	3.9		12.1	2.2	68.3	
May	77.6	83.6	−6.0	78.2	3.6	36.1	−3.0	11.1	79.4	3.1
June	78.3	81.8	−3.5	78.6	3.4	12.0	0.7	−1.6	77.8	0.3
July	80.0	82.0	−2.0	80.2	3.3	4.2	1.7	0.5	78.2	3.1
Aug	81.9	83.5	−1.6	82.1	3.2	2.5	1.9	1.6	79.8	4.3
Sept	80.2	85.3	−5.1	80.7	3.0	25.7	−1.7	1.9	81.7	2.2
Oct	82.7	83.7	−1.0	82.8	2.9	0.9	2.5	−1.3	80.3	5.5
Nov	78.8	85.7	−6.9	79.5	2.6	47.7	−3.9	2.1	82.5	13.4
Dec	78.0	82.1	−4.1	78.4	2.4	16.4	−0.8	−3.3	79.2	1.4
Jan	91.8	80.8	11.0	90.7	2.9	121.8	13.8	−1.1	78.1	187.2
Feb	94.0	93.6	0.4	94.0	2.9	0.2	2.2	12.3	90.4	12.7
Mar	94.0	96.9	−2.9	94.3	2.8	8.4	0.0	3.2	93.6	0.1
Apr	82.2	97.1	−14.9	83.7	2.0	221.3	−11.8	0.3	94.0	138.4
May	88.7	85.7	3.0	88.4	2.2	8.8	6.5	−10.6	83.4	28.3
June	90.9	90.6	0.3	90.9	2.2	0.1	2.2	4.8	88.2	7.5
July	91.7	93.1	−1.4	91.8	2.1	1.9	0.8	2.5	90.6	1.2
Aug	92.3	94.0	−1.7	92.5	2.1	2.8	0.6	1.0	91.6	0.5
Sept	97.3	94.5	2.8	97.0	2.2	7.7	5.0	0.6	92.2	25.7
Oct	97.1	99.2	−2.1	97.3	2.1	4.5	−0.2	4.6	96.8	0.1
Nov	104.6	99.4	5.2	104.1	2.3	27.1	7.5	0.3	97.1	56.7
Dec	108.4	106.4	2.0	108.2	2.4	3.9	3.8	6.8	103.8	20.7
Forecast		110.6		110.6	2.4			6.8	110.6	
		113.1						6.8	117.4	

Where,

X_t, α, t, and m are as used with SES, and,
β = Beta, a trend smoothing parameter subject to $0 \leq \beta \leq 1$

The forecast calculated in Table 9 is shown in Figure 19. For comparison I show SES of the first differences in the four columns to the right. Because of differences in initialization, the RMSE is calculated ignoring the first four periods. Selection of a forecast model among these two techniques is briefly discussed in Section IV of this chapter. With Holt exponential smoothing both α and β are restricted to values between zero and one.

a. Initializing Holt Table 9 calculations are not initialized. Since Holt involves estimating both a level and a trend, initialization is more complex than for SES. For monthly data, calculate the initial level, S_0, and the initial trend, B_0, using equations 22 through 25, where X refers to monthly observations. The forecaster must make appropriate adjustments to use this with data cumulated over other intervals.

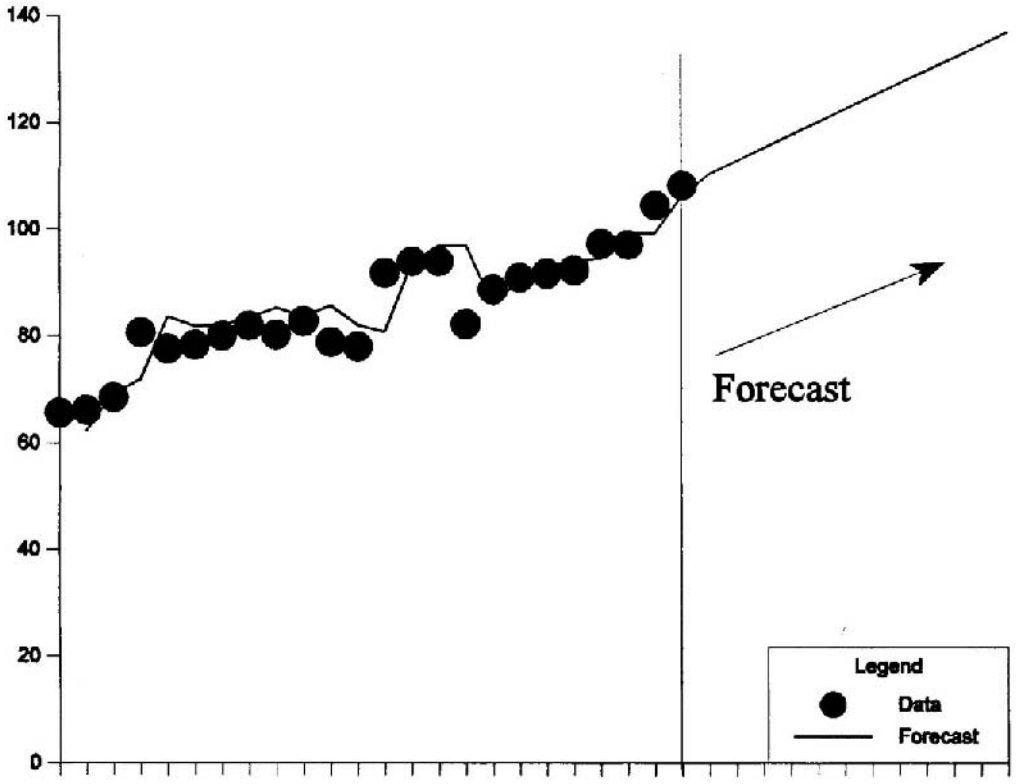


FIGURE 19 Holt exponential smoothing.

a. Calculate the average, $\bar{x}_j$ for j = periods 1 and 2 (such as two sequential years) of length L:

$$\bar{x}_j = \sum_{i=1}^{L} X_{i,j}/L \tag{22}$$

(that is, add the observations in each of the first two cycles and divide each by L)

b. Calculate the difference (D) between these two averages:

$$D = \bar{x}_2 - \bar{x}_1 \tag{23}$$

c. Divide this Difference by L to get an initial trend (B_0):

$$B_0 = D/L \tag{24}$$

d. Multiply B by $(L + 1)/2$ and subtract from $\bar{x}_1$ to get (S_0).

$$S_0 = \bar{x}_1 - B * (L + 1)/2 \tag{25}$$

For monthly data, $L = 12$ and $(L + 1)/2 = 6.5$. For quarterly data $L = 4$ and $(L + 1)/2 = 2.5$. Treat S_0 and B_0 as if they were calculated in the month prior to the month of the first observation, so the forecast value for the first month of actual data is $F_1 = S_0 + B_0$ (Equation 26).

TABLE 10 Grid Search for Holt Exponential Smoothing

		Beta (β)					
α, β		0.05	RMSE	0.1	RMSE	0.3	RMSE
Alpha (α)	0.1	0.1, 0.05	25.2	0.1, 0.1	27.7	0.1, 0.3	32.1
	0.2	0.2, 0.05	15.4	0.2, 0.1	17.7	0.2, 0.3	2.8
	0.4	0.4, 0.05	8.1	0.4, 0.1	10.2	0.4, 0.3	11.5
	0.6	0.6, 0.05	6.0	0.6, 0.1	7.4	0.6, 0.3	8.1
	0.9	0.9, 0.05	5.3	0.9, 0.1	5.9	0.9, 0.3	6.6

b. Selecting α and β The parameters for Holt Exponential Smoothing are selected following the same grid search process as with SES. Table 10 shows a typical grid along with the RMSE for each combination for the data shown in the previous example. With these results, one would select the α, β combination 0.9, 0.05.

C. Decelerating Trends

The trending data we have examined up to here involves linear trends, that is trends that can be approximately modeled using a line. It is also possible that trending data follows a curve. In Figures 20 and 21, the trend is reducing over time. There are many ways to conceptualize such data. It may be that the data are contained within some greater universe of data and cannot

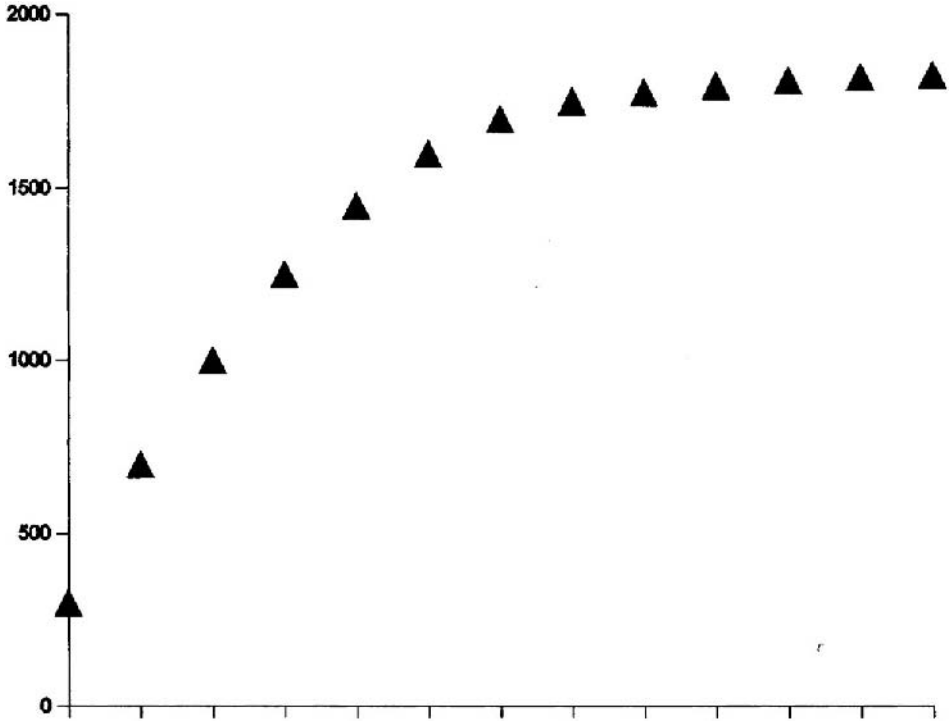


FIGURE 20 Decelerating growth.

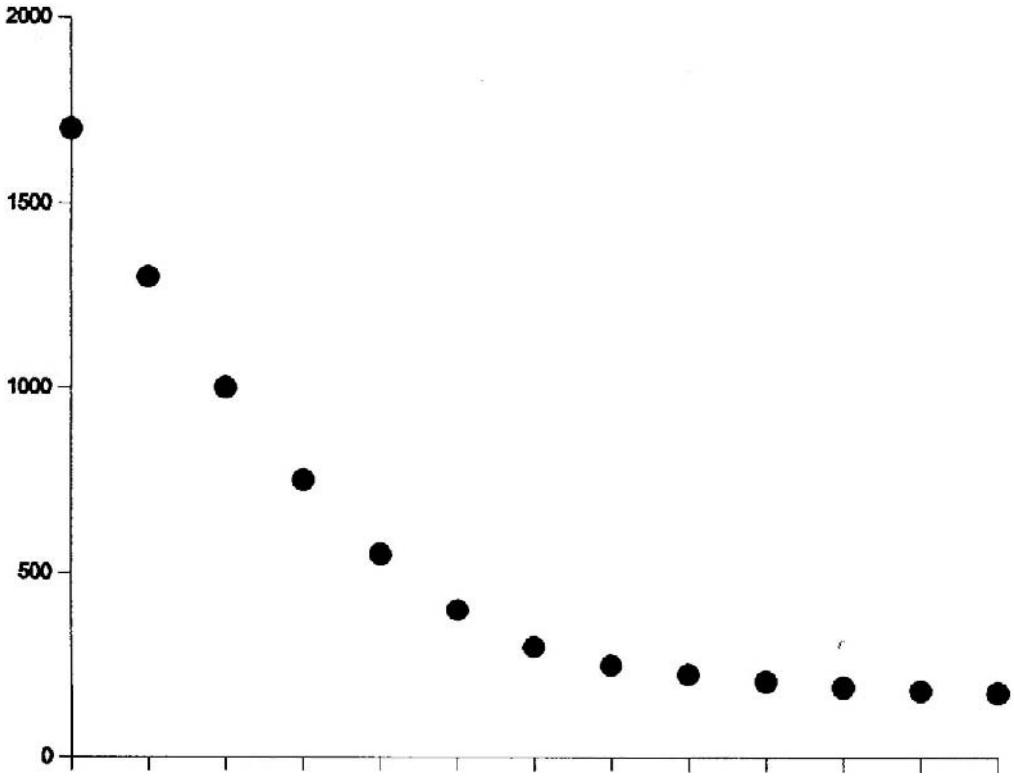


FIGURE 21 Decelerating decline.

exceed some maximum or minimum. For example, when forecasting the growth in employment in the sector of privatized government activities, there is a natural maximum roughly associated with current government employment. In such a case, examine the literature on growth curves (discussion of growth curves is beyond the scope of this chapter). The data in Figure 21 might also be thought to be dependent on the magnitude of its level component; this topic is briefly discussed in the next section, which addresses accelerating trends.

When trends decelerate during the historical period, a well-fit Holt model will produce an adequate forecast. However, sometimes the analyst anticipates that the trend will decelerate or decelerate further in the future. An approach that allows for this adjustment is known as the dampened trend method (Gardner and McKenzie, 1985):

$$F_{t+m} = \text{Forecast at time } t \text{ of time } t + m = S_t + \sum_{i=1}^{m} (\phi^i B_t) \tag{27}$$

$$S_t = \text{Level at time } t = F_t + \alpha e_t \tag{28}$$

$$B_t = \text{Trend at time } t = \phi B_{t-1} + \beta e_t \tag{29}$$

$$e_t = \text{Error at time } t = X_t - F_t \tag{30}$$

where, X_t, α, β, t, and m are as with Holt exponential smoothing, and,

ϕ = Phi, a dampening factor subject to $0 \leq \phi \leq 1$

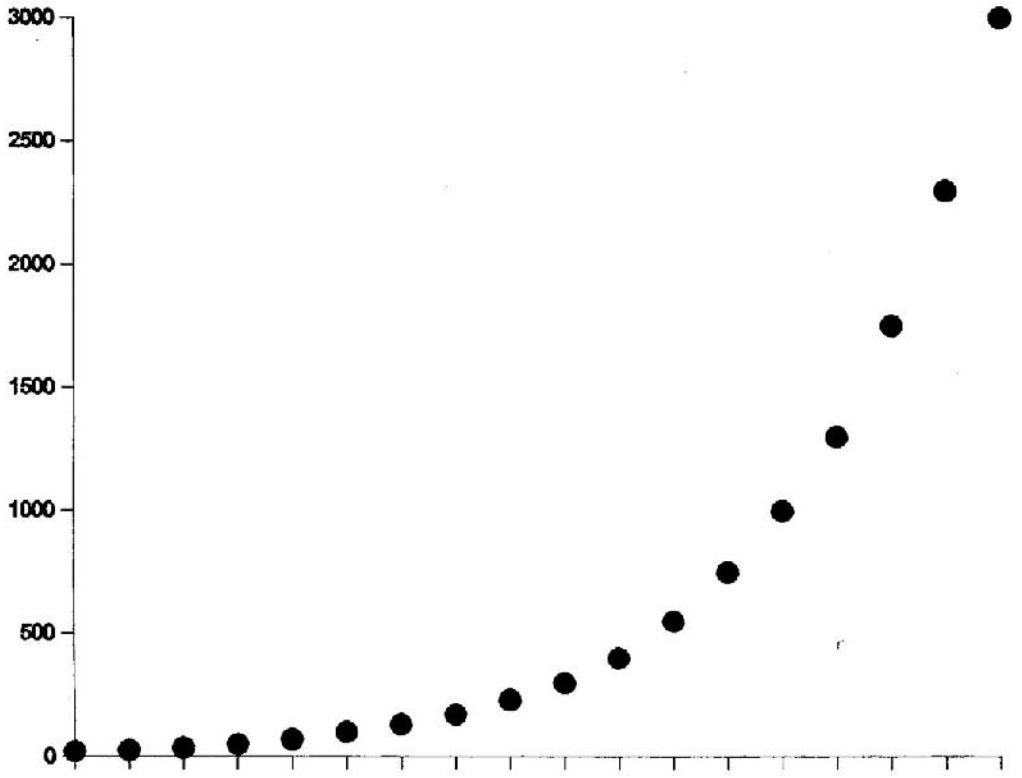


FIGURE 22 Accelerating growth.

This dampened trend technique is optimized as if it were the Holt technique. The dampening factor, ϕ, is not optimized. Instead, the forecaster uses it to implement a judgment that the trend will wither away over time. Smaller values of ϕ cause the trend to fade rapidly, while larger values allow the it to fade gradually.

D. Accelerating Trends

Figures 22 and 23 show data that have accelerating trends. Accelerating trends are very problematic for forecasting. A moment's reflection will show that no natural process can accelerate without limit. If the limit can be identified, an approach to forecasting these data is the use of growth curves as mentioned above. When data have a natural limit, and the use of growth curves is not desired, it is possible to emulate the effect of a growth curve using a dampened trend as discussed in the previous section. However, the result will only capture the linear and decelerating components of the trend. It will not capture the accelerating component.

When the rate of acceleration is fairly small, as with population growth or some cases of price inflation, it is possible for acceleration to occur over an extremely long period of time. In such cases, use of ratios might be relevant. These are discussed briefly; however, the reader should be very careful with these techniques. Ratios are appropriate only when the magnitude of the change depends on the magnitude of the level. Thus, the following discussion applies to data that may be similar to Figures 21 or 22. This approach is not appropriate for data that resembles Figures 20 or 23.

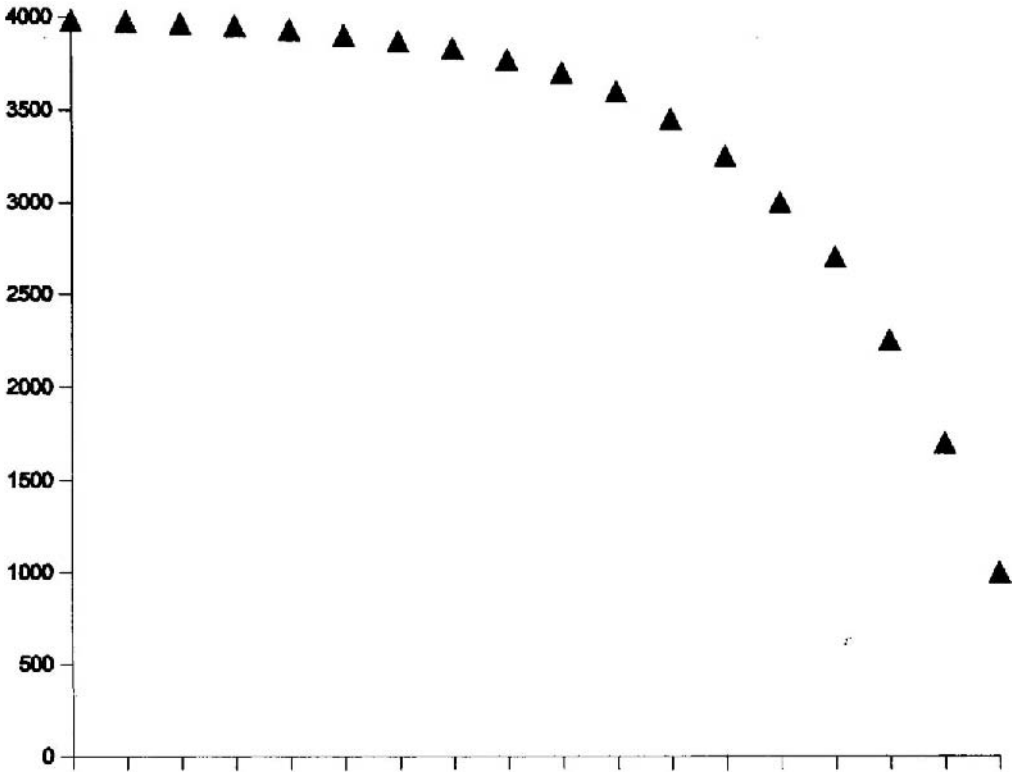


FIGURE 23 Accelerating decline.

Divide each observation by the prior observation $R_t = X_t/X_{t-1}$ (Equation 31); then forecast the resulting ratios (R_t) with a moving average or SES. Do not use Holt exponential smoothing or any other trending method for forecasting these ratios. As with first differences, a forecast using ratios is completed by reconstructing a forecast series that originates at the beginning of the data series to avoid adding unnecessary randomness into the forecast. The ratio approach can also be thought of as forecasting the percentage growth. A percentage (P) is simply a special way of representing a ratio, $P = (R - 1) * 100$ (Equation 32). Ratios may be easier to use because the reconstruction of the forecast series is less complex.

Another approach to use with caution is second differences (or ''second ordered differences''), which can be used with accelerating or decelerating trends. First calculate the first differences using equation 15. With accelerating or decelerating data, these differences still exhibit a trend. To eliminate this trend, calculate a second series of differences from the first series; calculate D_3'' (where D_t'' means a second difference ending in time period t) from D_2 and D_3 as follows $D_3'' = D_3 - D_2$, etc. (Equation 33). Now the data should not have a trend. With decelerating trends the second differences have the opposite sign of the first differences. Do not continue differencing the data. Forecast the second differences using SES or a moving average, then reverse the differencing process to reconstruct the entire data series.

If data resembles the curve in Figure 23, consider these alternatives:

1. Within a few time periods the process generating these data will cease.
2. The rate of accelerating decline is very small, so a linear approximation may provide

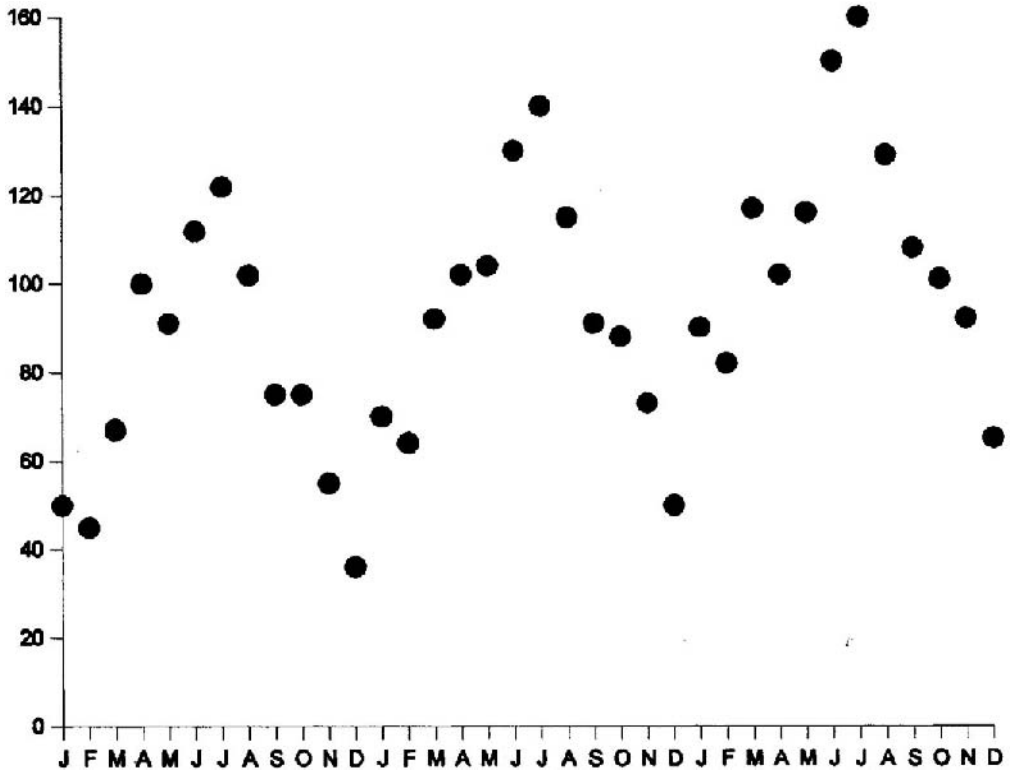


FIGURE 24 Annual seasonality.

an adequate forecast (use Holt exponential smoothing, SES of first differences, or a moving average of first differences).

3. The data are approaching zero and will not reach zero, so a dampened trend is appropriate.
4. The data have no natural lower limit, so second differences might be appropriate.

E. Seasonality

A cyclical pattern of special interest is seasonal variation. Seasonal patterns repeat over a fixed period of time such as a year (Makridakis et al., 1983). In Figure 24 we observe peaks in July and troughs in December. Seasonality may be easier to observe if the X axis of the graph is limited to the length of the suspected season and sequential cycles are graphed separately as shown in Figure 25. Data that tends towards the same ups and downs over each segment may be seasonal. If the overlapping series appear random, the data are not seasonal.

While seasonality is commonly thought of as an annual phenomenon, it is also possible to have seasonality within other time segments. Figure 26 demonstrates seasonality within weeks. This phenomenon arises when there is something special about the relationship between the data and the day of the week. Traffic is likely to exhibit seasonality within weeks. Figure 27 demonstrates seasonality within quarters. Data may reflect this sort of seasonality when management is particularly anxious to record data within the earliest quarter possible.

Sometimes seasonality is dependent on the level of the series at the time of the season.

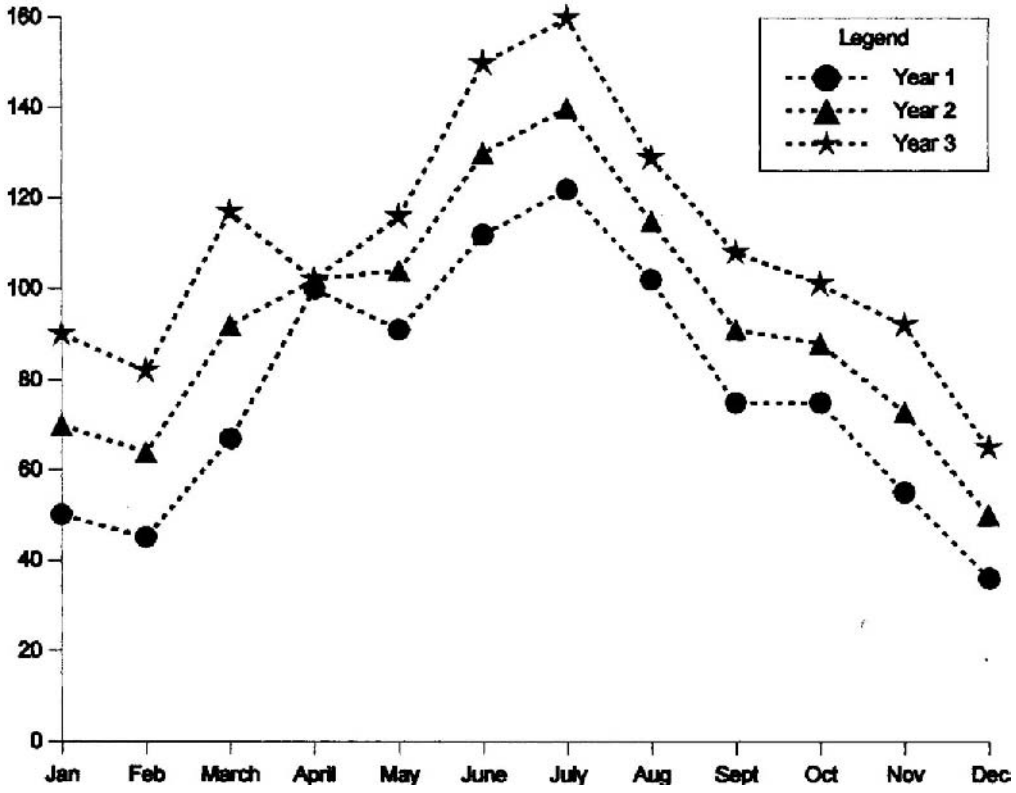


FIGURE 25 Annual seasonality: overlapping years.

Other times seasonality is unrelated to the level. Judge this by asking whether the seasonal difference is additive (more like ''50 units more in December'') or multiplicative (more like ''15% more in December''). Figure 28 shows multiplicative quarterly seasonality around linear growth of 10 units a month. With larger values of the level, the seasonal peaks and troughs get farther away from the line. Figure 29 shows additive quarterly seasonality around the same 10 units per month growth; in this figure the size of the peaks and troughs is unrelated to level.

1. Calculating a Simple Seasonal Index

Deseasonalizing means adjusting a series to remove seasonal impact. The following shows the calculation of a simple annual seasonal index for monthly data for both multiplicative and additive techniques. This technique requires a minimum of two seasonal cycles, but works better with three or more seasonal cycles.

a. First, calculate a double moving average as follows:

1. Calculate a 12 period moving average using equation 8 where $L = 12$.
2. Still using equation 8, calculate a 2 period moving average ($L = 2$) of the 12 period moving average. This new moving average is a 12×2 double moving average. For seasonal periods other than monthly, calculate an $L \times 2$ moving average where L is the number of periods for one seasonal cycle. For the rest of this explanation, the 12×2 moving average is labeled S_t.

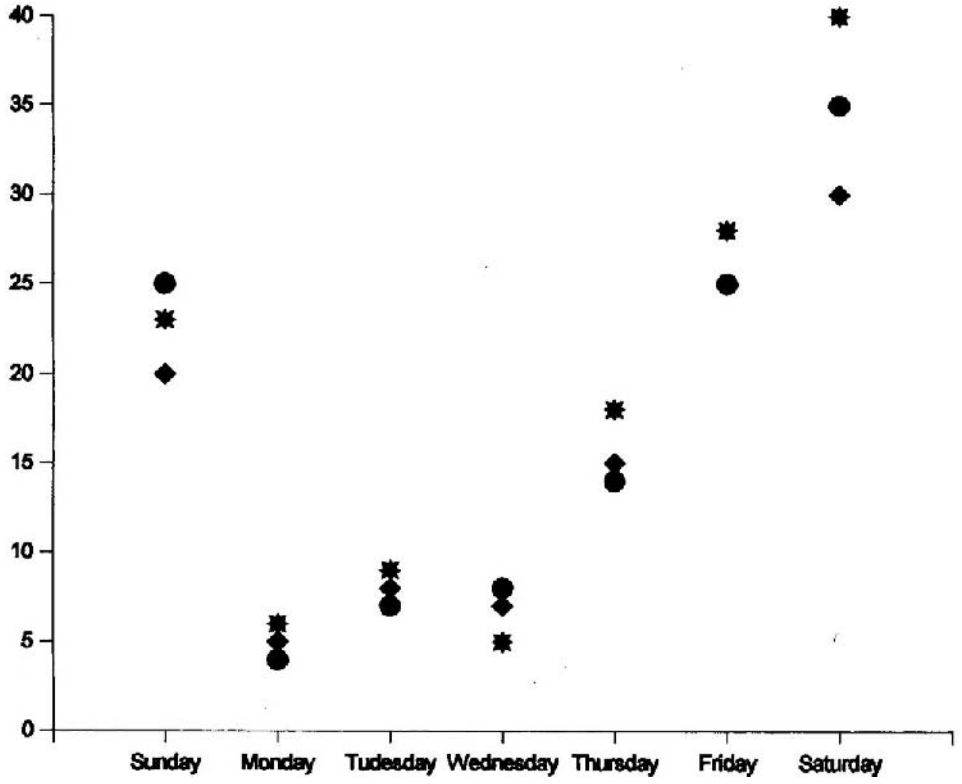


FIGURE 26 Seasonality within weeks.

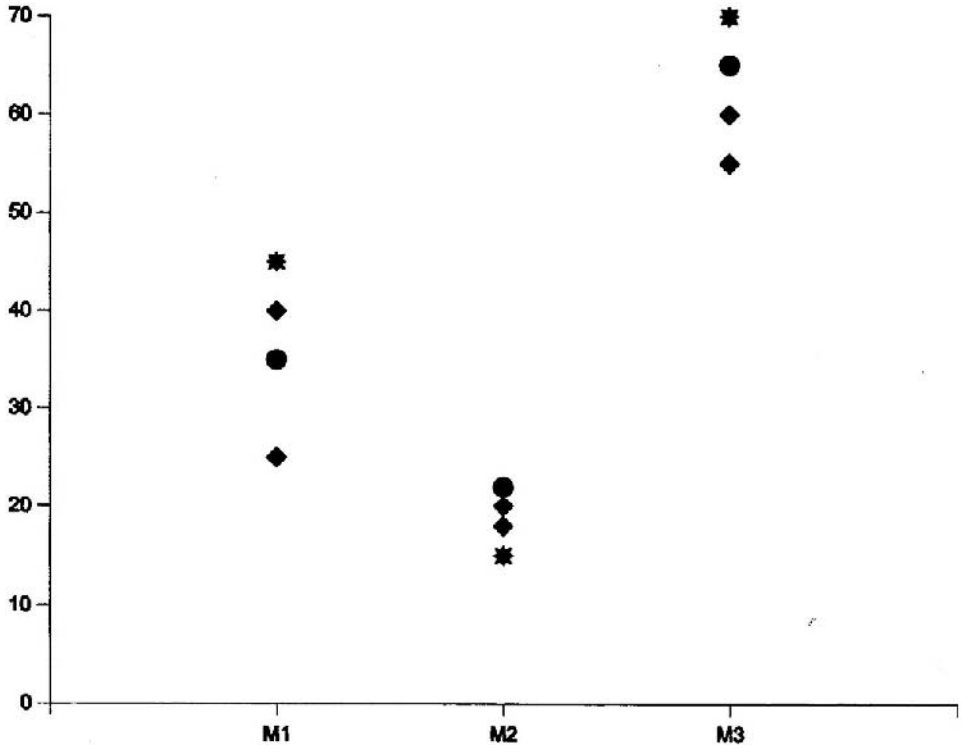


FIGURE 27 Seasonality within quarters.

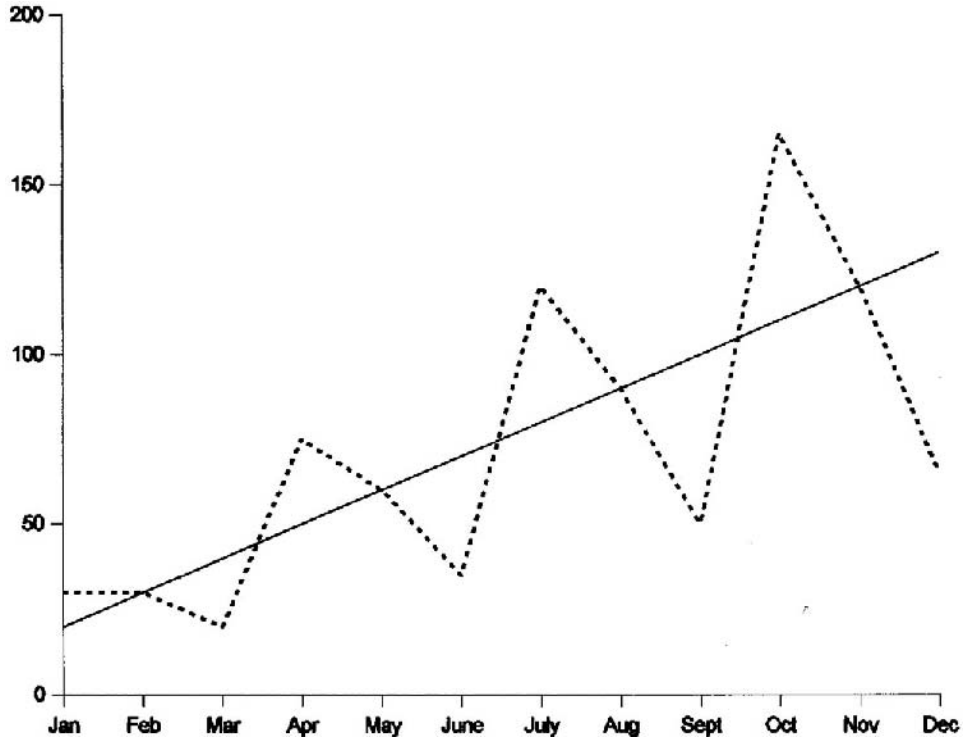


FIGURE 28 Multiplicative seasonality.

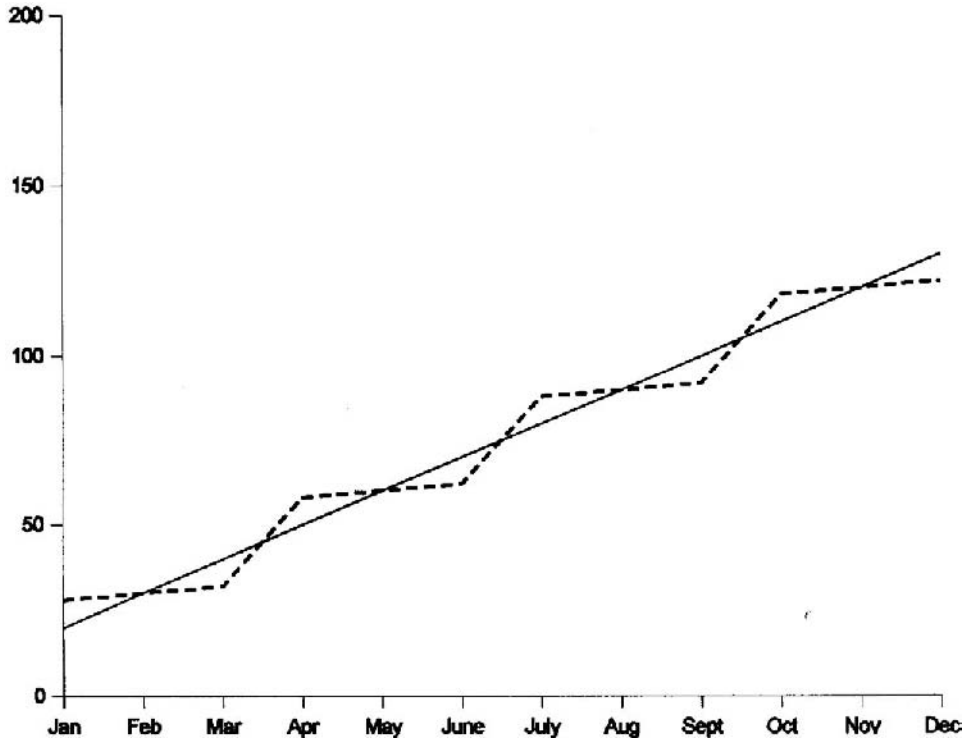


FIGURE 29 Additive seasonality.

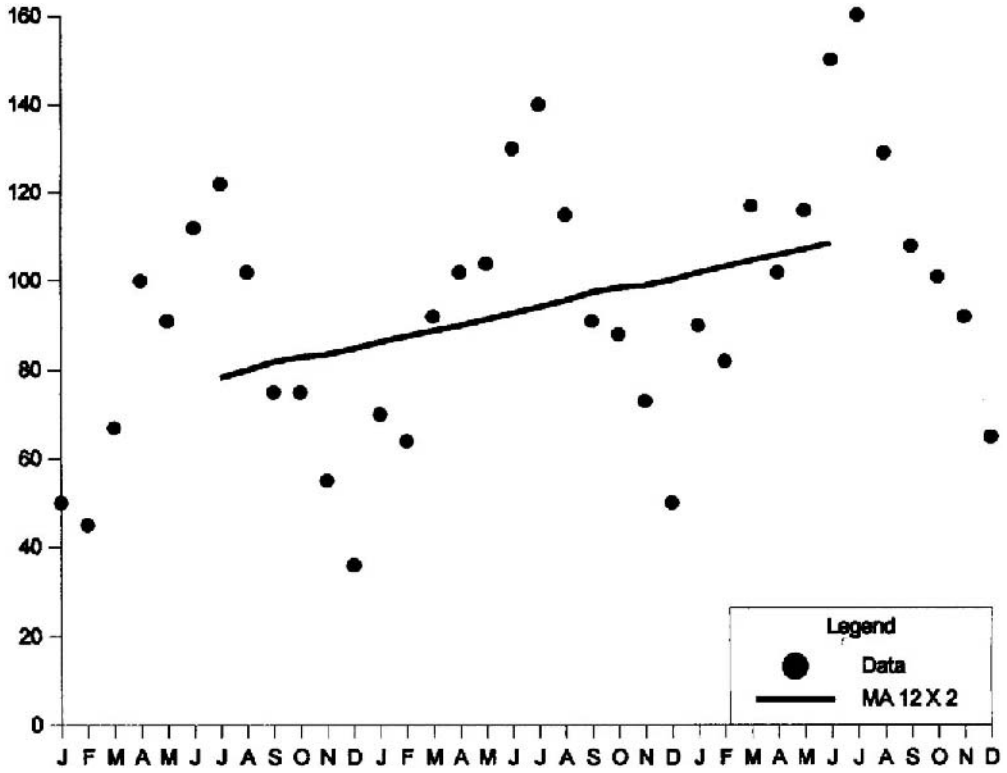


FIGURE 30 Trend-cycle component.

3. The center of a moving average is found by the expression $(L - 1)/2$ (Equation 34). In the case of 12 monthly periods, these two moving averages are centered, respectively, at period 6.5 and period 7. To calculate a seasonal index, enter the result of the S_t in the same row as the actual observation for its centered period (July for years starting in January). There are 12 fewer moving average values than raw values (the six observations at the beginning and the six at the end do not have enough observations for a 12×2 period moving average to be centered beside them).

These calculations produce an estimate of the trend-cycle in the data as shown by the solid line in Figure 30. The seasonal data are shown with the scatter plot. The centered trend-cycle estimate extends from the seventh period through the n-6th period.

b. To calculate a multiplicative seasonal index, proceed as follows:

1. Calculate an approximate index (I') by dividing each actual value by the S_t value as described in step a.

$$I'_7 = X_7/S_7,\ I'_8 = X_8/S_8,\ \ldots,\ I'_{n-6} = X_{n-6}/S_{n-6} \tag{35}$$

2. Average the index estimates for each month to get a smoother index (I)

$$I_{\text{JUL}} = (I'_7 + I'_{19} + I'_{31} + \cdots)/\text{Count of Julys}, \tag{36}$$

$$I_{\text{AUG}} = (I'_8 + I'_{20} + I'_{32} + \cdots)/\text{Count of Augusts}, \ldots$$

3. Divide the actual data by the index to obtain deseasonalized data (DESEAS)

$$\text{DESEAS}_t = X_t/\text{I}_t \tag{37}$$

c. For additive seasonality, follow these steps:

1. Calculate an approximate factor (I') by subtracting each S_t from the actual observation:

$$I'_7 = X_7 - S_7,\ I'_8 = X_8 - S_8,\ \ldots,\ I'_{n-6} = X_{n-6} - S_{n-6} \tag{38}$$

   If S is greater than X, the approximate factor should be negative.

2. Average the factor estimates for each month to get a smoother factor (I):

$$I_{\text{JUL}} = (I'_7 + I'_{19} + I'_{31} + \ldots)/\text{Count of Julys},$$
$$I_{\text{AUG}} = (I'_8 + I'_{20} + I'_{32} + \ldots)/\text{Count of Augusts}, \ldots \tag{39}$$

3. Subtract the factor from the actual data to obtain deseasonalized data (DESEAS):

$$\text{DESEASE}_t = X_t - I_t \tag{40}$$

The use of these equations is demonstrated in Table 11. In the column labeled Index and the column labeled Factor, the average of the I' is calculated in the boxed area, the values shown above and below that area repeat the values from the same months in the calculation area.

Figures 31 and 32 show the results of multiplicative and additive deseasonalization. The deseasonalized series is marked with triangles.

Forecast the deseasonalized data using a technique such as a moving average, SES, or Holt. To complete the forecast, re-seasonalize the data in order to know what to expect for various months. Re-seasonalize the multiplicative series by multiplying it by the seasonal factor, or re-seasonalize additive data by adding back the same additive factor.

This discussion has focused on annual seasonality of monthly data with the year beginning in January, but the actual data may begin in any month, be quarterly data, or have seasonality over some period other than a year. If appropriate adjustments are made, this method can be used with data cumulated over any interval and with any seasonal cycle. For example, if the data are quarterly rather than annual, the result is a total of 8 observations (4 for each year). Using 8 observations would require adjustments to the equations, such as getting the slope by dividing by 4 rather than 12.

1. Normalizing

Some forecasters suggest that multiplicative forecasting factors should be normalized to sum to 12 (or to L, the number of periods of one seasonal cycle) and additive factors should be normalized to sum to zero. Multiplicative factors can be normalized to sum to L by first summing them, second calculating the ratio L divided by the sum, and third multiplying each factor by the resulting ratio. Additive factors can be normalized to sum to zero by first calculating the average of the factors, then subtract this average from each factor, this normalizing process may cause a seasonal factor to change from positive to negative or vice versa, in which case normalizing may be misleading.

2. Aggregating Data

An alternative for working with seasonal data is to aggregate them across the season and forecast the aggregated data. For example, if the season is monthly within quarters, cumulate the data to quarters (four observations a year, each accumulated across three months). There are two important restrictions on cumulating data across seasons. First, the forecast must not need to be updated more frequently than allowed by the level of aggregation chosen. For example, a

TABLE 11 Seasonality

				Multiplicative seasonal			Additive seasonal		
Month	Data	MA 12	MA 12 × 2	Apx Ind	Index	Deseas	Apx Fct	Factor	Deseas
Jan	50				0.847	59.1		−14.2	64.2
Feb	45				0.762	59.1		−22.5	67.5
Mar	67				1.077	62.2		7.7	59.3
Apr	100				1.048	95.4		4.0	96.0
May	91				1.110	82.0		10.7	80.3
June	112				1.391	80.5		39.3	72.7
July	122	77.5	78.3	1.557	1.523	80.1	43.7	44.8	77.2
Aug	102	79.2	80.0	1.276	1.239	82.3	22.0	20.7	81.3
Sept	75	80.8	81.8	0.917	0.925	81.1	−6.8	−6.6	81.6
Oct	75	82.8	82.9	0.905	0.899	83.4	−7.9	−9.2	84.2
Nov	55	83.0	83.5	0.658	0.698	78.8	−28.5	−27.3	82.3
Dec	36	84.1	84.8	0.424	0.461	78.0	−48.8	−49.6	85.6
Jan	70	85.6	86.3	0.811	0.847	82.7	−16.3	−14.2	84.2
Feb	64	87.1	87.6	0.730	0.762	84.0	−23.6	−22.5	86.5
Mar	92	88.2	88.8	1.036	1.077	85.5	3.2	7.7	84.3
Apr	102	89.5	90.0	1.133	1.048	97.4	12.0	4.0	98.0
May	104	90.6	91.3	1.139	1.110	93.7	12.7	10.7	93.3
June	130	92.1	92.7	1.403	1.391	93.4	37.3	39.3	90.7
July	140	93.3	94.1	1.488	1.523	91.9	45.9	44.8	95.2
Aug	115	94.9	95.7	1.202	1.239	92.8	19.3	20.7	94.3
Sept	91	96.4	97.5	0.934	0.925	98.3	−6.5	−6.6	97.6
Oct	88	98.5	98.5	0.893	0.899	97.9	−10.5	−9.2	97.2
Nov	73	98.5	99.0	0.737	0.698	104.6	−26.0	−27.3	100.3
Dec	50	99.5	100.3	0.498	0.461	108.4	−50.3	−49.6	99.6
Jan	90	101.2	102.0	0.882	0.847	106.3	−12.0	−14.2	104.2
Feb	82	102.8	103.4	0.793	0.762	107.7	−21.4	−22.5	104.5
Mar	117	104.0	104.7	1.117	1.077	108.7	12.3	7.7	109.3
Apr	102	105.4	106.0	0.963	1.048	97.4	−4.0	4.0	98.0
May	116	106.5	107.3	1.081	1.110	104.5	8.7	10.7	105.3
June	150	108.1	108.7	1.380	1.391	107.8	41.3	39.3	110.7
July	160	109.33			1.523	105.1		44.8	115.2
Aug	129				1.239	104.1		20.7	108.3
Sept	108				0.925	116.7		−6.6	114.6
Oct	101				0.899	112.4		−9.2	110.2
Nov	92				0.698	131.8		−27.3	119.3
Dec	65				0.461	140.9		−49.6	114.6

forecast cannot be updated each month if the data are aggregated to quarters. Second, the analyst must not need to know about smaller units of data than the aggregated level. If data have been aggregated to quarters, the analyst cannot speak about monthly data.

3. *Second Differences*

Another alternative is to calculate the first differences of the seasonal period. Subtract observations that occur at the same point in two sequential seasons. For example, if the data follow annual seasonality, as with the previous discussion, calculate differences across years by taking the observation in January of year one and subtracting it from the observation in January of

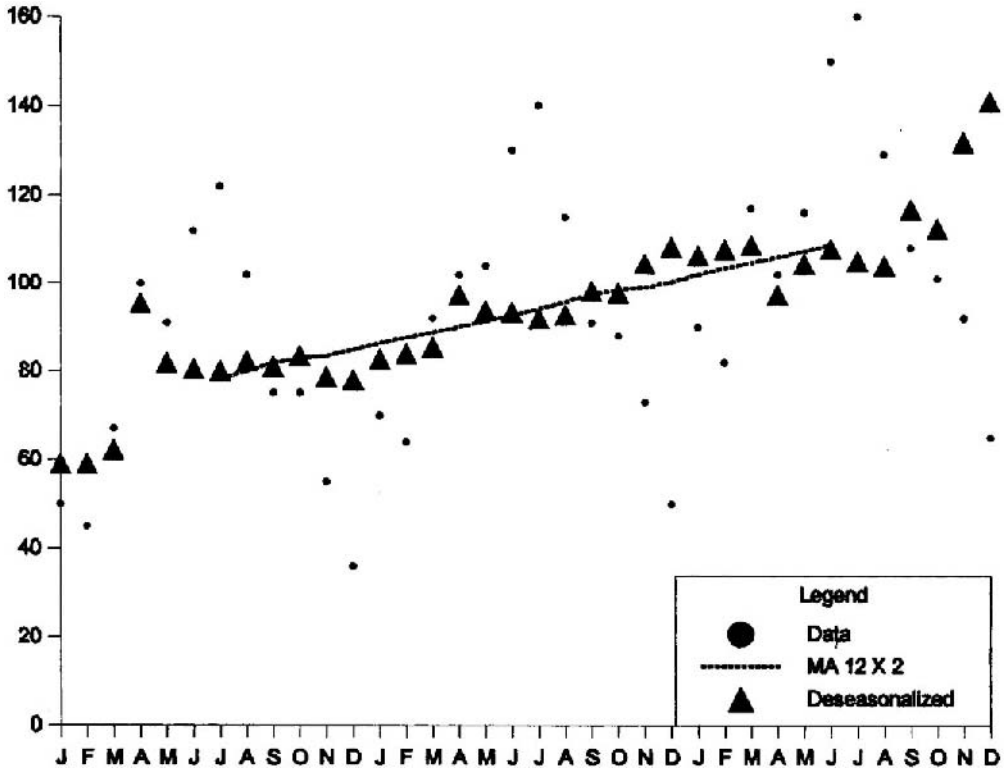


FIGURE 31 Multiplicative seasonal adjustment.

year two. For year three, subtract year two data from year three data, continuing until the data run out. The resulting data are no longer seasonal. Observations have been differenced (subtracted one from another) at the same point in the season, between them there was no seasonality. All the new observations are without seasonality. These differenced data will also reflect the same impact on trend as occurred when the first differences of the first period (that is, the sequential differences) were calculated. The differenced data can be forecast with SES or a moving average. Then the differencing must be reversed to produce the full forecast.

4. Winters' Three Parameter Method

A technique developed by Winters (1960) permits seasonality to update within the exponential smoothing model. Following is Holt-Winters exponential smoothing as modified by Williams (1987) to allow for independent parameter optimization:

$$F_{t+m} = \text{Forecast at time } t \text{ of time } t + m = (S_t + B_t * m) * I_{t+m-L} \tag{41}$$

$$F_t = \text{Forecast at time } t = (S_{t-1} + B_{t-1}) * I_{t-L} \tag{42}$$

$$S_t = \text{Level at time } t = S_{t-1} + B_{t-1} + \alpha e_t / I_{t-L} \tag{43}$$

$$B_t = \text{Trend at time } t = B_{t-1} + \beta e_t / I_{t-L} \tag{44}$$

$$I_t = \text{Seasonal Index at time } t = \gamma e_t / (S_{t-1} + B_{t-1}) + I_{t-L} \tag{45}$$

$$e_t = \text{Error at time } t = X_t - F_t \tag{46}$$

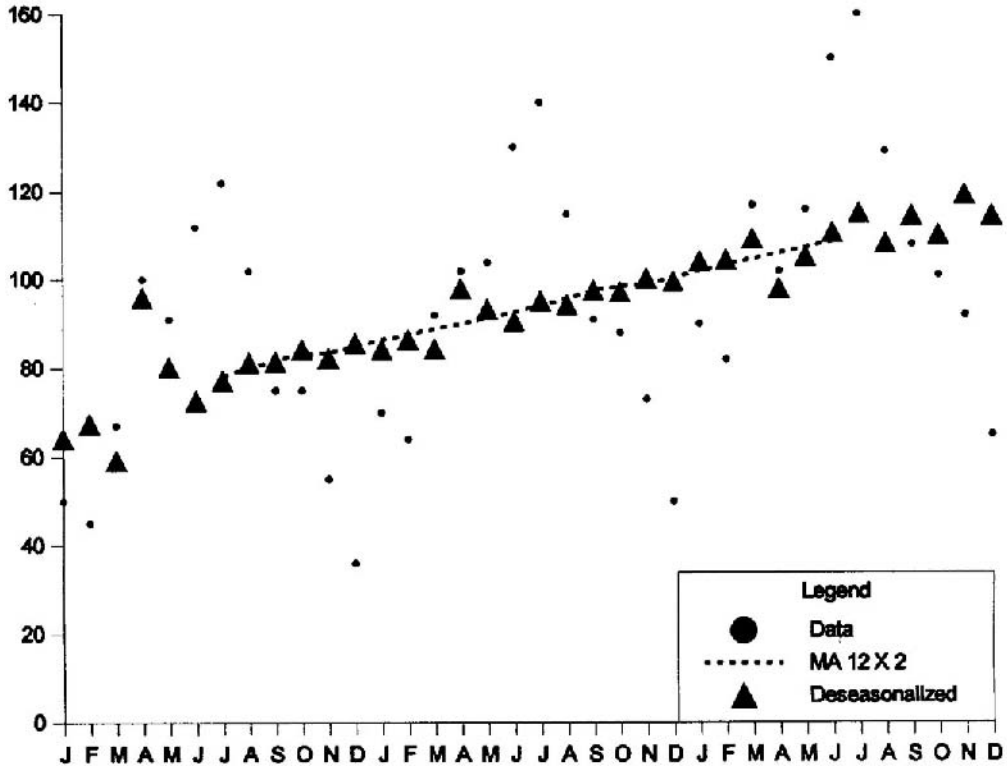


FIGURE 32 Additive seasonal adjustment.

where, X_t, α, β, t, and m are as with Holt exponential smoothing, and,

γ = Gamma, a seasonal smoothing parameter subject to $0 \leq \gamma \leq 1$

L = The length of the season (number of periods until the season repeats).

Since I_t updates the previous I_{t-L}, there are as many updating I factors as there are periods in the seasonal cycle, L. Also, unless otherwise initialized, I must be set at an initial value of 1, the value zero is invalid. Some forecasters suggest that the I values should be normalized to sum to L, and re-normalized with each update. At minimum the forecaster should take care that the factors do not vary radically from this norm.

Table 12 demonstrates the calculation of this equation using the seasonal data in Table 11. Initial seasonal factors are taken from the multiplicative seasonal index in Table 11. Figure 33 shows the results of these calculations.

a. Initialization Trend and level components can be initialized using Equations 22 through 25. Because each seasonal factor is updated only once each seasonal cycle, the self-initialization of the Winters' model can be slow, so initialization is recommended. A convenient way to calculate initial seasonal factors is to use the multiplicative seasonal factors from Equations 35 through 37 as calculated in Table 11. Equations 35 through 37 are more effective when there are at least two seasonal factors for each period (two Januaries, Februaries, Marches, etc.) to average. Because of the $L \times 2$ period moving average, they are available only if there are three seasonal cycles of data (for annual data, three years). If only two cycles of data are available, calculate the S_1 values without a moving average (thereby retaining the ability to average two seasonal factors) through these steps. First, calculate the trend and level using Equations

TABLE 12 Multiplicative Holt-Winters

Month	Data	F_t	e	α 0.6 S	β 0.01 B	γ 0.1 I	RMSE = 3.4
Jan	50		50.0	35.4	0.6	0.8	
Feb	45	27.44	17.56	49.86	0.82	0.81	
Mar	67	54.56	12.44	57.62	0.94	1.10	
Apr	100	61.35	38.65	80.69	1.31	1.11	
May	91	91.01	−0.01	81.99	1.31	1.11	0.00
June	112	115.89	−3.89	81.62	1.28	1.39	15.16
July	122	126.23	−4.23	81.23	1.25	1.52	17.86
Aug	102	102.18	−0.18	82.39	1.25	1.24	0.03
Sept	75	77.40	−2.40	82.09	1.22	0.92	5.74
Oct	75	74.89	0.11	83.38	1.22	0.90	0.01
Nov	55	59.04	−4.04	81.13	1.17	0.69	16.34
Dec	36	37.97	−1.97	79.74	1.12	0.46	3.87
Jan	70	68.45	1.55	81.96	1.14	0.85	2.39
Feb	64	67.34	−3.34	80.62	1.10	0.81	11.16
Mar	92	89.98	2.02	82.82	1.12	1.10	4.07
Apr	102	93.49	8.51	88.53	1.19	1.12	72.44
May	104	99.58	4.42	92.11	1.23	1.11	19.51
June	130	129.44	0.56	93.59	1.24	1.39	0.32
July	140	143.91	−3.91	93.28	1.21	1.51	15.30
Aug	115	117.04	−2.04	93.50	1.20	1.24	4.18
Sept	91	87.36	3.64	97.07	1.24	0.93	13.26
Oct	88	88.38	−0.38	98.05	1.23	0.90	0.15
Nov	73	68.81	4.19	102.91	1.29	0.70	17.57
Dec	50	46.89	3.11	108.27	1.36	0.45	9.67
Forecast		93.0		109.6	1.36	0.85	
		89.5		111.0	1.36	0.81	
		124.0		112.3	1.36	1.10	
		127.8		113.7	1.36	1.12	
		128.3		115.1	1.36	1.11	
		161.5		116.4	1.36	1.39	

22 through 25. Then, add B (the trend) back once for each period as shown in Equation 47. These values are then used in Equation 35.

$$S_1 = S_0 + B,\ S_2 = S_1 + B,\ S_3 = S_2 + \mathrm{B}, \ldots, S_{24} = S_{23} + B \tag{47}$$

b. Selecting Parameter Values Parameters can be selected through a grid search using a three dimensional grid, which can be represented on paper using a series of tables such as Table 10, one for each of several γ values.

c. Four-Models-In-One The Winters' model shown here is a slightly modified version of Holt-Winters Exponential Smoothing. One of the features of Holt-Winters (or this variate) is that it is four models in one. By setting γ to zero and initializing all the seasonal factors to 1, the model becomes Holt Exponential Smoothing. It allows the trend component of the model, but excludes the seasonality variation. By setting β to zero and initializing B to 0, the model becomes Winters Exponential Smoothing. It allows seasonal variation, but excludes trend. By setting both factors to these neutral values, the model becomes SES.

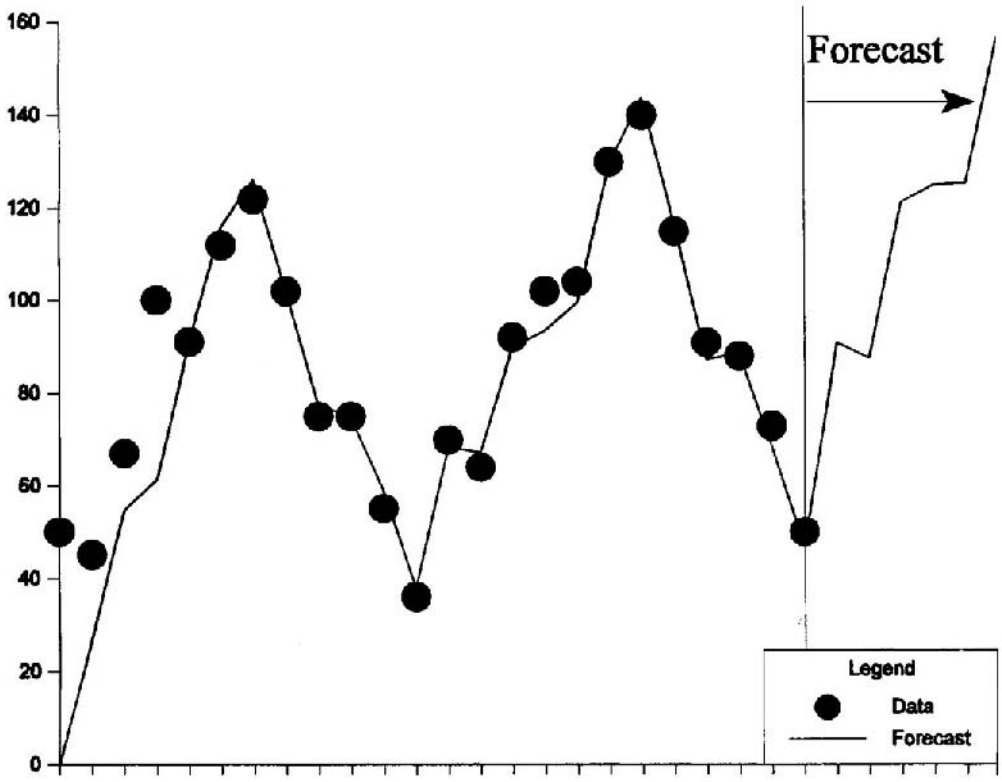


FIGURE 33 Holt-Winters exponential smoothing.

5. *Winters with Additive Seasonality*

When the seasonal pattern is additive, use a Winters' additive exponential smoothing model (Pfeffermann and Allon, 1989). The model shown below approximates the T. M. Williams' variation of the Winters model with additive seasonality.

$$F_{t+m} = \text{Forecast at time } t \text{ of time } t + \text{m} = S_t + B_t * m + I_{t+m-L} \quad (48)$$

$$F_t = \text{Forecast at time } t = S_{t-1} + B_{t-1} + I_{t-L} \quad (49)$$

$$S_t = \text{Level at time } t = S_{t-1} + B_{t-1} + \alpha e_t \quad (50)$$

$$B_t = \text{Trend at time } t = B_{t-1} + \beta e_t \quad (51)$$

$$I_t = \text{Additive seasonal Index at time } t = \gamma e_t + I_{t-L} \quad (52)$$

$$e_t = \text{Error at time } t = X_t - F_t \quad (53)$$

where X_t, α, β, γ, t, L, and m are as used with the Holt-Winters multiplicative model.

An example is shown in Table 13 using the same data as used with the Table 12. The additive factors from Table 11 are used as initial additive factors. The RMSE for the additive model is about 6% lower than for the multiplicative data. The forecast is shown in Figure 34.

a. Initialization This model should be initialized with 12 prior seasonal factors. Initial factors can be calculated using Equations 38 through 40. Use Equation 47 to calculate S_t values for Equation 38 only if necessary. Some discussion indicates that these seasonal factors should

TABLE 13 Additive Holt-Winters

				α 0.5	β 0.01	γ 0.1	
Month	Data	F_t	e	S	B	I	RMSE = 3.2
Jan	50		50.0	25.0	0.5	−14.2	
Feb	45	2.98	42.02	46.51	0.92	−18.32	
Mar	67	55.16	11.84	53.35	1.04	8.91	
Apr	100	58.39	41.61	75.19	1.45	8.16	
May	91	87.34	3.66	78.48	1.49	11.05	13.42
June	112	119.28	−7.28	76.33	1.42	38.58	53.07
July	122	122.54	−0.54	77.48	1.41	44.74	0.29
Aug	102	99.58	2.42	80.10	1.44	20.93	5.86
Sept	75	74.91	0.09	81.58	1.44	−6.62	0.01
Oct	75	73.81	1.19	83.61	1.45	−9.09	1.41
Nov	55	57.79	−2.79	83.67	1.42	−27.55	7.80
Dec	36	35.51	0.49	85.34	1.43	−49.53	0.24
Jan	70	72.60	−2.60	85.47	1.40	−14.43	6.74
Feb	64	68.55	−4.55	84.59	1.36	−18.77	20.68
Mar	92	94.86	−2.86	84.52	1.33	8.63	8.19
Apr	102	94.01	7.99	89.84	1.41	8.96	63.91
May	104	102.30	1.70	92.10	1.42	11.22	2.88
June	130	132.11	−2.11	92.47	1.40	38.37	4.43
July	140	138.61	1.39	94.57	1.42	44.88	1.93
Aug	115	116.91	−1.91	95.03	1.40	20.74	3.66
Sept	91	89.81	1.19	97.02	1.41	−6.50	1.42
Oct	88	89.34	−1.34	97.76	1.40	−9.22	1.80
Nov	73	71.61	1.39	99.85	1.41	−27.41	1.94
Dec	50	51.73	−1.73	100.40	1.39	−49.71	2.99
Forecast		87.4		101.8	1.39	−14.43	
		84.4		103.2	1.39	−18.77	
		113.2		104.6	1.39	8.63	
		114.9		106.0	1.39	8.96	
		118.6		107.4	1.39	11.22	
		147.1		108.8	1.39	38.37	

be renormalized to sum to zero with each update, at minimum, the forecaster should take care that these factors do not vary radically from this norm. Trend and level can be initialized using Equations 22 through 25.

b. Parameters and Multiple Models Parameters are selected as with the Holt-Winters multipliciative model, and, as with that model, setting any component to a neutral value eliminates the effect of that component from the model, leaving the other components intact. The neutral value for additive seasonal factors is zero.

IV. OPTIMIZING AND LOSS FUNCTIONS

Optimizing should not be limited to simple minimization of RMSE, or any loss function. Use of RMSE or other techniques has limitations which the forecaster should consider and account for before selecting a forecast model.

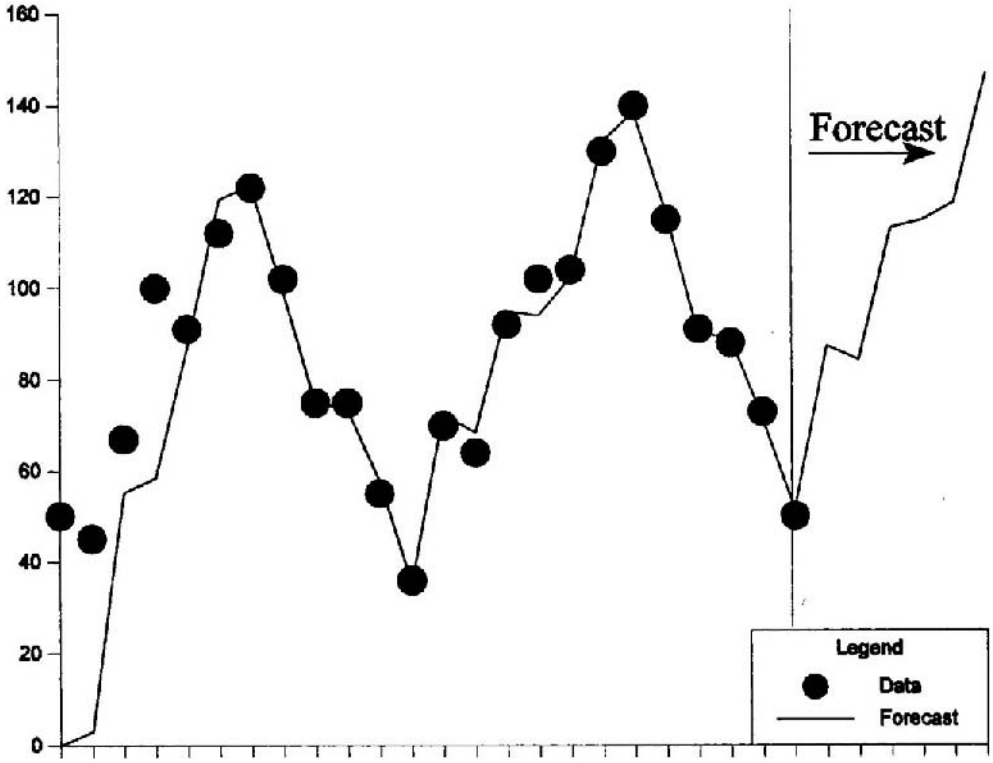


FIGURE 34 Additive Holt-Winters.

A. Limitations of Root Mean Squared Error

Early in this chapter we began to evaluate the relative accuracy of various forecast techniques through the interpretation of Root Mean Squared Error (RMSE); calculation is shown in Equation 7. RMSE is one of several commonly used loss functions whose role is to provide comparative information on forecast accuracy. For a single series RMSE is easily interpreted, a smaller RMSE implies the forecast more accurately predicts the data, a larger RMSE implies the opposite. However, there are several limitations on using RMSE:

1. Because the magnitude of RMSE depends on the magnitude of the data series, it is not useful for comparing forecast techniques among non-comparable series. Ordinarily this is a problem for forecast researchers, not forecasters. Nonetheless, a solution is to use a loss function that is independent of data series magnitude. One such loss function is the Symmetrical Mean Absolute Percent Error, SMAPE (Flores, 1986), which is the average of the Symmetrical Absolute Percent Error (SAPE). These are calculated as follows (where F, X, e, t, and n have their usual meanings in this chapter):

$$\text{SAPE}_t = |2e_t/(F_t + X_t)| \tag{54}$$

$$\text{SMAPE} = \sum_{t=1}^{n} \text{SAPE}_t/n \tag{55}$$

2. Forecasts are vulnerable to two kinds of error: inaccuracy, which is simply getting the wrong future value; and bias, which is systematically getting the wrong future value. RMSE

and SMAPE report inaccuracy. An excessive concern for reducing one of these loss functions can actually increase bias. There are two sorts of systematic wrong future values, first one might tend always to get the wrong value in a particular direction, such as always overestimating the value. A second error is tending always to get the opposite error of the previous error. The presence of either sort of systematic error suggests the forecaster might be able to improve the forecast. The tendency to always get the value wrong in a particular direction is of particular concern. If forecasts are accumulated over multiple periods, such as forecasting monthly data and summing it to annual estimates, these systematic errors augment each other, so the forecast becomes more inaccurate as more periods are added together. Unbiased forecasts should not exhibit this augmenting error.

Two statistics that help evaluate bias are mean error, ME, and the autocorrelation of errors, $\rho(e_t, e_{t-1})$. In a completely unbiased forecast, both of these statistics will be zero. Mean error (Flores, 1986) helps evaluate the tendency to get the value wrong in a particular direction:

$$\text{ME} = \sum_{t=1}^{n} e_t/n \tag{56}$$

Autocorrelation of error (Chatfield, 1978) evaluates any systematic error; $\rho(e_t, e_{t-1})$ can take on values ranging from -1 to 1. Values close to zero imply no systematic error. Negative values imply alternating errors, which might arise if the trend component of the forecast is over-responsive (β is too high in a Holt model), or if the data were inadequately decomposed before forecasting. Positive values imply successive errors with the same sign (the tendency to get the value wrong in a particular direction). This statistic is simply the Pearson's product moment correlation coefficient applied to successive errors, calculate $\rho(e_t, e_{t-1})$ as follows:

$$\bar{e} = \sum_{t=1}^{n} e_t/n \tag{57}$$

$$\sigma_{e(t)} = \sqrt{\left[\sum_{t=2}^{n} (e_t - \bar{e})^2/(n-1)\right]},\ \sigma_{e(t-1)} = \sqrt{\left[\sum_{t=1}^{n-1} (e_t - \bar{e})^2/(n-1)\right]} \tag{58}$$

$$\text{Cov}(e_t, e_{t-1}) = \sum_{t=2}^{n} [(e_t - \bar{e})(e_{t-1} - \bar{e})]/[n-1] \tag{59}$$

$$\rho(e_t, e_{t-1}) = \text{Cov}(e_t, e_{t-1})/(\sigma_{e(t)} * \sigma_{e(t-1)}) \tag{60}$$

The point of considering more than one loss function is that each loss function may be optimal at a different combination of parameter values. There is no algorithm for resolving this conflict. When different loss function suggest different optimal parameters, the analyst must choose a set of parameters based on the reasons for considering the various loss functions.

3. RMSE is vulnerable to outliers because it puts more weight on the deviations that are furthest from the forecast estimate (by squaring the deviations). One way to reduce the risk of vulnerability to outliers is to windsorize the data as discussed early in this chapter. Another way is to use median related loss functions; however, the forecaster may have considerable loss of information when looking at the median error of a time series. A third option is to consider several loss functions, including at least one bias related loss function.

4. When a loss function is calculated over the whole range of the data, it puts as much weight on accurately estimating the oldest data as it does on accurately estimating the data that is near the end of the series. However, the forecaster might be more interested in errors near the end of the data series. Also, certain errors may be particularly difficult to evaluate. For

example, when an uninitialized exponential smoothing model is used, the early errors tend to overwhelm the later errors, while when early observations are used to initialize an exponential smoothing model, the errors associated with these observations may be artificially reduced. Several approaches are available for addressing these issues. First, questionable errors (such as the first few errors in an uninitialized model) should be discarded from calculation of the loss function unless there simply are not enough observations to discard them. Second, the forecaster can concentrate special attention on the most recent errors. Summarize the loss function twice, once for all observations, and a second time for the last two years (assuming monthly data) observations. Excepting with $\rho(e_t, e_{t-1})$, the math for calculating a separate summary for the most recent data is quite easy. A third approach is to calculate weighted statistics. Again, the math is not difficult except with $\rho(e_t, e_{t-1})$: the error term (e^2 for RMSE, SAPE for SMAPE, or e for ME) of the most recent observation is assigned a weight of 1, and each prior error term is assigned some smoothly declining smaller weight. The more rapidly the weights decline, the more the focus of the loss function is shifted to recent observations. To calculate the weighted statistic, multiply the error term by the weight before summing, then divide by the sum of the weights rather than by n to calculate the mean error term. This is illustrated with mean deviation (where ω_t is the weight for the observation at time t):

$$\text{ME} = \sum_{t=1}^{n} (\omega_t * e_t) \Big/ \sum_{t=1}^{n} = \omega_t \tag{61}$$

5. An examination of the use of RMSE shows that it is calculated for the one-period-ahead forecast. The forecast techniques adjust with each new observation; while this improves the forecast, it limits the number of observations that are available for evaluating any particular projection. In fact, only the projection to the next period is available for evaluation, projections into later periods reflect additional updates by the time the actual data are available for comparison. So, the loss function evaluates only how well the forecast projects ahead a single period. Projection into subsequent periods is taken on faith. Two options are available to forecasters:

a. Hold out data from the forecast model for evaluation (Armstrong, 1985). Instead of optimizing the parameters with all data entered into the calculation, hold out the last year's data (or however much data are associated with the furthest horizon to be projected) when estimating parameters. Project through the hold out period. Then compare the results with the actual data held out. If the forecast model makes a satisfactory projection of the hold out period, add these data to the calculations without updating the parameters. Then make a new forecast of the future period. Obviously, this approach would be difficult to implement with every forecast update; however, this approach may be appropriate when preparing particularly critical forecasts. It may also be worthwhile to systematically evaluate all forecasts using this approach on an annual schedule.
b. A second option is to deemphasize the use of loss functions. In Figure 35 we see the SES of the first differences forecast compared with the Holt forecast as previously shown in Table 9. The Holt forecast is previously demonstrated in Figure 19. The comparative RMSE values for Holt exponential smoothing and SES of the first differences are 5.3 and 5.1 respectively. By implication, the SES model produces the better forecast. However, since these data are artificial, it can be known with certainty that the true growth rate of these data is approximately 1.3 units per period. The SES

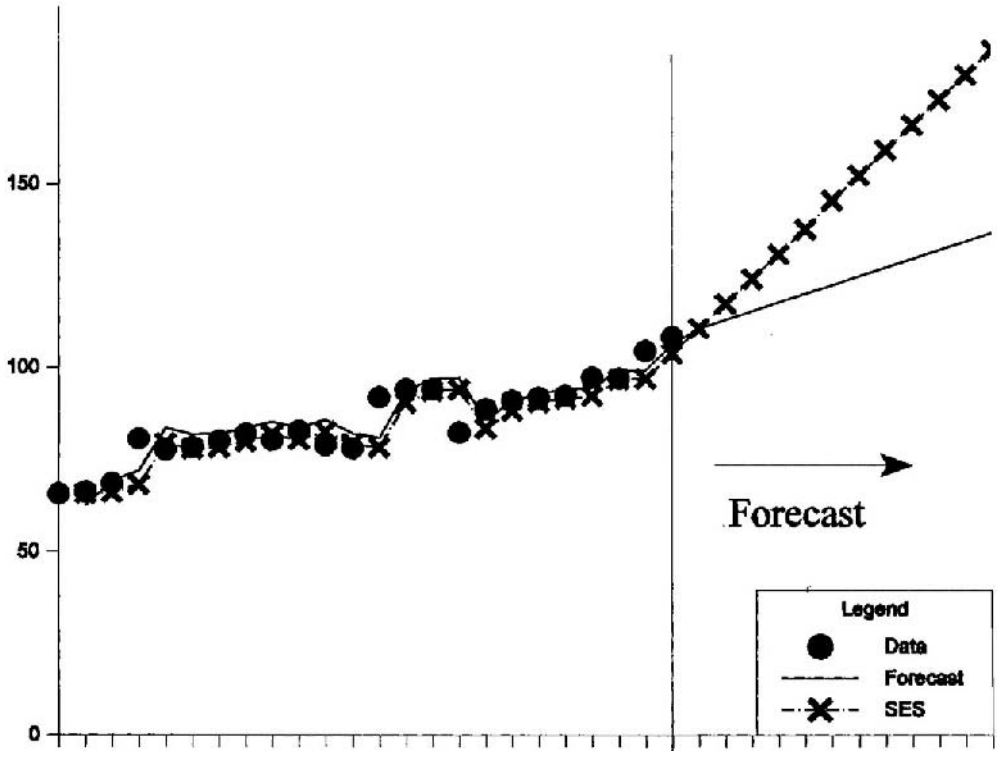


FIGURE 35 Comparison: Holt vs. SES of first differences.

model projects growth of 6.8 units per period at the end of the series, while the Holt model projects 2.4 units per period at the end of the series. Further the SES estimate of the trend ranges from −10.6 to 12.3, while the Holt estimate ranges from 2.0 to 3.9 (except following a few upshifts in the data, the Holt model gradually decreases the estimate of trend). The SES model of the differences rapidly responds to variation in the data, but this response leads to one-period-ahead forecast success, hence the slightly lower RMSE. If the forecaster is interested in projecting only through the next period, he or she should select the SES model based on the RMSE results. However, if the forecaster is interested in multiple horizons, he or she can balance the information provided by the loss function with other information. In particular, he or she might notice that the trend estimated through the SES forecast of differences frequently changes by relatively large amounts and ask how he or she can to rely on a forecast that will likely be considerably different with just a few more updates. By balancing such common sense considerations against the one-period-ahead loss function evaluation, the forecaster can improve his or her chances of making a reasonable forecast.

B. Rules of Thumb

While the optimization approach assumes that exponential smoothing parameters should be fit to the historical data, some forecasters argue otherwise. For example, Armstrong argues that

there is no evidence that fitting the historical data produces better forecasts. He recommends the forecaster use judgment in selecting parameters. It is likely that the best approach is a balance between optimizing with historical data and use of rules of thumb such as these: (1) Armstrong (1985) recommends high α values when the process that generates the data is unstable, low α values when there is a high risk of measurement errors, and decreasing α values for shorter cumulation periods (months are short, years are long). While the first two of these recommendations appear reasonable, the rationale for the third is less clear. (2) It is my experience that β values should be extremely low, particularly when using the variant of Holt exponential smoothing shown in this chapter. High β values (even relatively small β values, such as 0.2 might be thought to be relatively high) lead to unstable forecasts.

V. COMBINING FORECASTS

There is evidence that forecasts can be improved through the simple averaging of equally reasonable forecasts made with different techniques. (Makridakis and Winkler, 1983; Clemen, 1989) Averaging of forecasts is not difficult. It consists of making a forecast through each of several reasonable methods, and then averaging the results. An example of two forecasts made with reasonable methods might be the two forecasts in Table 9, with a lower α value for the SES forecast.

VI. UPDATING AND MONITORING

So far, I have focused on making an initial forecast. The steps for making such a forecast include collecting data, analyzing and decomposing it, selecting a model, optimizing the model, making projections, and possibly averaging those projected values. Sometimes a forecaster can then make a report and move on to his next project. However, in many situations, this is only the beginning of a forecasting effort. After completing the initial forecast, the forecaster finds that the data generating process generates more data, which leads to the need for updating and monitoring.

A. Updating

Updating is adding one or more new observations to the historical data to get a new forecast. An advantage of simple forecast techniques is that they allow relatively easy updating. Updating does not require re-optimizing the forecast model. Most analysts re-optimize only if they suspect that the forecast model is no longer fit.

Armstrong (1985) says that frequent updating is important for accuracy, but, uncharacteristically, he provides no evidence to support this claim. Nevertheless, this view makes sense. The frequency of the updates depends on several factors. The forecast should be updated at least as frequently as required for users. Also, it cannot be updated more frequently than data become available. If the process for updating is not labor intensive, it might be best to update as often as possible.

Updating produces a new forecast. However, the new forecast may not differ significantly from the old forecast. Forecasts, like other statistical estimates, are contained in confidence intervals. For most of these techniques, production of confidence intervals is complex, and there is evidence they are unreliable (Makridakis et al., 1987). One of the advantages of frequent

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

updating is that it concretely demonstrates the uncertainty of the forecast. While the most recent forecast update becomes the best estimator of the future values, it is also uncertain. The forecaster should not feel obliged to interject repeated minor changes to forecasts into complex decision making processes, such as legislative sessions. Instead, by demonstrating the history of updating, the forecaster can provide the forecast user with useful information on how certain the forecast is. Once a forecast is used in formal decision making setting, the update might be thought of as part of the monitoring process, useful for asking whether the estimate is so changed as to require interrupting the decision-making with new estimates.

B. Monitoring

Three forms of monitoring are shown here, tracking signals, the Wineglass graph, and the Outlook graph.

1. Tracking Signals

The simplest method of monitoring a forecast (other than just updating) is to use a tracking signal while updating. A tracking signal is a statistic that alerts the forecaster that the forecast is out of control, that is, that the model fit needs reevaluation. A typical tracking signal is the smoothed error tracking signal (McClain, 1988).

$$\mathrm{MAD}_t = \delta|e_t| + (1 - \delta)\ \mathrm{MAD}_{t-1} \tag{62}$$

$$E_t = \theta e_t + (1 - \theta)E_{t-1} \tag{63}$$

$$\mathrm{SE}_t = |E_t/\mathrm{MAD}_t| \tag{64}$$

Where, MAD = Mean Absolute Deviation, MAD_t = a smoothed estimate of MAD for time period t, e = error, E_t = a smoothed estimate of error at time period t, δ and θ = smoothing constants for MAD_t and E_t, respectively, and SE_t = Smoothed Error at time period t.

The values for δ and θ are set arbitrarily, but δ should be set low, such as $\delta = 0.05$. A possible equation for the determination of θ is $\theta = 1/(N(1 + 3\alpha))$ (Equation 65), where N is the number of time periods the forecaster is willing to risk overlooking an out-of-control situation and α is the level parameter of an exponential smoothing model. This equation suggests that δ should be higher for highly variable data than for less variable data. The forecaster watches for an increase, particularly a sharp increase, in SE_t which signals a need to fit the forecast model through a new grid search.

2. Wineglass

Assume that an analyst made a monthly forecast for the current year. The year has now begun and the analyst wants to know how well the forecast is holding up. Wineglass (Wu et al., 1992) is a tracking tool that graphically demonstrates forecast accuracy under such circumstances. It depends on having several types of information: (1) the availability of monthly forecasts for the year, and (2) access to the previous year's monthly data series for non-seasonal data, or at least two previous year's information for seasonal data. For non-trending non-seasonal data the Wineglass graph can be produced through these equations:

$$X_{i,j} = \text{the actual monthly observation for month i in year } j. \tag{65}$$

$$X_{*,j} = \text{the sum of actual monthly observations in year } j = \sum_{i=1}^{12} X_{i,j} \tag{66}$$

$F_{i,j}$ = the forecast for month i in year j. For non-trending non-seasonal data, in historic years estimate $F_{i,j} = (1/12)X_{*,j}$, [Eq. 67] for the forecast year, $F_{i,j}$ should be the actual forecast. (67)

$$F_{*,j} = \text{the sum of the monthly forecasts in year } j = \sum_{i=1}^{12} F_{i,j} \tag{68}$$

I = a subscript for the Ith month.

$g_{i,j}$ = ratio of total actual to total forecast year to date = for month i in year j,

$$\sum_{i=1}^{I} X_{i,j} \Big/ \sum_{i=1}^{I} F_{i,j} \tag{69}$$

$$g_{*,j} = X_{*,j}/F_{*,j} \tag{70}$$

$$\omega_{*,j}^2 = 1/11 \sum_{i=1}^{12} [(X_{i,j} - F_{i,j})^2/(F_{i,j}F_{*,j})] \tag{71}$$

$$VW_{I,j+1} + 1 = \omega_{*,j}^2 \left(\sum_{i=I+1}^{12} F_i \Big/ \sum_{i=1}^{I} F_i \right) \tag{72}$$

ξ = Z value of the normal distribution for 1/2 the area associated with a confidence interval, e.g., ci = 80%, ξ = 1.282, smaller values imply lower tolerance for error.

Upper_1, Lower_1 = Bound for the period I (upper and lower) $= 1 \pm \xi\sqrt{(\text{VW}_1)}$ (73)

The calculation of these equations are shown in Table 14 (the forecast is artificial).

As shown in Figure 36, the Wineglass chart shows the ratio of year-to-date actual-to-forecast (g_i) along with tolerance boundaries that demonstrate whether the forecast is ''on track,'' that is, the likelihood of attaining a cumulative value consistent with the original forecast within the tolerance level. If g_i is outside the tolerance boundaries as in Figure 36, the apparent conclusion is that the forecast will not be attained within the tolerance level. There will be more error in the ultimate actual number than the forecast user is willing to tolerate. Calculation of the Wineglass graph as shown here applies only to non-trending non-seasonal data. For other data equation 67 must be replaced with actual historical forecasts and equation 71 must be revised to follow Wu et al.'s (1992) Equation 5.

3. *Outlook*

Given year to date experience, what does the whole year's data series look like now? A graph that handles this is the outlook graph (Wu et al., 1992). Beginning with the Wineglass variance and g_i ratios, outlook can be calculated as follows:

ξ = the outlook confidence level.

$$F_I^M = g_i F_{*,j} = \text{the medium outlook forecast as of period I} \tag{74}$$

$$VO_I = g_i^2 \text{VW}_I F_{*,j}^2 \tag{75}$$

TABLE 14 Wineglass

Prior year	X_i	$X_{*,j}$	F_i	$F_{*,j}$	$(X_i - F_i)$	$(X_i - F_i)^2$	$F_{i,j} * F_{*,j}$	$(X_i - F_i)^2/F_{ij} * F_{*,j}$	Sum last col	11	ω^2
Jan-92	1067		1089.9		−22.9	525.2	14255020	0.000			
Feb-92	983		1089.9		−106.9	11431.2	14255020	0.001			
Mar-92	909		1089.9		−180.9	32730.8	14255020	0.002			
Apr-92	931		1089.9		−158.9	25254.5	14255020	0.002			
May-92	1511		1089.9		421.1	177311.2	14255020	0.012			
Jun-92	1197		1089.9		107.1	11466.8	14255020	0.001			
Jul-92	1053		1089.9		−36.9	1362.8	14255020	0.000			
Aug-92	1160		1089.9		70.1	4911.7	14255020	0.000			
Sep-92	871		1089.9		−218.9	47924.5	14255020	0.003			
Oct-92	1057.5		1089.9		−32.4	1050.8	14255020	0.000			
Nov-92	1057.5		1089.9		−32.4	1050.8	14255020	0.000			
Dec-92	1282	13079	1089.9	13079	192.1	36896.0	14255020	0.003	0.0247	11	0.0022

The contribution of the preceding calculations is the resulting number $\omega^2 = 0.0022$. *Continuing*:

Forecast year	X_i	$\sum_{i=1}^{I} X_{i,j}$	F_i	$\sum_{i=1+1}^{I} F_{i,j}$	g_i	$\sum_{I=1+1}^{12} F_i$	$\sum_{I=1}^{12}/\sum_{i=1}^{I} F_i$	VW_I	$\sqrt{VW_I}$	ξ	Upper	Lower
Jan-93	1468	1468	1450	1450	1.012	15950	11.00	2.4%	15.6%	1.282	1.199	0.801
Feb-93	988	2456	1450	2900	0.847	14500	5.00	1.1%	10.5%		1.134	0.866
Mar-93	950	3406	1450	4350	0.783	13050	3.00	0.7%	8.1%		1.104	0.896
Apr-93	1009	4415	1450	5800	0.761	11600	2.00	0.4%	6.6%		1.085	0.915
May-93	1484	5899	1450	7250	0.814	10150	1.40	0.3%	5.6%		1.071	0.929
Jun-93	1042	6941	1450	8700	0.798	8700	1.00	0.2%	4.7%		1.06	0.94
Jul-93	1133	8074	1450	10150	0.795	7250	0.71	0.2%	4.0%		1.051	0.949
Aug-93	1183	9257	1450	11600	0.798	5800	0.50	0.1%	3.3%		1.043	0.957
Sep-93	1278	10535	1450	13050	0.807	4350	0.33	0.1%	2.7%		1.035	0.965
Oct-93			1450	14500		2900	0.20	0.0%	2.1%		1.027	0.973
Nov-93			1450	15950		1450	0.09	0.0%	1.4%		1.018	0.982
Dec-93			1450	17400								

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

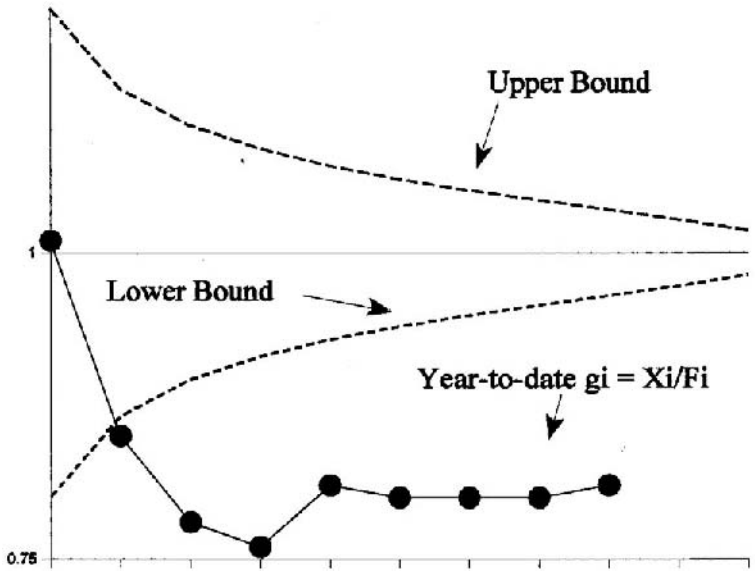


FIGURE 36 Wineglass.

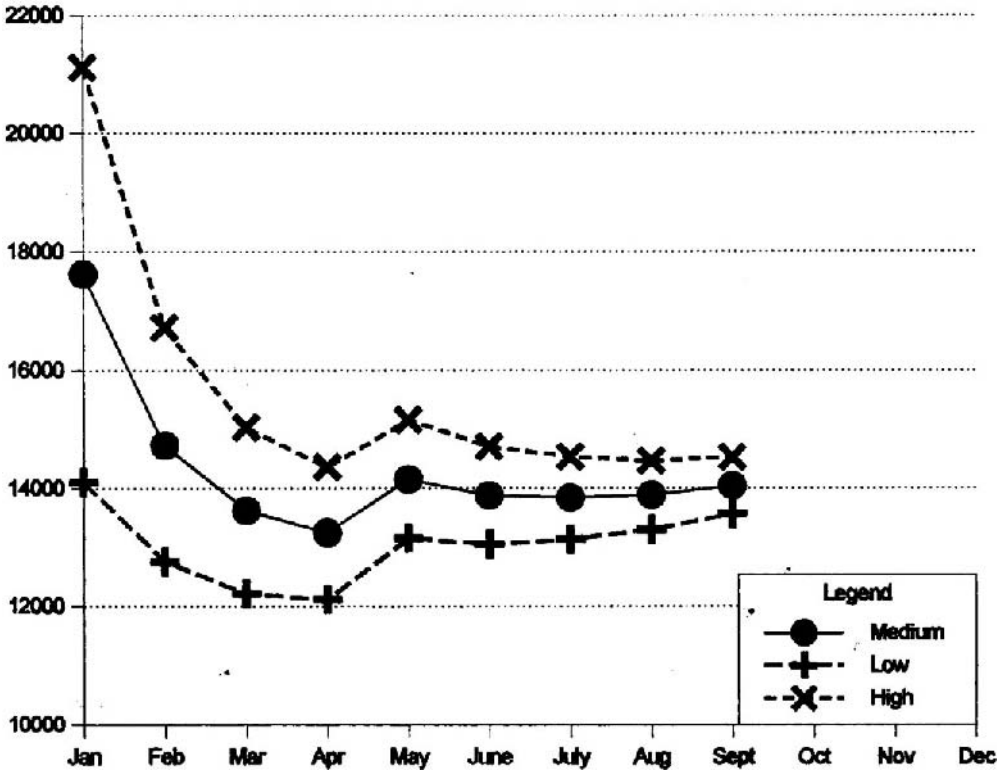


FIGURE 37 Outlook.

TABLE 15 Outlook

Forecast year	g_i	$F_{*,j}$	VW_I	$VO_I = g_i^2 VW_1 F_{*j}^2$	$\sqrt{VO_I}$	$\frac{F_I^M}{17400}$	ξ	F_I^L	F_I^H
Jan-93	1.012	17400	2.4%	7509828	2740.4	17616	1.282	14103	21129
Feb-93	0.847		1.1%	2388647	1545.5	14736		12755	16717
Mar-93	0.783		0.7%	1225048	1106.8	13624		12205	15043
Apr-93	0.761		0.4%	771892	878.57	13245		12119	14371
May-93	0.814		0.3%	617348	785.71	14158		13150	15165
Jun-93	0.798		0.2%	423962	651.12	13882		13047	14717
Jul-93	0.795		0.2%	301050	548.68	13841		13138	14545
Aug-93	0.798		0.1%	212088	460.53	13886		13295	14476
Sep-93	0.807		0.1%	144693	380.39	14047		13559	14534

$$F_I^H = F_I^M + \xi\sqrt{VO_I} = \text{the high outlook as of period I} \tag{76}$$

$$F_I^L = F_I^M - \xi\sqrt{VO_I} = \text{the low outlook as of period I} \tag{77}$$

Calculations are shown in Table 15. The resulting graph is shown in Figure 37.

Above each month on the X axis, the outlook graph reports low, medium and high estimates of the annual forecast. As the unknown portion of the year diminishes, the range between low and high diminishes. This graph reports not only how confident the forecaster remains in the prior forecast, but also what the current likely forecast would be.

These last two graphs have the advantage of communicating complex information graphically such that it is easily interpreted. Disadvantages include that they are somewhat complex to establish, the choice of ξ is somewhat arbitrary, and, as designed, they work only with monthly level data in forecasts that cumulate to years and only once the year has begun. A reasonable forecast monitoring strategy could be to use frequent updating with a tracking signal as the primary source of monitoring prior to the forecast year, then to supplement this strategy with Wineglass and Outlook once the forecast year has begun. For other sorts of data, focus forecast monitoring efforts on updating and tracking signals.

VII. SUMMARY

A recurrent theme throughout this chapter is that the forecaster should know his data. For example, decomposition depends on forecaster knowledge of the data generating process. Decomposition also takes advantage of any prior knowledge the forecaster has of any component series. Finally, through decomposition, the forecaster comes to better understand both the data generating process and many of the component processes. Preparing the data for forecasting consists of analyzing the data and sorting out its variability, and forecasting consists, primarily, of taking advantage of that knowledge. Finally, updating and monitoring brings the forecaster into constant contact with the data series whereby he learns to anticipate how and when it changes. For effective forecasting, this familiarity with the data is the most important tool the forecaster can bring to the forecast.

ACKNOWLEDGMENTS

I would like to thank Don Miller and the editors of this book for their useful suggestions concerning this chapter. Any errors are my own responsibility.

REFERENCES

Ammons, D.N. (1991). *Administrative Analysis for Local Government: Practical Applications of Selected Techniques*, Carl Vinson Institute of Government, Athens, Georgia, pp. 68–73.

Armstrong, J.S. (1985). *Long-Range Forecasting, From Crystal Ball to Computer*, 2d ed., John Wiley & Sons, New York.

Ashley, R. (1988). ''On the Relative Worth of Recent Macro-economic Forecasts,'' *International Journal of Forecasting*, *4*: 363–376.

Ashley, R. (1983). ''On the Usefulness of Macroeconomic Forecasts as Inputs to Forecasting Models,'' *Journal of Forecasting*, *2*: 211–223.

Chatfield, C. (1978). ''The Holt-Winters Forecasting Procedure,'' *Applied Statistics*, *27*: 264–279.

Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*: 559–583.

Flores, B.E. ''A Pragmatic View of Accuracy Measurement in Forecasting,'' *Omega*, *14*: 93–98 (1986).

Gardner, Jr., E. and E. McKenzie (1985). ''Forecasting Trends in Time Series,'' *Management Science*, *31*: 237–1245. (1985).

Makridakis, S. and M. Hibon (1991). ''Exponential smoothing: The Effect of Initial Values and Loss Functions on Post-sample Forecasting Accuracy,'' *International Journal of Forecasting*, *7*: 317–330.

Makridakis, S., S.C. Wheelwright, and V.E. McGee (1983). *Forecasting: Methods and Applications*, 3rd ed., John Wiley & Sons, New York.

Makridakis, S., A. Anderson, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982). ''The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition,'' *Journal of Forecasting*, *1*: 111–153.

Makridakis, S., M. Hibon, E. Lusk, and M. Belhadjali (1987). ''Confidence Intervals an Empirical Investigation of the Series in the M-competition,'' *International Journal of Forecasting*, *3*: 489–508.

Makridakis, S. and R.L. Winkler (1983). ''Averages of Forecasts: Some Empirical Results,'' *Management Science*, *29*: 987–996.

McClain, J.O. (1988). ''Dominant Tracking Signals,'' *International Journal of Forecasting*, *4*: 563–572.

Pfeffermann, D. and J. Allon (1989). ''Multivariate Exponential Smoothing: Method and Practice,'' *International Journal of Forecasting*, *5*: 83–98.

Vollmann, T.E., W.L. Berry, and D.C. Whybark (1993). *Integrated Production and Inventory Management: Revitalizing the Manufacturing Enterprise*. Business One Irwin, Homewood, Illinois, p. 71.

Williams, T.M. (1987). ''Adaptive Holt-Winters Forecasting,'' *Journal of the Operational Research Society*, *38*: 553–560.

Winters, P.R. (1960). ''Forecasting Sales by Exponentially Weighted Moving Averages,'' *Management Science*, *7*: 324–342.

Wu, L. S.-Y., J.R.M. Hosking and J.M. Doll (1992). ''Business Planning Under Uncertainty,'' *International Journal of Forecasting*, *8*: 545–557.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

16

Demographic Techniques for Cohort Analysis and Population Trends

Deirdre M. Mageean
University of Maine, Orono, Maine

I. INTRODUCTION

The need for the public administrator to be demographically literate is greater today than ever before. Informed decision making at many levels requires a knowledge of demographic trends and influences in society. This is as true for the local town manager or planner as it is for those employed at the state or federal level. Areas such as housing, labor force participation, school planning, health service delivery, support services for the elderly, transportation, emergency planning and political redistricting all entail demographic analysis. The sub-discipline of demography which primarily deals with such analysis is known as applied demography. Applied demography focuses on the use of demographic data, methods and perspectives to facilitate decision making regarding practical problems. These problems tend to arise in the realms of business and government, particularly at the state and local level. Although demographic analysis is a very important part of private sector analysis it has a longer history in the public sector where there is a long tradition of using demographic and closely related socioeconomic information for the purposes of analysis, planning and reporting (for a recent and excellent example of work in this area see the collection of case studies in Kinter et al., 1994).

In this chapter we focus on the fundamental and commonly used tools in the areas of population composition, trends, estimation, projection, and cohort analysis. We also examine the basic data sources used most extensively in such analysis. As products from the Census Bureau have become more accessible, primarily on C.D.s, demographic data are now available to the general public and the smallest towns. A working knowledge of these sources and techniques gives the public administrator essential tools for decision-making.

II. DATA SOURCES

The main sources which the public administrator is likely to draw on are the Census and vital statistics. Others which may be used as supplementary data are tax records, school enrollments, car registration data, and housing starts. These latter are sometimes referred to as symptomatic data, a term used to describe variables which reflect change in population size, and are frequently used in population estimation and projection.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

A. Census Statistics

The 1990 Census of Population and Housing was the twenty-first decennial enumeration of the United States. Since 1790, the date of the first census, the constitution has provided for these enumerations in order to reapportion the House of Representatives. The modern census can be said to have begun in 1940 with the incorporation of the housing component and the introduction of sampling techniques to supplement information obtained from the long-form questionnaire which is sent to every household. The official date of enumeration is April 1st and the basis of enumeration is the ''usual place of residence,'' usual in that a person resides there most of the year. Hence, college students are enumerated in their dorms or apartments, not their parental home, because they reside in their college most of the year.

In essence the census provides a statistical portrait or snapshot of the nation. The amount and complexity of information contained in the census has changed over the years as demand for information has increased but it continues to provide the one comprehensive, detailed and most reliable source of information for decision makers. With these rich data we can identify the key characteristics (size, composition, and growth) of areas large and small and see how they are changing. We can investigate four main areas—housing, households, population, and

TABLE 1 Short-Form Questionnaire and Portions of the Long-Form Questionnaire Subjects

Population	Housing
Household	Number of units in structure
Sex	Number of rooms in unit
Race	Tenure (owned or rented)
Age	Value of home or monthly rent paid
Marital status	Congregate housing (meals included in rent)
Hispanic origin	
The following additional subject items appear on the ''long-form'' or ''sample'' questionnaire	
Social characteristics:	Condominium status
Education (enrollment and attainment)	Plumbing and kitchen facilities
Place of birth, citizenship, and year of entry to the United States	Telephone in unit
	House heating fuel
Ancestry	Source of water and method of sewage disposal
Language spoken at home	Vehicles available
Migration (residence in 1985)	Year structure built
Disability	Year moved into residence
Fertility	Number of bedrooms
Veteran status	Farm residence
	Shelter costs, including utilities
Economic characteristics:	
Labor force	
Occupation, industry, and class of worker	
Place of work and journey to work	
Work experience in 1989	
Income in 1989	
Year last worked	

Source: U.S. Census, 1990 Census of Population and Housing Tabulation and Publication Program, July 1989.

economic characteristics. Additionally we can examine such topics as transportation and schooling.

Census information is obtained from two questionnaires—a short form which is sent to the majority of households and a long-form questionnaire which is sent only to a sample of the population. The questions on the long form contain the questions from the short from and a set of larger questions which supplement that information. Table 1 shows the subject coverage of the census and distinguishes population items from housing items and 100 percent items from sample items.

The modern census is essentially two censuses conducted concurrently. The first, a census of population, covers demographic, economic, and social characteristics of individuals and households while the census of housing collects information on equipment, financial and structural characteristics of housing units. The use of the sample long form questionnaire provides a cost-effective means of expanding the information collected by the census. In 1990 the long form was sent to approximately one out of every six households in the nation, providing a sample of approximately 16–17 percent. The sampling fraction is sufficient to provide good statistical reliability, even for small geographic areas. However, as a sample the data have a certain amount of sampling error. Most users of census data ignore the sampling error and usually this is not a problem. However, the informed user may wish to be informed of the level of accuracy surrounding sample data especially if important decisions are to be based on the information. Hence, the notions of sampling error and statistical significance (discussed in previous chapters) are important in demographic information. Each published report from the Census Bureau (both the decennial census and the Current Population Reports conducted in the interdecadal period) contains an appendix which explains how to calculate the standard error and confidence levels for the data.

The enormous amount of census information is available at many levels of geographical unit from regions, states, counties and municipalities to block groups and census tracts. Not all information is available at all levels. In general the more detailed data are available for areas higher in the geographic hierarchy. The informed user should be aware of the basic hierarchy of census geography as well as the file structure based on that geography. An excellent overview can be found in Myers (1992).

III. THE CURRENT POPULATION SURVEY

Comprehensive and rich though the census may be the information can quickly become outdated. To provide for the need for up to date and continuous data the Census Bureau conducts the Current Population Survey (CPS) which obtains information from about 66,000 households every month. The CPS covers a wide range of subjects and is particularly useful for information on employment, the labor force and income, migration and school enrollment trends. It is, however, more limited in use. Although the sample size is sufficient to provide reliable information at the national and regional levels, as well as for some large states it is of no use for smaller geographical units such as counties, places and minor civil divisions. Like the decennial census the reports are now available on CD as well as tape and printed copies.

A. Vital Statistics

The National Vital Registration System provides for the registration of five vital events—birth, death, fetal death (stillbirth), marriage and divorce. Not all events are registered in all states

but most State offices of vital statistics maintain and produce summary tabulations for the state and counties, cities and towns within their jurisdiction. At the national level the National Center for Health Statistics administers the National Vital Registration System and publishes the annual *Vital Statistics of The United States.*

The data obtained from these respective sources provide the core information used in demographic analysis. To these data are applied the techniques of analysis of population composition and change, population estimation and projection and cohort analysis. It is to these techniques that we now turn.

IV. AGE AND SEX COMPOSITION

We begin with what might be termed the staple elements of population analysis—age and sex composition. These, along with the elements of population change, births, deaths, and migration, are the basis of the more sophisticated techniques of estimation and projection.

Every population has a different age and sex composition and the manner in which population is distributed is among the most useful and revealing of population data. Such distribution, i.e. the proportion of males and females, young and old, can have considerable implications for a populations's socioeconomic needs. Some populations are young, for instance a suburb with young families or a college town, while others, such as a retirement community in Florida or Arizona, will be old. Consequently these populations will have very different needs for schooling, medical services and shops. They will have different proportions of their population which are economically active and will have different recreational and crime profiles. The simplest representation of population distribution is the population pyramid, or age-sex pyramid, which displays the proportions of males and females in each year group (typically five years). The sum of all the age-sex groups equals 100 percent of the population. It is called a pyramid because of its frequently triangular shape with larger numbers of young people at the bottom and fewer people among older age cohorts as mortality takes its toll. In reality the classic pyramid shape is now seen mainly in developing countries which still have relatively high fertility. Developed countries such as the United States more typically display the missile shape in Figure 1 which reflects lower fertility levels and greater longevity among its population.

Over representation of age groups or gender will result in distortions in the pyramid which can be clearly seen in Figure 2 which represents the college town of Orono, Maine, home of the University of Maine. The large number of university students resident in dorms and apartments in the town affect the population profile. As can be seen in Table 2 and Figure 2 the population has a disproportionately large percentage of persons in the 15–24 age group. The large student body, the majority of whom are male, also affects the sex ratio of the population. The sex ratio is the ratio of males to females in a given population and is usually expressed as the number of males for every 100 females. Finally, the large cohort of young people shifts the median age of the population downward. With a median age of 21.5 the town is considerably ''younger'' than neighboring communities where the median age is 32, closer to the 1990 U.S. national average of 32.6 years. Conversely a town with a significant retirement community would have a median age higher than the national average. In 1995 the median age of the U.S. population was 34.3, the highest ever recorded and is projected to rise to 35.7 in 2000 and peak at 38.5 in 2035. This increase in median age is driven by the aging of the cohort of births between the years 1946–1964, known as the baby boomers. The term cohort refers to a group of people sharing a common demographic experience who are observed through time. Demographers use birth, marriage and education cohorts for analytic purposes with the most commonly used cohort being the birth cohort. A birth cohort refers to people born in the same year or

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

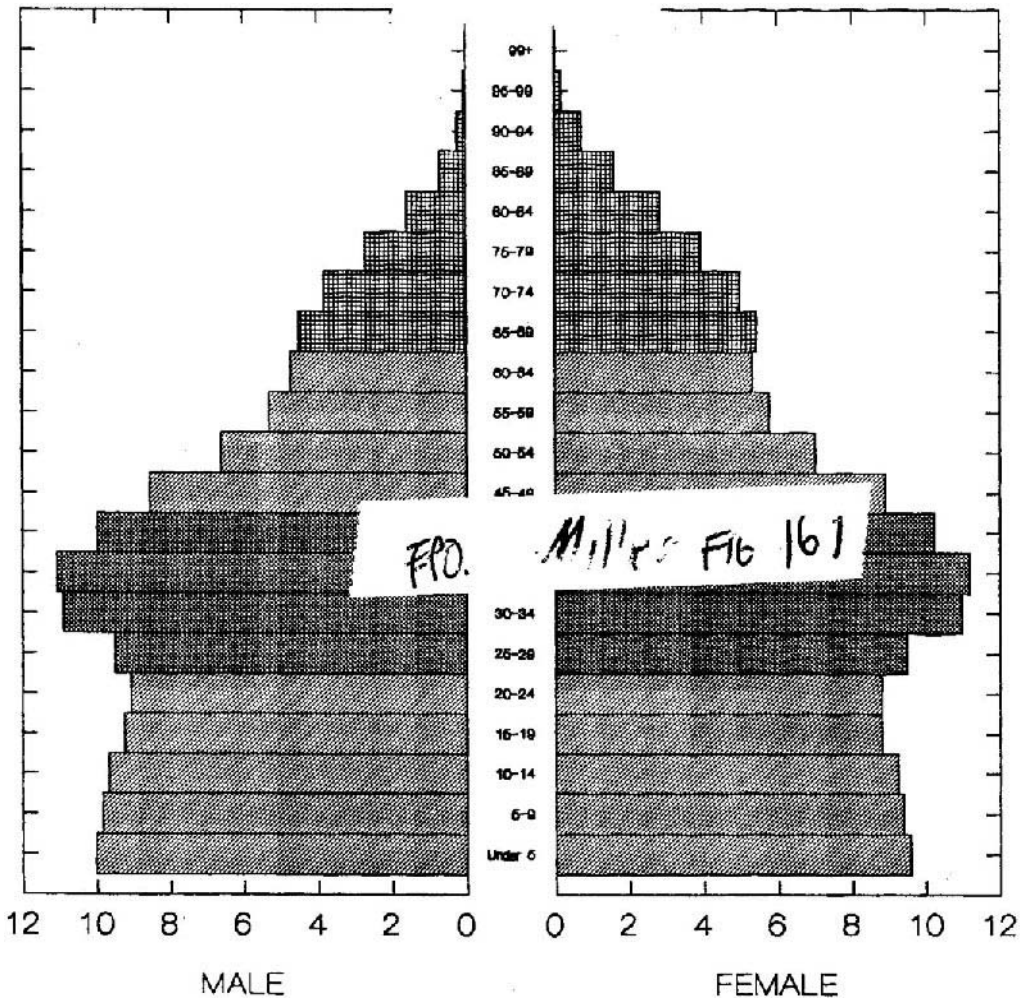


FIGURE 1 Population Pyramid of the U.S. resident population, 1995. The shaded areas display the baby boom cohort and the cohort of elderly (65+). From U.S. Bureau of the Census, Population Paper Listings PPL -41, reproduced in *Statistical Abstract of the United States, 1996* (116th edition). Washington, D.C. 1996.

period, such as the baby boomers who were born during the years 1946 through 1964. Cohorts are easily identified in population pyramids such as Figure 1 where we can clearly see the baby boom cohort and the cohort of elderly, i.e. those 65 years and older. Cohort analysis, a tool in strategic planning which is discussed later in this chapter, refers to the identification of cohorts and the study of their behaviors, attitudes and characteristics across time.

One final tool in the analysis of population distribution is the dependency ratio. This is the ratio of persons in the ''dependent'' or non-working ages (variously defined as those under 15 or 17 and those 65 years and over) to those in the working or economically active years of 18–64. It is a convenient but crude indicator of the economic burden which the economically active population must carry. Separate figures can be computed for the youth dependency ratio and the elderly dependency ratio as shown in Table 3. These figures indicate that the dependency

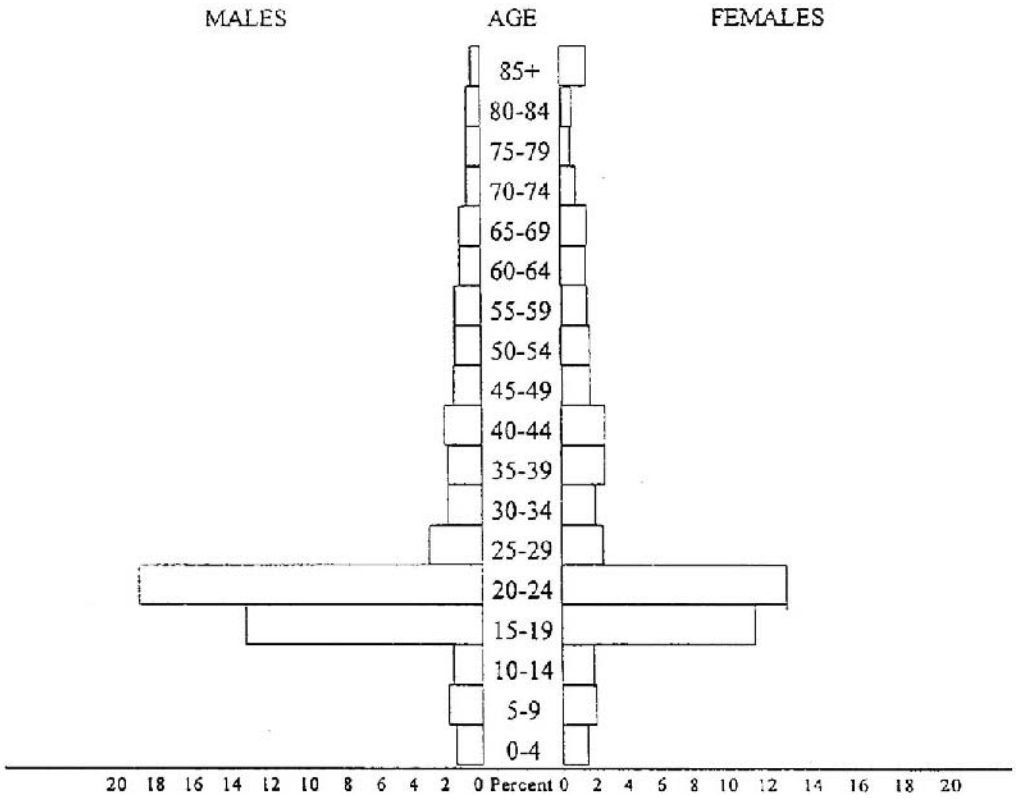


FIGURE 2 Population Pyramid of the resident population of the Town of Orono, Maine, 1990. From U.S. Bureau of the Census, 1990 Census, Summary Tape File 3A.

ratio would fall initially from its 1995 level of 63.7 to 60.2 in 2010. Then, as the baby boomers begin to reach 65, the ratio is projected to increase to 68.2 by 2020, 78.7 by 2030, and 79.9 by 2050. The elderly dependency ratio is projected to rise from its 1995 level of 20.9 (currently an all-time high) to 35.7 by 2030. At no time, however, through 2050 would the overall dependency ratio be as high as that in the 1960s with its large number of babies born during the baby boom. These ratios can be conducted at all geographical levels for which age is provided.

A. The Elements of Change in Populations

Population change has three components: births, deaths, and migration. In developed industrial societies where death rates are very stable change is largely brought about by changes in the birth rate or natural increase and migration. For small areas and over short periods of time migration is the most important component and can be a powerful force in local demographic change. Demographers measure most of these vital events in rates, that is the number of events expressed per 1000 population in a given year. For instance, the death rate (also known as the crude death rate) for the United States in 1990 was measured as:

$$\frac{\text{Number of deaths}}{\text{Total population}} \times \text{K} = \frac{2{,}162{,}000}{250{,}885{,}000} \times 1000 = 8.6$$

TABLE 2 Population by Age and Sex, Orono, 1990

Frequencies	Male	Female	Percentages	Male	Female
Under 1 year	27	21	Under 5 years	1.3	1
1 and 2 years	57	46	5 to 9 years	1.8	1.7
3 and 4 years	52	40	10 to 14 years	1.7	1.5
5 years	34	51	15 to 19 years	12.6	11.8
6 years	42	43	20 to 24 years	18.6	13.2
7 to 9 years	115	87	25 to 29 years	3.1	2.1
10 and 11 years	67	73	30 to 34 years	1.8	1.9
12 and 13 years	81	56	35 to 39 years	1.8	2.2
14 years	34	32	40 to 44 years	1.9	2.1
15 years	26	25	45 to 49 years	1.5	1.5
16 years	31	39	50 to 54 years	1.3	1.6
17 years	32	35	55 to 59 years	1.3	1.3
18 years	389	386	60 to 64 years	0.9	1.2
19 years	849	762	65 to 69 years	1	1.2
20 years	702	598	70 to 74 years	0.5	0.9
21 years	574	418	75 to 79 years	0.4	0.6
22 to 24 years	695	380	80 to 84 years	0.4	0.7
25 to 29 years	331	223	85 years and over	0.2	1.1
30 to 34 years	187	198	Total	52.1	47.6
35 to 39 years	189	230			
40 to 44 years	203	218			
45 to 49 years	158	157			
50 to 54 years	135	167			
55 to 59 years	142	141			
60 and 61 years	34	39			
62 and 64 years	67	87			
65 to 69 years	105	123			
70 to 74 years	61	99			
75 to 79 years	39	74			
80 to 84 years	43	79			
85 years and over	26	119			

Source: U.S. Bureau of the Census, 1990, Summary Tape File 3A.

Because these rates are, as their name suggests, mere crude measures, they are usually calculated at more precise levels such as age specific rates or, in the case of death, as cause specific rates. Fertility is the event which requires the most specific calculations. These rates are not described in this chapter but can be found in most good introductory population texts such as that of the Population Reference Bureau (1991). A word of caution about these crude rates–because they are affected by an area's population characteristics, particularly its age structure, the researcher should allow for these differences before reaching conclusions about that area's health, economic or environmental conditions.

Most researchers in public administration will be concerned with obtaining a measure of population change and then deriving estimates of projections of the population. The first step is to compute the population equation which measures change over time. It is expressed as:

$$P2 = P1 + (B - D) + (I - O)$$

TABLE 3 Number of Dependents per 100 Persons Age 18 to 64 Years: 1900 to 2500[a]

Year	Total dependents	Under age 18	Age 65 and over
		Estimates	
1900	79.9	72.6	7.3
1910	73.2	65.7	7.5
1920	72.0	64.0	8.0
1930	67.7	58.6	9.1
1940	59.7	48.8	10.9
1950	64.5	51.1	13.4
1960	82.2	65.3	16.9
1970	78.7	61.1	17.6
1980	64.9	46.2	18.7
1985	61.9	42.6	19.3
1990	62.0	41.7	20.3
		Projections	
1995	63.7	42.8	20.9
2000	62.4	41.8	20.5
2010	60.2	39.0	21.2
2020	68.2	40.4	27.7
2030	78.7	43.0	35.7
2040	79.7	43.1	36.5
2050	79.9	43.9	36.0

[a]Middle series as of July 1, resident population.
Source: Current Population Reports, Series P-25, Nos. 311, 519, 917, 1095, 1127, and Table 2.

where P2, the population at time 2, is equal to P1, the population at time 1, plus births, minus deaths, plus in-(im)migrants less out-(e)migrants.

Thus, using data for the U.S:

$$250{,}878{,}000 = 248{,}168{,}000 + (4{,}179{,}000 - 2{,}162{,}000) + (853{,}000 - 160{,}000)$$

U.S. pop. '91 U.S. pop. '90 Births '90 Deaths '90 immigration '90 emigration '90

If the value of one of the terms is unknown when the others are known, then that value can easily be computed. If only net migration is known, and not the actual numbers of in and out migrants then NM can be substituted for (I − O), adding or subtracting depending on whether it represents a net gain or loss.

Natural increase is the difference (usually a surplus) between births and deaths. NI = B − D. The difference between the birth rate and the death rate is then known as the rate of natural increase (RNI). Thus,

$$\text{RNI} = \frac{\text{B} - \text{D}}{\text{P}} \times \text{k}.$$

In the United States in 1990 the rate of natural increase was equal to

$$\frac{4{,}179{,}000 - 2{,}162{,}000}{250{,}885{,}000} \times 1000 =$$

$$\frac{2{,}017{,}000}{250{,}885{,}000} \times 1000 =$$

$$.008039 \times 1000 = 8.04$$

This is the difference between the birth rate (16.7) and the death rate (8.6).

V. DATA ESTIMATION AND PROJECTION

A. Population Estimation

As the importance of demographic data in decision making in the public and private sector has been increasingly recognized there has been a rapid rise in the demand for such data particularly in the estimation and projection of population. Estimates of population are made by the Census Bureau and by states to meet the need for information used in planning and allocation of grants, funds, and resources. A second area of increased demand has come from small geographic units. Here the need is for estimates to help in such decisions as the location of fire stations, the expansion of schools and the construction of homes and shops. Short term estimates are essential because they fill in the gaps between the decennial counts. They are likely to become even more important if the coverage and content of Census 2000 is reduced. There are now available a range of tools for providing estimates which have been refined over the years. Most draw on data collected for administrative purposes which have proven useful in estimating current population size and in providing an indication of changes in the size of the population residing in an area.

The range of techniques available for estimation include the Component Method, the Census Bureau's Administrative Records Method, the Censal Ratio Method, the Housing Unit Method and the Ratio-Correlation Method. Each of these methods has its advantages and disadvantage (Rives, 1995). Choosing between them is sometimes a matter of data availability. In this section we cannot review all methods. Instead we illustrate the procedure by an examination of the component method, commonly used by the Census and states.

The procedure commonly used to estimate the population of counties and subcounty areas is called the component method because it uses the three basic demographic components of population change—births, deaths, and migration. The relationship between these components was outlined in the population equation above and, as noted there, if we know the number of births and deaths and the volume and direction of migration that has occurred between a base date (usually the last census) and the estimate date (the date for which we wish to obtain a population estimate) then we can very simply determine the population for a given date. This approach can be used for a range of subpopulations or cohorts. When applied to cohorts it is known as the Cohort-Component approach.

The most reliable data in the computation are the births and deaths as these are generally available from state departments and maintained on an annual basis. One important note here is that the birth and death data should be recorded for place of residence not place of occurrence because population estimates are made on a residence base. If births and deaths were recorded in a town in which a major hospital was located, and these vital events were recorded there, then the data would be distorted. The more problematic data are those of migration. Changes of address are not maintained on a continuous basis in this country. As a result migration data

usually has to be estimated using some variable such as school enrollment or tax records which is symptomatic of change in a community. The nature of these symptomatic data may vary according to the unit for which estimates are required.

The procedure used by the Census Bureau for estimating the population of each subcounty area is known as the Administrative Record method (U.S. Bureau of the Census, 1980). In order to estimate the population of these sub-county areas for 1988 the component procedure used each of the components of change for the time period July 1, 1986, to 1988 to add to a July 1, 1986 population base in order to estimate July 1st, 1988. Births and deaths were estimated using birth and death rates adjusted to reported resident births and deaths for counties. Net internal migration was developed using a number of methods. First, Federal income tax return data were used to measure the net number of exemptions moving in and out of each jurisdiction. Secondly, these were then converted into a rate based on the total number of exemptions on income tax returns, and, this rate was applied to the actual population in the jurisdiction exposed to moving in order to estimate the number of migrants. Third, data from the Immigration and Naturalization Service were used to make adjustments for net immigration from abroad. Finally, allowance was made for changes in the number of persons in large group quarters (college dormitory populations, inmates of institutions, military barracks, and populations aboard ships).

Once each of the components of population has been estimated separately, they are combined into an estimate of the total resident population for each area. The formula used to compute the total resident population is:

$$RESP = HHPOP + HHMIG + B - D + IMMIG + SPECPOP$$

where

RESPOP = resident population on the estimate date,
HHPOP = household population on the base date,
HHMIG = net migrants in households for the period,
B = births for the period
D = deaths for the period
IMMIG = immigrants from abroad for the period, and
SPECPOP = special populations on the estimate date.

Once the total resident population estimates for all areas within a county are computed they are controlled to independent estimates of the population at the county level. This has been proven to increase the accuracy of the estimates. Tests of methods have also shown that averaging these estimates using independent methodology tends to increase the accuracy of the estimates. The census estimates are averaged with the estimates prepared by state agencies that participate in the Federal State Cooperative Program for Population estimates. Some of these states prepare estimates based on a housing unit method. In this method, estimates of the number of occupied housing units are developed first for the estimate date. The number of housing units is then multiplied by the average number of persons per household to provide an estimate of the population in households. An estimate of special populations not in housing units is then added in order to obtain an estimate of the total resident population in the area. Other states use the Component Method 11, the Regression Method, the Driver License Address Change Composite Migration Estimating Method, the Administrative Records Method or some average of the above.

The accuracy of the population estimates for sub county areas is assessed by the Census Bureau by comparing the estimates of a decennial year, e.g 1980 or 1990 to the decennial census

TABLE 4 Selected Measures of the Accuracy of Subcounty Population Estimates: 1980

Size of area (population)	Number of areas	Average absolute percent error	Percent positive errors	Less than 10%	10.0 to 19.9%	20% or more
Total	35,644	15.2	48.5	51.9	24.5	23.6
Less than 100	2,425	35.1	55.1	21.4	20.0	58.6
100–499	11,085	19.8	52.8	37.5	26.9	35.6
500–999	6,613	13.2	46.5	52.2	27.7	20.2
1000–2499	7,141	11.6	43.9	58.6	26.3	15.1
2500–4999	3,348	9.6	43.0	66.7	22.6	10.6
5000–9999	2,212	8.3	45.8	72.3	20.6	7.1
10,000–24,999	1,740	6.5	51.7	80.6	14.4	4.9
25,000–49,999	636	5.5	52.7	84.9	11.9	3.1
50,000–99,999	284	4.5	46.5	93.3	6.0	0.7
100,000 and over	160	3.9	36.9	95.6	4.4	—

Source: *1988 Population and 1987 Per Capita Income Estimates for Counties and Incorporated Places: Northeast Current Population Reports*, Service P-26, No. 88-NE-SC.

figures for that year. Table 4 presents the results of the comparison done between a set of April 1, 1980 estimates and the 1980 decennial census counts. As can be seen from the table as the size of the population of subcounty areas decreased, generally the spread in the errors increased.

The estimates of the population of counties are made independently of the estimates developed for sub-county areas. For the majority of counties the estimates are based on an average of estimates developed from the Component Method 11, the Regression (ratio-correlation) Method, and the Administrative Records Method.

Those using population estimates, especially novice users, should always remember that estimates (and projections) are educated guesses, and therefore contain error. Researchers should proceed with caution and remember that the data are approximations, not facts. This is especially important when they are used for crucial decision-making purposes, such as the location of hospitals. The consequences of error can be serious and disputants often find themselves in court!

B. Population projections

A second tool used in the decision-making process and long range planning is population projections. An obvious distinction between estimates and projections is that while estimates are for a point in the recent past for which population census or register data are not available, projections are for some point of time in the future. However, as Long points out, a more fundamental difference between estimates and procedures is that while the estimate, like the projection, is based on a previous census the estimate contains a ''reality check''—actual information related to changes in the population (Long, 1993).

Conventional population projections are an estimate of the future size of the population subdivided by age and sex. As with population estimates there are a variety of methods which can be employed (Smith, 1994). The most elementary population projections begin with population estimates at two or more time points, or with the population size and either birth and death rates. The net reproduction rate can be substituted for the latter. More precise or sophisticated models require an initial population and a series of fertility and mortality rates that can be used

to project survivors by age and births in future periods. The most widely used method is the Cohort Component. This method, like the component estimation method, makes use of the population equation where population at time 2 (P2) is now the population projected at some future date and P1 is the population at the base year from which the projection starts. Usually these projections are computed on an age-sex specific basis, and in some instances (such as the census projections) computed for race and ethnic group. In the method employed by the Census Bureau for national level projections, 1995 to 2050 six sets of data were required to generate the projection figures (U.S. Bureau of the Census, 1996). These are:

1. Base-year population
2. Projected fertility rates
3. Projected survival rates
4. Future net immigration statistics
5. 1990 inflation/deflation rates
6. Armed Forces overseas population

The numbers in this projection are based on an estimated July 1, 1994, resident population consistent with the population enumerated in the 1990 census, and are projected forward using with alternative assumptions for future fertility, life expectancy, and net immigration levels. The components of change are projected separately for each birth cohort. The base population is then advanced each year by using projected survival rates and net immigration by single year of age, sex, race, and Hispanic origin. A new birth cohort is added each year to the population by applying the projected fertility rates by race and Hispanic origin to the January 1st female population.

Because of the breakdown by race and Hispanic origin the computation is more complex than the usual component projection method. Nevertheless, it is quite straight-forward. Each data set is organized into 16 different race/ethnic/sex matrices with a cell for each year of age from 0 to 100 and over. The sum of all the cells in all 16 matrices equals the total population. The method proceeds as follows:

> Starting with a July 1, 1994 modified population estimate based on the 1990 census, each cell is inflated by Demographic Analysis to correct for persons not included in the population count in 1990. Then, each age/race/ethnic/sex cell is survived forward to July 1, 1995, by applying the appropriate survival rate. The population under 1 is created by first calculating the population of women exposed to the risk of childbearing. Generally, this involves averaging the July 1, 1994, and July 1, 1995, inflated female population of each race/ethnic group by single years of age between the years of 14 through 49. Then, the corresponding age/race/ethnic specific fertility rate is applied to this averaged population to produce, after aggregation, the total number of births by race/ethnicity for that 12 month interval. The assumed sex ratio for each group is then used to divide the births into males and females. Then factors from a 1990 census file showing the race and/or origin reported for children in families with parents of differing race and/or origin were applied to the births. Finally, the number of births by sex and race are survived forward to July 1, 1995.
>
> After the births are calculated, net immigration by age/sex/race/ethnicity is added. Then the movement of the population of Armed Forces overseas is applied to the population by detailed group. Next, the population is deflated to be consistent with the 1990 census count. Finally, the 16 groups are summed and subsequently displayed. The same set of procedures, when applied to the July 1, 1995, population would generate the July 1, 1996, population. This process is continued through the year 2050 (U.S. Bureau of the Census, 1996, p. 26).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

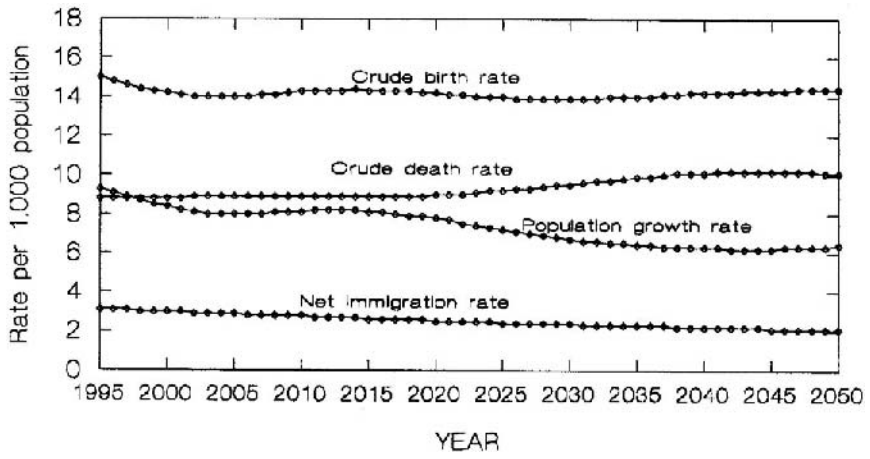


FIGURE 3 Components of Population Change: 1995 to 2050 showing the estimated annual levels of Net Growth, Births, Deaths, and Net Immigration: 1995 to 2050 (Middle Series). From Table 1. Annual Projections and Components of Change for the United States: 1995 to 2050 (Middle Series). Current Population Reports, *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050*. Series P25–1130, 1996.

The validity and usefulness of such a method of projection depends on the accuracy of the base population and of the accuracy of the predicted trends in fertility, mortality, and migration. The cohort-component method described above also depends in large measure upon the assumptions about the demographic processes (fertility, life expectancy and migration) which constitute the population dynamics. Figure 3 displays the projected levels of growth in births, deaths and migration for the period 1995 to 2050 for what is termed a ''middle series'' projection. In a model which predicts national level trends these assumptions have to allow for differences between different ethnic groups.

Finally, because of the tentative nature of the assumptions about these components it is usual to present projections at three levels; ''high'', ''medium'' and ''low''. The ''medium'' projection is what is generally presented as most likely to occur while the ''high'' and ''low'' projections represent plausible upper and lower bounds to future population change. Table 5 displays the data for projections for all three levels for the United States, 1995 to 2050 while Figure 4 graphically displays the differences between the series of projections. The principal fertility, mortality and migration assumptions for each of the three levels of projections are shown in Table 6.

The population projections produced by the Census Bureau produce useful information at the national and state levels and provide a scenario which can aid decision makers. Generally these decision makers are aware of the need for such projections in long range planning in such areas as planning for elementary and secondary school facilities. Frequently, however, they may have difficulty in applying these general projections to the specific populations in which they are interested. In such instances it is necessary for the user to adapt the generic projections. This can most easily be done by applying some procedure which will forecast the future demand for, say, school places, taking into account the projected size, distribution, and composition of the population to be served. The generic age-sex projection can be transformed by applying rates, ratios and proportions that make them more relevant to the decision-maker's needs. Thus

TABLE 5 Total Resident Population: 1900 to 2050[a]

Year	Lowest series	Middle series	Highest series
		Estimates	
1900	(X)	76,094	(X)
1910	(X)	92,407	(X)
1920	(X)	106,461	(X)
1930	(X)	123,077	(X)
1940	(X)	131,954	(X)
1950	(X)	151,868	(X)
1960	(X)	179,979	(X)
1970	(X)	203,810	(X)
1980	(X)	227,225	(X)
1985	(X)	237,924	(X)
1990	(X)	249,402	(X)
		Projections	
1995	262,798	262,820	262,846
2000	271,237	274,634	278,129
2005	276,990	285,981	295,318
2010	281,468	297,716	314,571
2020	288,807	322,742	357,702
2030	291,070	346,899	405,089
2040	287,685	369,980	458,444
2050	282,524	393,931	518,903

[a] In thousands as of July 1, resident population.
X: Not applicable.
Sources: Tables 1 and 3 in Current Population Reports, Series P-25, No. 311, 519, 917, 1095, and 1127.

the generic age-sex projection is converted into a projection for a specific ''function'' such as a service or a facility. A functional population projection is an age-sex projection that has been transformed or otherwise incorporated into formulas that forecast the future supply or demand for some particular purpose (Kono, 1993). Examples of such functional projections are, the future size of the labor force, the future size of high school enrollments, the future number of households, the future needs of community services and facilities and future requirements for food, energy and other resources.

Educational planning projections are an example of functional projections. For instance, the number of students who will be attending elementary and secondary schools (k-12) is a function of two factors, the number of children of school age and the proportion of those children who will actually enroll in public schools rather than in private schools or in home schooling. Projections for the former are available from the Census Bureau, so the problem is to estimate the enrollment proportions. The U.S. Department of Education's National Center for Education Statistics (NCES) addresses this task for national and state level enrollments through the combination of grade retention and enrollment rate methods (U.S. Department of Education, 1996).

The grade retention method starts with 6-year-olds entering first grade and follows their progress as a cohort moving through subsequent public schooling. Transition from grade to grade is confronted as a ''survival'' or retention rate, the fraction of the earlier year who enter the following grade. The second approach uses the enrollment rate for the ages involved, expressed as the proportion (or percentage) of the population of each age group that actually enrolls

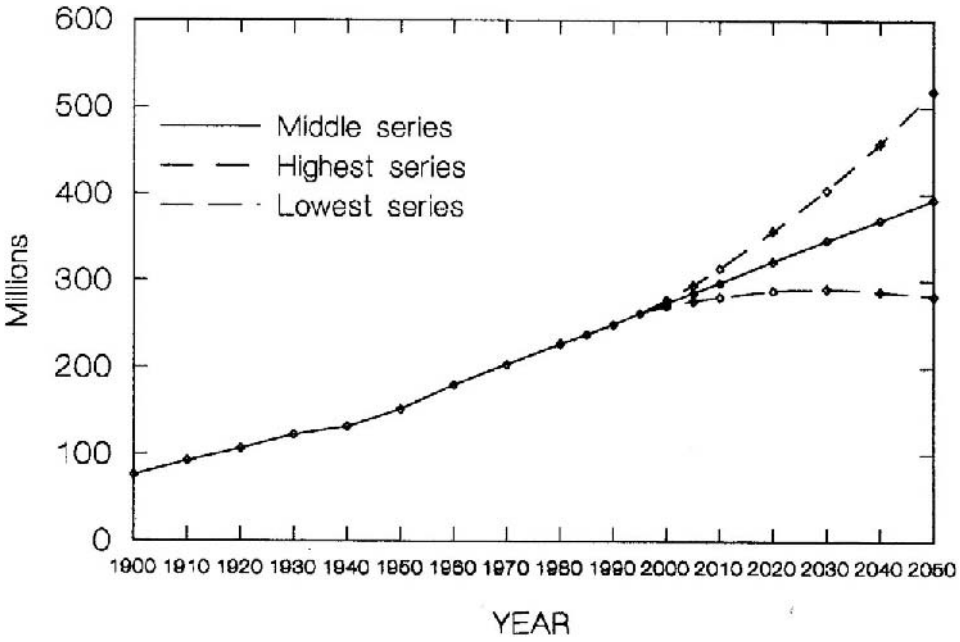


FIGURE 4 Total Resident Population of the U.S.: 1900 to 2050 comparing three series of projections (High, Middle and Low). From Table C, Current Population Reports, *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050*, Series P25–1130, 1996.

in public schooling. Note that these are alternative routes to the same end point. Suppose one starts (as does NCES) with projected enrollments in kindergarten and first grades: one applies the projected enrollment *rates* for these grades to the population projections of 5- and 6-year olds produced by the Bureau of the Census. From this point one can either a) apply the expected retention rates from first grade to second grade, from second grade to third grade, and so on, to estimate numbers in future years or b) one can continue to apply projected (grade-specific) enrollment rates to population projections for this cohort in subsequent years, to arrive at estimates of each grade's enrollments. Combining the two methods to obtain a composite estimate of future enrollments is superior to either alone: each method has its particular biases and a composite tends to limit their respective effects. Both methods assume that past trends in factors affecting enrollment will continue over the period of forecasting and so is vulnerable to changes in migration, mortality, drop-out rates, and use of private schools and home schooling. State-level projections are particularly vulnerable to changes in migration patterns which may change dramatically over relatively short periods. In practice long-term projections appear to be more successful when based on the enrollment rate method, probably because migration rate shifts between states are better captured by Bureau of Census population projections. The grade retention method yields somewhat better estimates over shorter projection periods by virtue of its sensitivity to short-term population movements. The composite estimate of projected enrollment is then made as a weighted function of the two methods, using the formula

$$E_t = B_t X_R(t) + (I - b_t) X_E(t)$$

where E_t is the composite projection for year t (from now)

X_R (t) is the projection for year t based on the grade retention method

TABLE 6 Principal Fertility, Mortality, and Net Immigration Assumptions in Each Projection Series

Principal series name	Fertility	Life expectancy	Net immigration
Middle	Middle	Middle	Middle
Lowest	Low	Low	Low
Highest	High	High	High
Low fertility	Low	Middle	Middle
High fertility	High	Middle	Middle
Low life expectancy	Middle	Low	Middle
High life expectancy	Middle	High	Middle
Low net immigration	Middle	Middle	Low
High net immigration	Middle	Middle	High
Zero net immigration	Middle	Middle	Zero

Source: Population Projections of the United States by Age, Sex, Race and Hispanic Origin: 1995 to 2050. Current Population Report, Series P-25-30.

X_E (t) is the projection for year t based on the enrollment rate method

and b_t is the time specific weight to be applied to the grade retention method, varying linearly from 1.0 for a one year lead time ($t = 1$) to 0.0 for a ten year lead time ($t = 10$), and is zero for all greater lead times. Thus a one-year projection is based exclusively on the retention method and ten-year and longer-term projections are based exclusively on the enrollment method estimates, with both methods contributing to intermediate-term estimates.

The estimated total enrollment in elementary schools EG at time t can be written as

$$EG_t = K_t + E_t + \sum G_{jt}$$

where t denotes the time of interest

K_t is the enrollment at nursery and kindergarten level
E_t is the enrollment in elementary special and ungraded programs
G_{jt} is the enrollment in grade j at time t

and the summation operator $\sum$ is taken over grades 1–8 ($j = 1, 2, \ldots, 8$).

In this formula K_t is estimated from

$$K_t = R_t\,(K) \times P_{5t}$$

where $R_t(K)$ is the enrollment rate for nursery and kindergarten and P_{5t} is the population of five-year-olds at time t;

$$E_t = R_t(E) \times \sum P_{it}$$

where $R_t(E)$ is the enrollment rate for elementary and ungraded programs

and P_{it} is again the population of age i at time t
and $\sum$ is taken over grades 1 to 8;

G is estimated as

$$G_{jt} = R_{jt} \times G_{j-1,t-1}$$

where R_{jt} is the retention rate for grade j (i.e. the proportion of grade j in year t that were in grade $j - 1$ in year $t - 1$).

Thus the enrollment in each higher grade is estimated recursively, starting with a direct estimate for the first grade value G_{it} at any time t.

Similarly the projected total enrollment in secondary grades (9–12) at time t, SG_t, is

$$SG_t = S_t + PG_t + \sum G_{jt}$$

where S_t is enrollment in secondary special and ungraded programs

PG_t is enrollment in postgraduate programs in secondary schools and
G_{jt} is as above, now summed over grades 9–12 ($j = 9, 10, 11, 12$).

In this formula

$$S_t = R_t\,(S) \times \sum P_{it}$$

where R_t (S) is the enrollment rate for secondary special and ungraded programs and P_{it} is again the population of age i at time t;

$$PG_t = R_t\,(P) \times P_{18t}$$

where $R_t(P)$ is the enrollment rate for postgraduate programs and P_{18t} is the population aged 18 and older at time t.

In detail the NCES estimates develop a smoothed version of the projection estimates before combining with the retention estimates described above.

VI. COHORT ANALYSIS

In the sections above we have referred to groups of individuals who experience the same event within the same time interval, such as baby boomers, the elderly or elementary school children. These are all examples of cohorts. A cohort is defined as ''those people within a geographically or otherwise delineated population who experienced the same significant life event within a given period of time'' (Glenn, 1977: 9). In general experience its commonest manifestation is probably ''The Class of 19—'' whilst in the social sciences cohorts are most frequently defined by year of birth, yielding a birth cohort. Demographers use birth, marriage and education cohorts for analytic purposes. Cohort analysis is a method of research developed by demographers to analyze properties of such groups over time and is a tool in strategic planning. Two types of question are often at the focus of a cohort analysis. First, what changes in characteristics or attitudes of people develop with age? This has been the commonest application of the technique, with studies *interalia* of voting affiliation and pattern, political conformity, and mental ability. Second, does the behavior or attributes of people in a specific cohort differ from those of people in other cohorts? (Discussion of the behavior of ''Baby Boomers'' is implicitly based on such a focus, with the idea that people in this cohort (born from 1946 through 1964) behave differently from their ancestors as a result of being born in those specific years.)

Cohort analysis is most powerful when based on a standard cohort table (Table 7). In such a table each cohort is defined by its starting period, and on each subsequent occasion on which the cohort attributes are re-measured, tracking of a further cohort is initiated at the original age of the first. Thus in Table 7 the initial cohort is the 1952–1956 age group of 15–19 year-olds who experienced a homicide rate of 6.2 (per 100,000); a second cohort of 15–19 year-olds commenced in 1957–1961, when the first cohort, now five years older, experienced a higher rate of 13.6 homicides per 100,000 at age 20–24 years-old; the third is started at 15–19 years

TABLE 7 A Cohort Table for Homicide Frequency (deaths per 100,000)

	Period				
Age group	1952–1956	1957–1961	1962–1967	1968–1971	1971–1976
15–19	6.2	7.5	8.6	15.1	17.1
20–24	11.8	13.6	14.2	22.9	25.5
26–29	12.4	11.9	13.6	19.3	22.2
30–34	10.8	10.6	10.9	15.5	16.9
35–39	9.4	8.8	9.1	12.5	13.4
40–44	7.7	6.8	7.1	9.6	10.2
45–49	6.1	5.7	5.5	7.3	7.4

Source: Smith, M.D. Increases in youth violence: age, period, or cohort effect. Presented at the American Sociological Meeting. Boston, MA. 1979. Table reproduced in D. Knoke and P.J. Burke. *Log Linear Models*, Sage University Paper, 20. Newbury Park, CA. Sage Publications, 1983.

a further five years later in 1962–1967, when the second cohort is of age 20–24 and the first is of age 25–29, and so on. Such a table is demanding of data, in that data must be available for exactly the right years, but has many advantages. In particular, movement along the diagonals tracks each cohort in time as it ages, movement down each column corresponds to effects between age in a given period, and movement across each row reveals effects at constant age as each cohort of that age is replaced by a successor. Thus there are three distinct classes of effects on the response variable(s) of interest: there are age effects associated with biological age; there are cohort effects due to factors associated with birth cohort; and there are period effects due to factors influential at any given period (Glenn, 1977). In Table 7 the data along most rows show a distinct increase over time in homicide rates experienced at any fixed age, and the data within columns show markedly higher homicide rates at all ages in 1968–1971 and 1971–1976. Proceeding from top left to lower right along any diagonal corresponds to tracking the fate of any particular cohort.

In principle the idea of cohort analysis is to determine the extent of influence of each of these three classes of factors. In practice this is made difficult by three statistical issues. First, each entry in a cohort table is a sample statistic rather than a parameter of the underlying population. It is straightforward to demonstrate the occurrence of departures of the data from an underlying null model (or null hypothesis) but it is very difficult to prove that a given pattern did not result from sampling variation. Second, as a cohort ages it loses some members to mortality and to migration. This means that even if cohorts are sampled on any given sampling occasion in a perfectly random manner, there exists the possibility that population changes have been nonrandom; then the members of the cohort on later dates are a systematically biased subset of the cohort composition on the earlier measurement dates. This implies that cohort analysis will be most effective for populations that are relatively closed, with little in-migration or out-migration.

The third issue is more complex to explain. Since the cohort attributes are examined by sampling each cohort at different times, each set of measurements reflects the influence of all three classes of effect (age, period, and cohort). Within each diagonal each set of measurements is the outcome of the change in age of the members of that cohort since previous measurement and is also the outcome of those measurements being made in the current period. Similarly each column is the joint outcome of factors associated with age and cohort identity: one has a cross-sectional analysis in which one wishes to attribute the differences between rows to the age effect

but where in reality the different age groups are from different cohorts. Within each row age is fixed but the measured effects may originate in the cohort or in the period attribute of the sample. Such cross-correlation means that the effects of any factor of interest are inevitably confounded with the effects of one other factor.

As noted above, studies of aging have provided one of the most frequent applications of cohort analysis. Age is typically characterized by chronological age but in reality reflects a suite of correlated changes with maturation and experience. Maturation (or biological aging) refers to the sequence of physiological changes which occur with increase in years. Associated with these physiological changes is psychological aging in which personality changes involving attitudes, values and behavior develop with age. These changes are further paralleled to a greater or lesser extent by what has been termed ''social aging,'' the sequences of changes in status and relationships to others that occur with increase in age. In general studies of these factors have been indexed almost exclusively to chronological age, making it difficult to separate out the influences of these several dimensions of aging. Thus the effects of being ''young at heart'' are hardly accessible to investigation through cohort analysis.

Where effects present are largely due to aging, the cross-sectional pattern with age should be approximately parallel in different periods. Glenn (1977) describes such patterns as ''surprisingly frequent,'' though most tables show evidence of at least two effects. Pure cohort effects are, not surprisingly, extremely rare.

Table 7 provided an example of a standard cohort table in which age and period increments are matched. In practice many cohort datasets involve multiple age classes (the cross-sectional component) measured at multiple dates (the cross-sectional component) measured at multiple dates (the longitudinal component) that differ in spacing from that between age groups. With such data Glenn (1977) recommends construction of two tables. One displays the values of the dependent variable for a given cohort at different times. The other displays trend data for each age level. Each table emphasizes particular patterns within the dataset, though the problems of cross-correlation already mentioned persist to confound interpretation.

A. Data Sources

Cohort analysis is typically a secondary analysis of extant social survey data gathered for other purposes, it being essentially impossible to contemplate a designed survey with the necessary extent and temporal span otherwise needed. Thus analysis usually focuses on data in national sample surveys repeated at intervals. At least four sources of American data meet this repetition criterion.

The first is the periodic data of the Bureau of the Census decennial censuses and the Current Population Surveys.

The second source of survey data particularly amenable to cohort analysis are the General Social Surveys of the National Opinion Research Center in Chicago. Designed to allow trend determination over extended time spans, these surveys have expressly addressed the availability of repeat surveys, and the codebooks for the Surveys contain appendices detailing the previous usage in earlier surveys of questions re-visited in later ones. The data are thus keyed to ready use in cohort analysis.

The other two major sources of repeated survey data are the Roper Public Opinion Research Center (Williamstown, MA) and the Institute for Social Research at the University of Michigan (Ann Arbor, MI). Glenn (1977) discusses several of the advantages of these surveys, particularly as to academic access, and describes several of the indexing tools available to identify questions recurrent in the surveys conducted by these two institutions and therefore amenable to cohort analysis.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

B. Data Limitation

Several limitations critical to cohort analysis emerge under detailed review of potential data sources, such that questions repeated in multiple instances of particular surveys turn out to be essentially incomparable across periods. One such problem lies in the changing demographic representation of survey samples. Some of the earliest Gallup polls, for instance, were intended to predict the outcome of elections and the samples were deliberately designed to reflect the composition of the voting populace. As a result, segments of the population less likely to vote—women, blacks, southerners, and people with low education—were largely omitted from the samples. In later years most survey organizations shifted to sampling in line with the composition of the population as a whole (Glenn, 1977).

Another common problem in cohort analysis is where particular questions are preceded by a ''filter.'' Glenn (1977) notes that surveys of views on school desegregation were frequently—but not invariably—preceded by a question as to whether the respondent had children of school age. Since responses by a filtered and by an unfiltered sample of respondents are likely to differ, comparability is violated and needs to be restored e.g. by appropriately filtering the response database to correct for that bias.

A third source of incomparability arises where the range of allowable answers in response to a common question changes with time. Respondents are, for example, likely to respond differently when the extreme option for an answer is ''not at all satisfied'' than when it is ''not very satisfied,'' even if all other options are unchanged. An extensive literature shows that the presence of such ''hedge'' terms is associated with different views as to the truth of the hedged statement (Zadeh, 1966). Where such a change in phrasing occurs between time periods, there exists a risk that any change in response level detected between periods is a consequence of the change in wording rather than of a temporal pattern in the phenomenon of interest.

A fourth source of error arises with data with a significant degree of seasonal variation. Thus voting behavior may vary with the weather, views on air safety may be different in the immediate aftermath of a major air crash, and support for campaign finance reform may alter with the level of television coverage prevailing. A related source of variability may lie with response bias, where respondents supply inaccurate information to the interviewers, sometimes deliberately and sometimes inadvertently. For example, people have been shown to give different responses to interviewers depending on their ethnicity so that a shift in the ethnic proportion of interviewers over time is likely to result in a shift in mean response.

A further source of error in cohort analysis may arise with changes in social attitudes towards particular behaviors. It is well-established that respondents tend to over-report socially acceptable behaviors and to under-report behavior and attitudes that are frowned upon. If approval of certain behaviors or attitudes change over time, cohort analysis may determine an increase in the incidence of those behaviors or attitudes as the cohort ages, where the true pattern is one of stasis.

The results of cohort analysis can in some cases be adjusted for these biases. Thus if women are under-represented in a regional sample, and their true representation in the region is known e.g. from Census data, then the observed responses by women can be weighted in a formula such as

$$R = \sum w_i \times f_i / \sum w_i$$

where f_i is the observed frequency of response by the ith sex, w_i is the true proportion of the ith sex in the region, and R is the true response rate for the combined population of men and women. Similar numerical adjustment is possible wherever the true representation is known and

the response rate from the group of interest within the survey is known. Similar logic applies to the measurement of attributes in a regional population.

As noted above, changes in the composition of a cohort may arise as a result of mortality, such that the attributes of a cohort measured at a later date are the attributes of the survivors, with this subset of the original cohort membership perhaps differing in the attribute of interest from those who died. Dependent variables arguably linked to mortality require special attention in this respect: Appleton et al. (1996) provide a clear example of how ignoring age-related mortality can lead to nonsensical conclusions.

C. Significance Testing and Sampling Variation

As noted above, cohort analysis is typically conducted using data gleaned from extant surveys designed for other purposes. This can create problems when testing the statistical significance of the results of cohort analysis. National samples-and particularly the earlier ones-were often derived by ''quota control'' methods, in which a sample is expressly enhanced in respect of under-represented components until the desired level of representation has been reached. More recent surveys have followed a stratified sampling design that yields a full probability sample. This is achieved by drawing an initial sample from a list of geographic units e.g. in the United States typically of Metropolitan Statistical Areas (MSAs) and non-urban areas of similar size. The elements of this initial sample are termed Primary Sampling Units and are typically used in a number of sequential surveys before being replaced. Each PSU is in turn sub-divided into a set of smaller units about the size of a city block in an urban area; in less populated areas these units are spatially larger but about the same in terms of population. Finally, the population of interest within these blocks are sampled, either following random sampling or through a quota control method. The ''population of interest'' may be either households or people, and the sample cases are correspondingly different.

Using such a multi-stage process has implications for the sampling uncertainty. In particular, if a dependent variable is well-correlated with the sample clusters, variance inflation may occur. For example, if there exists substantial residential segregation on the basis of wealth or education or race, the variance estimates for variables correlated with these aspects will be inflated. This inflation can be as high as four-fold, though in practice standard errors of many variables are perhaps only 25% larger than would obtain from truly random sampling (Glenn, 1977). Moreover, if bias is not an issue, sampling variance for the mean is greatly reduced the larger the available sample, so sampling uncertainty is an issue only for peripheral analyses focussed on smaller subsets.

One should also recall the issue of cross-correlation of variables described above. There is perhaps less point in assessing the statistical significance of a correlation if the inevitable presence of a confounding variable precludes realistic interpretation of the results. Indeed Glenn (1977) argues for what is really a weight-of-evidence approach to cohort analysis, suggesting that self-consistency of multiple patterns, agreement of results with prior theory, and the presence of systematic patterns across multiple time periods all make it unlikely that an analysis lacks significance. Although these points are true, reliance on them implies that the role of cohort analysis is largely confirmatory, except when matching a theoretical prediction for the first time. A different view is that cohort analysis allows study of broad-scale pattern and trend inaccessible to the type of local detailed studies amenable to rigorous statistical analysis. This view sees an important role for cohort analysis in hypothesis generation. In most analyses elements of both lines of thinking are likely to be present.

Several statistical approaches are available for use in cohort analysis, including analysis

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

of variance and covariance, use of dummy variable regression, and log-linear analysis. In general these methods explicitly represent potential interaction between variables and attempt to partition variance across factors and interactions. At least three difficulties need to be faced when adapting these approaches to cohort analysis. First, most of these methods model linear effects (sometimes in transformation space) and may misrepresent as interaction between effects what are really the results of non-linearities. One might ask, for example, whether it is realistic to assume in the interests of applying a linear model that people's attitudes change at a constant rate with age? Second, many regression models assume that effects are additive e.g that period effects are the same for all cohorts and ages (and similarly for age and cohort as dependent). However, if older people are more rigid in their attitudes, it follows immediately that period effects are not independent of age and statistical models requiring such an assumption are flawed from the outset. It is possible to model nonlinear effects and interactions, but most such approaches require one to specify the *form* of the relevant function. Subsequent results are then as likely to reflect error in the choice of function as an effect of interest.

These limitations are not insurmountable but they do raise the standards of statistical knowledge needed for robust analysis of cohort data. Since to a large extent cohort analysis is being conducted on secondary data and because suitable panel data cannot be obtained, the question of balance between statistical rigor and the data rigor always bears consideration.

Simple mathematical modeling of effects may provide greater insight than does statistical modeling. Any change of interest suggested in cohort analysis derives from intra-cohort change and from the succession of cohorts, with this latter the combined outcome of addition and subtraction of individuals. These different sources of change cannot be unequivocally segregated in cohort analysis, instead requiring a full panel study (see below) for resolution. However, some approximate estimates of the likely magnitude of the different sources can sometimes be obtained through elementary modeling.

Suppose we have an observed change in a response variable between two time periods (say ten years apart), measured in adults aged 20–70 years. Then the response of, say, 30-years-olds in the second period is the (sampled) response of people who were in the 20-year-olds cohort in the earlier period. If there has been no intra-cohort change between the two periods, we expect the 30-year-olds to retain the response they had as 20-year-olds, the 40-year-olds to retain the response they had as 30-year-olds, and so on. Additionally, we expect a stationary age distribution under these conditions. Hence we can estimate the outcome in the absence of intra-cohort change by assigning to each cohort in the second period the response it had ten years earlier. Only for the 20-year-olds in the second period is there no estimate of their earlier response, since they have only just entered the sample. Since by hypothesis we have age stationarity, we can weight these responses by use of Equation 1 above, with w_i and f_i here respectively the proportion of the population in the *i*th age group on the first date and the corresponding response for that age group. This procedure thus allows only one source of change in the response, namely the change in age distribution as the cohort ages, and the estimate R from equation 1 is then the change assuming no intra-cohort change between periods. This estimate can then be subtracted from the observed change in response across the cohort to obtain an estimate of the contribution of intra-cohort change.

Using an analogous procedure for cohort succession one can estimate the contribution of succession to the observed change. If the sum of the two contributions is close to the value observed, one may conclude it is unlikely that interaction is present and one can use the relative size of the two contributions as a measure of the strengths of their respective effects. If, on the other hand, their sum exceeds the measured value, the two effects are likely to be correlated whilst if their sum is markedly less than the measured effect an interaction is probable. However, in such modeling one needs to consider the likely effect of sampling variability on the results,

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

and conclusions about the relative size of intra-cohort and cohort successional changes, of correlation, and of interaction need to be evaluated against the uncertainty in the individual response values measured.

D. Alternatives to Cohort Analysis

Panel studies resemble but are not identical to a cohort analysis. In a panel study the same individuals are studied on multiple occasions and the total change of interest experienced by the panel is measured. In an intra-cohort study, on the other hand, samples of individuals from the cohort are studied at each time period, without any constraint on constant membership between samples: meeting the defining criterion for cohort membership, and not prior membership in the sample form previous occasions, is the only requirement for sampling in cohort analysis. The key difference is that cohort analyses examine only the aggregate values of the variables of interest where a panel study documents individual responses and therefore the extent to which responses in different segments of the population may offset each other. This last is the major advantage of panel studies over cohort analysis, particularly if relevant attributes of the individuals involved in these offsets are known: the correlates of the individual responses can then be identified and used in assessment of likely causation. Outside this gain, panel studies may be no more powerful than cohort analysis: they share with such analysis the confounding of age and period effects and are also particularly expensive to conduct because of the necessity of tracking individual panel members over extended periods of time.

Cross-sectional studies attempt to infer patterns of temporal change from the simultaneous observation of samples of different-aged individuals at a given moment. If age differences are dominant relative to cohort and period effects, such comparison yields effective insight into the correlates of aging. The problem is, of course, to know whether such dominance is in fact the case. Conducting cross-sectional analysis for two (or more) periods can be informative here, in that conclusions should be largely independent of period if age effects are indeed dominant (see above). Cross-sectional analysis can readily incorporate the influence of other attributes of the sampled populations, such that the putative effects of age can be related to each of such ancillary variables. For example, one can consider the inferred aging effects in highly educated versus poorly educated members of the sample.

Retrospective study is the third alternative channel of approach to the problems open to cohort analysis. For social scientists such studies are typically through interview records of people's recollections of past events. If participant recollections could be standardized to the dates desired of a cohort analysis, the resulting database could be the basis for a cohort or a panel study. In practice, however, faulty recollection of events and, perhaps more importantly, the irregular spacing of significant events within each participant's timeline probably make such a study transformation impossible. In addition, a retrospective study shares with cohort analysis the confounding of age and period influences.

REFERENCES

Appleton, D.R., J.M. French, and M.P.J. Vanderpump (1996). ''Ignoring a Covariate: an Example of Simpson's Paradox,'' *The American Statistician, 50* (4): 340–341.

Glenn, N.D. (1977). *Cohort Analysis*. Beverly Hills, Sage Publications.

Kinter, H.J., T.W. Meriick, P.A., Morrison, and P.R. Voss (1994). *Demographics*: *A Casebook for Business and Government*. Boulder, Westview Press.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Kono, S. (1993). ''Functional Population Projections,'' *Readings in Population Research Methodology, Vol. 5, Population Models, Projections and Estimates*, New York, United Nations Population Fund.
Long, J.F. (1993). ''Population Estimation,'' *Readings in Population Research Methodology, Vol. 5, Population Models, Projections and Estimates*, New York, United Nations Population Fund.
Myers, D. (1992). *Analysis with Local Census Data: Portraits of Change*. Boston, Academic Press.
Haupt, A. and T.T. Kane (1991). *Population Handbook*, 3rd edition, Washington, D.C., Population Reference Bureau.
Rives, Jr., N.W., W.J. Serow, A.S. Lee, H.F. Goldsmith, and P.R. Voss (1995). *Basic Methods for Preparing Small-area Population Estimates*,'' Madison, University of Wisconsin.
Smith, D.P. (1994). *Formal Demography*, New York, Plenum Press.
U.S. Bureau of the Census (1980). *Population and Per Capita Money Income Estimates For Local Areas*: *Detailed Methodology and Evaluation* Current Population Reports, Series P-25, No. 699, Washington D.C., U.S. Government Printing Office.
U.S. Bureau of the Census (1996). *Population Projections of the United States By Age. Sex, Race, and Hispanic Origin*: *1995 to 2050*, Current Population Reports P25–113, Washington, D.C., U.S. Government Printing Office.
U.S. Department of Education (1996). *Projections of Education Statistics to 2006*, National Center for Education Statistics 96–661, Washington, D.C., U.S. Government Printing Office.
Zadeh, L.A. (1996). ''Fuzzy Logic = Computing with Words,'' *IEEE Transactions on Fuzzy Systems*, *4*: 103–111.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

17

Multivariate Regression Analysis in Public Policy and Administration

Elizabeth A. Graddy
University of Southern California, Los Angeles, California

I. INTRODUCTION

Multiple regression offers analysts one of the most powerful and useful tools for quantitative analysis. With the exception of descriptive statistics, it is the most widely used quantitative method. There are three primary reasons for its popularity. First, it is accessible. Regression analysis is relatively easy to use and understand, and estimation software is widely available. Second, the multivariate linear specification is robust. Many relationships have been found empirically to be linear. The linear specification is the simplest, and thus always appropriate as a first approximation of a causal relationship. Moreover, we often do not know enough about the relationship under study to justify an alternative (nonlinear) specification. Third, the results of regression analysis have proven to be very useful, both for predicting or forecasting and for explanation (i.e., determining the causes of a phenomenon). It is the ability of multivariate regression to control for confounding influences on the relationship under study that makes it a particularly powerful tool for explanation.

Multiple regression has been used to explore a wide range of phenomena in public administration and public policy. Examples include work on organizational structure (Graddy and Nichol, 1990) and organizational behavior (Robertson, 1995); local service delivery issues including contracting (Ferris, 1988), volunteering (Sundeen, 1990) and coproduction activities (Sundeen, 1988); intergovernmental questions (May and Burby, 1996); evaluations of programs (Devaney, Bilheimer and Schore, 1992) and laws (Graddy, 1994); and a variety of issues focused on specific policy areas, e.g., health policy (Greenwald et al., 1984; Mann et al., 1995).

Given its power and usefulness as a methodology and the broad range of public sector issues about which it has provided insight, consider what regression analysis is and how one can effectively use it. Assume we want to quantitatively analyze a relationship between two or more variables. We need a set of observations for each variable, and a hypothesis setting forth the explicit form of the relationship. The set of observations, a *sample*, is chosen from the population of interest. The variable we wish to explore is called the *dependent* variable (denoted Y). The variables that are believed to cause or influence Y are called *independent* variables (denoted as Xs).

The model we will explore in this chapter is a multivariate linear relationship between X and Y, or:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i \quad (1)$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where

- Y denotes the dependent variable[1]
- X_j denote the independent variables, j = 1,2, . . . , k
- β denote the coefficients that measure the effect of the independent variables on the dependent variable
- ϵ denotes the error term, which represents stochastic, or random, influences on the dependent variable

Consider a simple example involving the determinants of crime. Assume first that we believe that crime increases in densely populated areas; in other words, the crime rate depends on population density. Specified as a bivariate linear relationship, this becomes:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2}$$

where

- Y is the crime rate, and X is a measure of population density
- β_1 is the average change in the crime rate associated with a unit change in population density
- β_0 is the average crime rate when the independent variable is zero

We need to include an error term (ϵ) in our specification *even* if we believe that density is the sole determinant of the crime rate because there may be minor random influences on Y other than X, and there may be random measurement error. ϵ represents the difference in the *observed* Y (the data) and the *expected* Y, *E*(Y), in the population, which is based on the model.

Estimating Equation 2 using data on 1980 crime rates and population density for 62 New York counties[2] reveals that $\beta_0 = 3859$ and $\beta_1 = .18$. Therefore, the estimated value of Y for a given X is: 3859 + .18X. The interpretation of the estimated equation is:

- the crime rate (the number of reported offenses per 100,000 population) increases an average of .18 (the value of β_1) for each unit increase in population per square mile
- the expected crime rate when population density is approximately zero is 3859 (the value of β_0)

The estimated line represents the average crime rate for a given population density, and is plotted in Figure 1.

Obviously, factors other than population density can affect the crime rate. Consider a slightly more complex model in which crime also depends on opportunity (as measured by the high school dropout rate and per capita income). Our model becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{3}$$

where X_2 = the high school dropout rate, and X_3 = per capita income

The interpretation of the coefficients in this multivariate regression equation changes in a subtle, but important way, from the interpretation in the simple regression model. The coefficients now provide the average change in the crime rate associated with a unit change in their respective variables *holding the other independent variables constant*. For example, β_1 is the average change in the crime rate associated with a unit change in population density (X_1) with both the dropout rate and income held constant. β_0 provides the average crime rate when *all* the independent variables equal zero.

Estimating Equation 3 using the New York data generates the following parameter estimates: $Y = 675 + .12X_1 + 210X_2 + .27X_3$. Thus by including other determinants of the crime

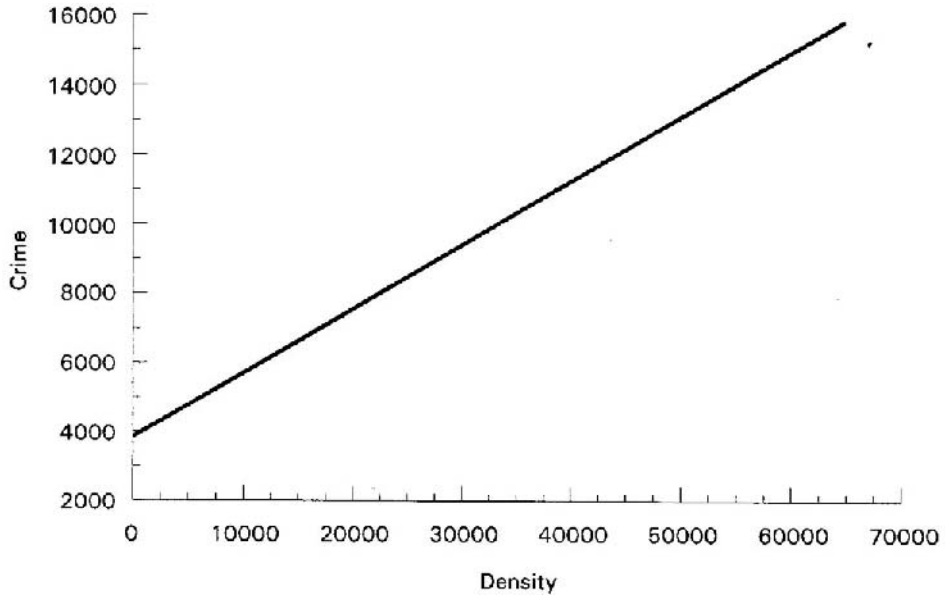


FIGURE 1 Estimated line.

rate, we find that the impact of population density is reduced. With the dropout rate and income held constant, a unit increase in population density raises the crime rate by only .12.

In general, for the multiple regression model described by Equation 1 β_j provides the average change in the dependent variable associated with a unit change in X_j with the other independent variables held constant. This means that we can estimate the effect of X_j *alone* on Y. Multiple regression thus allows for the statistical control of confounding variables, and this is one of its most powerful characteristics. It allows us, for example, to isolate the effects of a teenage pregnancy prevention program on the subsequent rate of teenage pregnancies, controlling for known (or suspected) determinants of teenage pregnancies. The parameter estimate associated with the program isolates its effect.

This ability to control for confounding influences is particularly important when the relationship of interest cannot be easily separated from other effects without using statistical techniques. For example, assume we want to estimate the effects of gender discrimination on salaries in an organization, but we know that many of the higher paid males have been in the organization longer than their female colleagues. We can estimate this relationship using multivariate regression analysis by including seniority (and other determinants of income) as independent variables in the model. The parameter estimate on gender will isolate the effects, if any, of gender discrimination.

The effective use of multivariate regression analysis requires an understanding of how to estimate the regression model and interpret the results, and an understanding of the assumptions that underlie the model and their implications for the credibility of the results. The remainder of the chapter is devoted to these topics. Section II explains the intuition behind estimation. Section III explains how to evaluate estimation results. Section IV presents the assumptions that underlie the use of regression analysis for statistical inference, and Sections V through VIII address how assumption violations are identified and their implications. Section IX concludes the chapter. Table 1 provides a list of the symbols that will be used in this chapter.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 1 List of Symbols

Population parameters (unobserved)		Estimated parameters (from sample)	
Name	Symbol	Name	Symbol
Regression coefficient	β	Estimated coefficient	$\hat{\beta}$
Variance of estimated coefficient	$\sigma^2_{\hat{\beta}}$ or $VAR(\hat{\beta})$	Estimated variance of the estimated coefficient	$s^2_{\hat{\beta}}$
Standard deviation of the estimated coefficient	$\sigma_{\hat{\beta}}$	Standard error of the estimated coefficient	$s_{\hat{\beta}}$
Error term	ϵ	Residual	e
Variance of the error term	σ^2_{ϵ} or $VAR(\epsilon)$	Estimated variance of the error term	s^2
Standard deviation of the error term	σ_{ϵ}	Standard error of the equation	s
Expectation operator	$E(\cdot)$		

II. ESTIMATION

The correct use of regression analysis requires at least an intuitive understanding of how the multivariate regression model is estimated. We will develop this understanding using a simple regression model.

Recall our first model of the crime rate as a function of only population density. Assume we believe this model, $Y = \beta_0 + \beta_1 X + \epsilon$, to be true. Note, that we *cannot* observe the value of the parameters (β_0 and β_1), but we *can* observe Y and X (in this case, the crime rates associated with different population densities). We would like to use the information we have (a set of *n* observations on Y and X)[3] to estimate β_0 and β_1, because these parameter estimates will provide a quantitative description of the relationship between X and Y.

More precisely, the parameter estimates will provide an *estimated* value of Y for a given value of X. In this example, the estimated β_0 is 3859 and the estimated β_1 is .18, therefore the estimated value of Y is 3859 + .18X. Estimated values will be denoted with a hat, so $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is the estimated version of our model.

The *estimated* value of Y associated with a particular value of X will usually not be the same as the *observed* value of Y associated with the same value of X. For example, our estimated equation predicts a crime rate of 3880.6 for areas with a population density of 120. One of the observations, however, is a county with a population density of 120 and a crime rate of 3235.7. This difference of 644.9 between the estimated and observed crime rate is called the residual.

Residuals are denoted e, so the residual associated with a particular observation *i* is e_i (for each sample observation, $i = 1,2, \ldots, n$). It is useful to distinguish the residual e from the error term ϵ:

- $e = Y_i - \hat{Y}_i$ represents the deviation of the *observed* Y from the *estimated* line. e is an observable variable.
- $\epsilon = Y_i - E(Y_i)$ represents the deviation of the *observed* Y from the expected line. ϵ is a conceptual variable, because the expected relationship is not observable.

Given that we want to estimate β, how do we do it? Consider our data on crime and density, which are plotted in Figure 2.

If we draw a line through these data points, we are in effect estimating β. The values of β associated with this particular line (e.g., $\acute{Y} = \acute{\beta}_0 + \acute{\beta}_1 X$) can be read off the axis using any 2 points on the line. The line, of course, does not perfectly fit the data; no line will unless Y

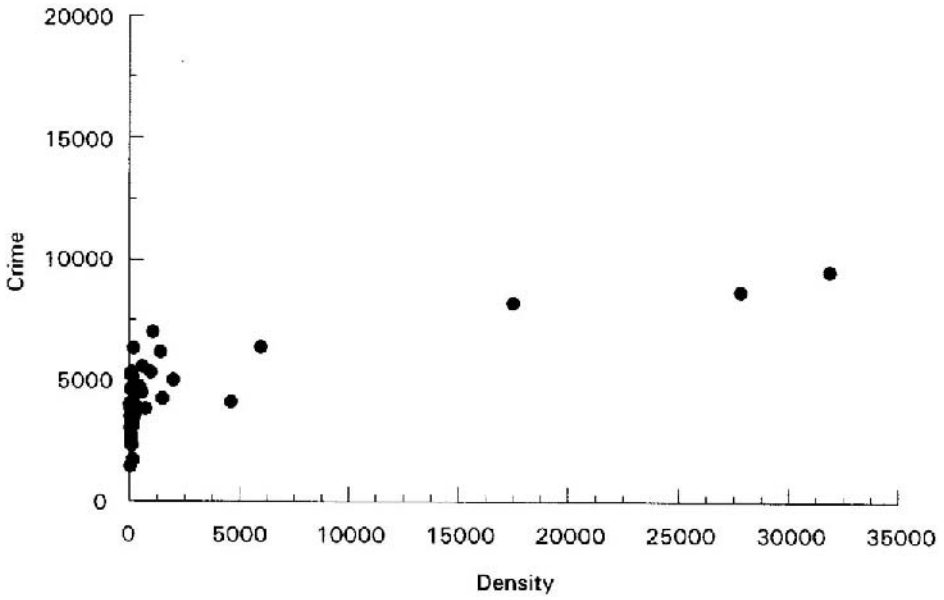


FIGURE 2 Scatter plot.

and X have an exact linear relationship. These differences between the points on the line (the predicted value of crime rate) and its observed value (the data point) are the residuals (e).

Notice that we could draw many other lines through these same data points, and each line would estimate a different β and produce a different set of residuals. How do we know the ''correct'' line to draw? i.e., what are the best estimates of β that can be derived from these data?

Obviously, we would like the predicted values of the dependent variable to be as close as possible to its actual values. In other words, we would like to minimize the difference between Y_i and $\hat{Y}_i$ for all observations of Y, or:

$$\min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)$$

This criterion, however, will not yield a unique line. Moreover, it has the disadvantage of allowing large positive errors to cancel large negative errors, so some lines satisfying this criteria could have very large residuals.

This problem could be avoided by either minimizing the absolute value of the residuals, or the squared residuals. Large residuals would then be penalized regardless of their sign. Minimizing the squared residuals has the added advantage of penalizing large outliers much more proportionally than small ones. Moreover, the squared term is easier to manipulate mathematically than the absolute value term.

The criterion of minimizing the sum of the squared residuals to obtain the best line through a set of data, or more precisely:

$$\min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

TABLE 2 Estimation Output

Variable	Coefficient	Std. error	t-ratio	Prob |t| ≥ x
Constant	674.68	981.9	.687	.49474
X1	.11902	0.2907E-01	4.095	.00013
X2	210.34	113.4	1.855	.06870
X3	.27481	0.8171E-01	3.363	.00137

is called *ordinary least-squares estimation* (or OLS). This is the method used in regression analysis, and the criterion can be shown to yield the following unique estimates of β:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{6}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{7}$$

Equations 6 and 7 are the OLS formula for the parameters in the simple two-variable regression model.

Parameter estimation in the multivariate regression model is analogous to the simple case, but considerably more laborious. Estimation requires the simultaneous solution of a system of linear equations. More precisely, for a model with *k* independent variables, the OLS estimates of β_0 through β_k are determined by the following set of equations:

$$\beta_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 - \ldots - \hat{\beta}_k \overline{X}_k$$

$$\sum_{i=1}^{n} y_i x_{i1} = \sum_{i=1}^{n} x_{i1}^2 \hat{\beta}_1 + \sum_{i=1}^{n} x_{i1} x_{i2} \hat{\beta}_2 + \ldots + \sum_{i=1}^{n} x_{i1} x_{ik} \hat{\beta}_k$$

$$\sum_{i=1}^{n} y_i x_{i2} = \sum_{i=1}^{n} x_i x_{i2} \hat{\beta}_1 + \sum_{i=1}^{n} x_{i2}^2 \hat{\beta}_2 + \ldots + \sum_{i=1}^{n} x_{i2} x_{ik} \hat{\beta}_k$$

$$\sum_{i=1}^{n} y_i x_{ik} = \sum_{i=1}^{n} x_{i1} x_{ik} \hat{\beta}_1 + \sum_{i=1}^{n} x_{i2} x_{ik} \hat{\beta}_2 + \ldots + \sum_{i=1}^{n} x_{ik}^2 \hat{\beta}_k$$

where:

$$y_i = Y_i - \overline{Y} \; x_{ik} = X_{ik} - \overline{X}_k$$

The most important function of a regression software package is to solve this system of equations for $\hat{\beta}_0$ through $\hat{\beta}_k$.

To estimate a multivariate regression model using one of the many available software packages (e.g., SAS, SPSS), one need only input the data and the model statement (formatted as required by the software). The output will include a variety of summary statistics on the dependent variable (e.g., mean and standard deviation) and the estimated model (e.g., R^2, *n*), followed by the estimates of the coefficients and their associated standard errors and t-statistics.

For example, the multivariate model of crime rates discussed in §I was estimated using the software package LIMDEP.[4] The portion of the output that contains the parameter estimates is presented in Table 2. The second column provides the estimated coefficients that were discussed in §I. The third column provides the standard errors, estimates of the standard deviations

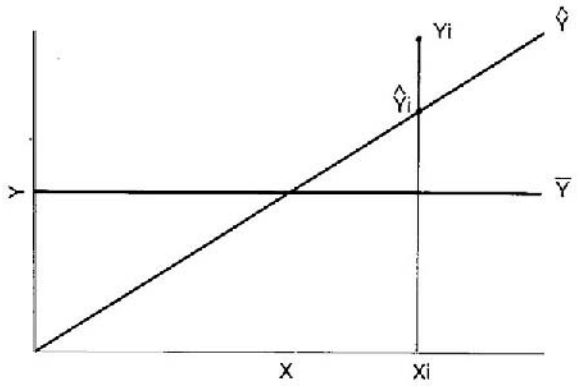


FIGURE 3 Predicting Y.

associated with the estimated coefficients. The interpretation of the other information and how one evaluates an estimated model in general is discussed in the next section.

III. EVALUATION

Once we obtain estimates of our regression model coefficients, we must evaluate the results. There are two aspects to the evaluation: how well does the regression model fit the data? how well do the estimated coefficients conform to our *a priori* expectations?

A. Goodness-Of-Fit

After obtaining coefficient estimates, the obvious question to ask is: how well does the model fit the data? or, equivalently, how well does the regression model explain variations in the dependent variable?

Let's begin by considering the general problem of predicting some variable, Y. If we only have observations on Y, then the best predictor of Y is the sample mean. For example, if we want to predict an individual's weight and the only information available is the weight of a representative sample of individuals, then the sample mean is the best predictor of the individual's weight. However, what if we believe that height is related to weight? Then, knowing an individual's height (X) should improve our predictions of weight (Y).

Consider a particular observation, Y_i. Without knowing X_i, the best guess for Y_i would be the sample mean, $\overline{Y}$, and the error in this guess is $Y_i - \overline{Y}$.

By using knowledge of the relationship between X and Y, we can improve that prediction, knowing X_i leads to the prediction of $\hat{Y}_i$ (Figure 3). So we have ''explained'' part of the difference between the observed value of Y_i and its mean. Specifically, we have explained $\hat{Y}_i - \overline{Y}$. But, $Y_i - \hat{Y}_i$ is still unexplained. To summarize:

$Y_i - \overline{Y}$ = the total deviation of the observed Y_i from $\overline{Y}$
$\hat{Y}_i - \overline{Y}$ = the portion of the total deviation explained by the regression model
$Y_i - \hat{Y}_i$ = the unexplained deviation of Y_i from $\overline{Y}$

We can calculate these deviations for each observation. If we square them (to avoid cancel-

lation of deviations with opposite signs), we can calculate the total deviation and the explained and unexplained portions for all observations. Specifically, the sum-of-squared deviations are:

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \textit{total deviation}$$

$$\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \textit{explained deviation}$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \textit{unexplained deviation}$$

Our goal in predicting Y is to explain the deviation of the observed values from the sample mean. Recall that the *unexplained* portion of these deviations is the quantity that OLS estimation minimizes. Therefore, one measure of how well the regression model explains the variation in the dependent variable is the ratio of explained to total deviation. This measure is called the *coefficient of determination* or R^2. Specifically,

$$R^2 = \frac{\textit{explained deviation}}{\textit{total deviation}} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \tag{15}$$

R^2 is the *proportion* of the variance in the dependent variable explained by all the independent variables. A high R^2 implies a good overall fit of the estimated regression line to the sample data. Since $0 \leq R^2 \leq 1$, an R^2 close to 1 indicates a very good linear fit (most of the variance in Y is explained by X), while an R^2 near 0 indicates that X doesn't explain Y any better than its sample mean. R^2 is thus an easy to understand, and almost universally used, measure of the goodness-of-fit of the regression model.

Obviously, when we estimate a regression model we would like a high R^2, since we want to explain as much of the variation in Y as we can. We must be careful, however, not to blindly pursue the goal of a high R^2. There are three issues to consider in an evaluation of an R^2.

First, a high R^2 does not necessarily mean a *causal* explanation, merely a statistical one. For example, consider the following model of the amount of household consumption in year *t*:

$$Y_t = \beta_0 + \beta_1 Y_{t-1}$$

This model is likely to have a very high R^2, but last year's consumption (Y_{t-1}) does not *cause* this year's consumption. The two are merely highly correlated because the key causal variable (income) is likely to be similar over the two years.

Second, a low R^2 may simply indicate that the relationship is not linear. For example, if one attempts to fit a line to the relationship $Y = X^2$ for X ranging from -100 to 100, as illustrated in Figure 4, the R^2 for the OLS estimated line will be 0 even though the relationship between X and Y is an exact *nonlinear* one.

Finally, if a high R^2 is the only goal, it can be achieved by adding independent variables to the regression model. Additional independent variables cannot lower R^2. In fact, if there are $n - 1$ independent variables (where *n* is the number of observations), then $R^2 = 1$ regardless of the independent variables. For example, if there are only two observations and a model with one independent variable, then the line will fit the data perfectly regardless of which independent variable is used. Adding a third observation will destroy the perfect fit, but the fit will still be

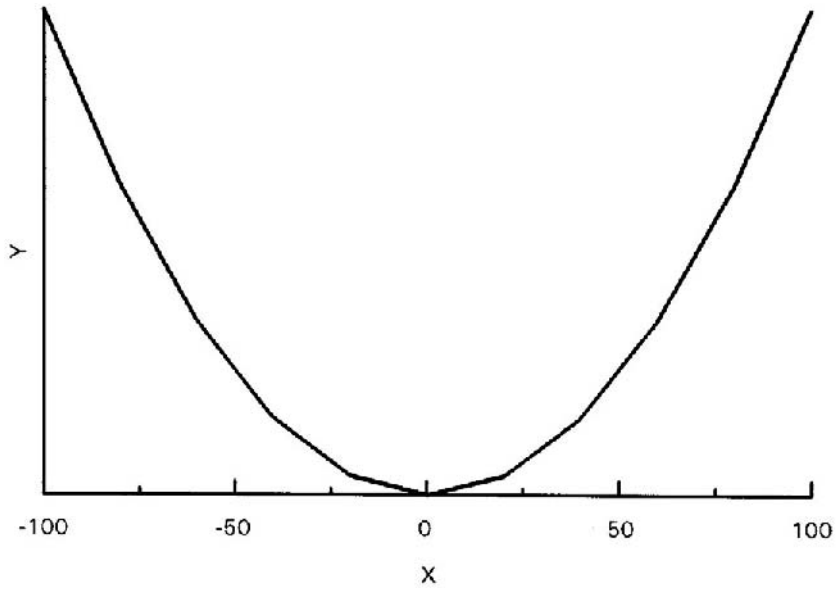


FIGURE 4 Y = X Squared.

good simply because there is only 1 observation to ''explain''. This difference between the number of observations (n) and the number of independent variables (k) is called the *degrees of freedom* (df). Specifically:

$$df = n - k - 1$$

The greater the degrees of freedom, the more reliable or accurate the estimates are likely to be.

A second measure of the goodness-of-fit of an estimated model is the F-test. The F statistic is the ratio of the explained deviation to the unexplained deviation, adjusted for the degrees of freedom, or:

$$F = \frac{(\textit{explained deviation})/k}{(\textit{unexplained deviation})/(n\text{-}k\text{-}1)} \tag{17}$$

This statistic is used to conduct a significance test for the overall fit of the estimated equation. A ''high'' F value indicates it is unlikely that we would have observed the estimated parameters if *all* the true parameters are zero. Intuitively, the set of independent variables has little explanatory power if the explained variation is small relative to the unexplained variation.

Most software packages provide the results of the F-test. But, note that it is a relatively weak test. R^2 provides more information, and is thus the more common measure of model goodness-of-fit.

B. Coefficient Expectations

Determining the extent to which coefficient estimates conform to our *a priori* expectations has two elements—statistical significance, and expectations about signs and magnitudes.

1. Statistical Significance

The first assessment of a coefficient estimate should be to determine if it is statistically different from zero, for this reveals whether a relationship was found in the sample data. Statistical significance is assessed using tests of whether the observed sample estimate is likely to have occurred even if the population value is zero. The appropriate significance test for individual coefficients in a regression analysis is the *t-test*. The t-test is based on the t-statistic, which is computed as:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \; j = 1, 2, \ldots, k \tag{18}$$

where s is the standard error of the estimated coefficient, $\hat{\beta}_j$.

The t-statistic associated with the null hypothesis of β *equal to zero* is routinely reported with the regression results in most regression software packages (see, for example, Table 2, column 4). If not, it is easily calculated by dividing the parameter estimate by its standard error.

A *high* absolute value of t_j indicates it is *unlikely* that we would have observed the sample estimate *if* $\beta_j = 0$ in the underlying population. If one obtains a high *t* value, the inference is that $\beta_j \neq 0$. The magnitude of *t* considered to be ''high'' depends on the degrees of freedom and the selected probability level (the level of significance). These critical t-statistic values can be found in collections of statistical tables or econometric textbooks, but increasingly the probability level associated with the *t* value is generated by the regression software and included as a standard part of the regression results (e.g., see Table 2, column 5).

Selecting the probability level that denotes statistical significance is a subjective decision. A widely-used value is 5%. If one obtains a t-statistic larger than the critical value for the 5% level (larger than an absolute value of 2.0 for most samples), the interpretation is that only 5% of the time would we expect to observe this particular value of the coefficient estimate *if* the true coefficient is zero. The implication then is that $\beta \neq 0$. If one prefers more certainty before making this inference, then a lower probability level (e.g., 1%) can be selected. If one is comfortable with a larger potential error, then a higher probability level (e.g., 10%) can be selected.

Finally, note that it is certainly not necessary that all coefficients ''pass'' a t-test. Failure means that the hypothesized relationship was not observed in this particular sample, not that there isn't a relationship in the population under study. Such ''negative'' information is, in fact, quite useful. If researchers repeatedly find no effect across different samples, this becomes persuasive evidence that the hypothesis, and its underlying theory, need to be reconsidered.

2. Signs and Magnitudes

Once the statistically significant coefficients have been identified, their sign and magnitude should be checked against *a priori* expectations. In most cases, we will have a specific expectation with respect to the sign of the coefficient, because sign expectations derive directly from our hypotheses. For example, if we expect increasing population density to increase criminal activities, we expect a positive coefficient on density; if we expect crime to decrease with an increased police presence, we expect a negative coefficient on our measure of police presence. If the estimated coefficient sign differs from the expected one, it indicates a problem with the underlying theory, variable measurement, or the sample.

We are less likely to have strong expectations about magnitude. Our understanding of most public sector processes is not so well developed that we can identify the magnitude of impacts *a priori*. Nevertheless, an examination of magnitudes can offer important information about the expected impact of interventions, as well as their cost effectiveness, e.g., the reduction

in teenage pregnancies expected to result from an additional $100K spent on a prevention program. We may also care about relative magnitudes—how the magnitude of one regression coefficient compares to another. These are considered next.

C. Relative Importance of Independent Variables

Sometimes, we want to make statements about the *relative* importance of the independent variables in a multiple regression model. For example, recall our multivariate model of crime as a function of density, the high school dropout rate and income (Equation 3). $\hat{\beta}_2$ is larger than $\hat{\beta}_1$; does this imply that the dropout rate (X_2) is more important than population density (X_1) in determining the crime rate? It does not; one *cannot* directly compare the magnitudes of estimated regression coefficients, because the variables are measured in different units. The absolute value of a coefficient is easily changed simply by changing the measurement scale of the variable.

In order to compare the impact of different independent variables on the dependent variable, we must calculate *standardized coefficients* (sometimes called beta coefficients). These are produced by standardizing each variable (by subtracting its mean and dividing by its estimated standard deviation), and then estimating the model using the standardized values.

$$(Y_i - \overline{Y})/s_Y = \alpha_1(X_{1i} - \overline{X}_1)/s_{X_1} + \ldots + \alpha_k(X_{ki} - \overline{X}_k)/s_{X_k} + \epsilon_i \tag{19}$$

Since all the variables are now measured in the same units (all standardized variables have a mean of zero and a variance of one), their coefficients can be directly compared. If the magnitude of $\hat{\alpha}_1$ exceeds $\hat{\alpha}_2$ then one can state that X_1 has a greater influence on Y than X_2.

Standardizing the variables in our crime model and estimating a model analogous to Equation 19 yields the following estimates of $\hat{\alpha}$:

$$\hat{\alpha}_1 = .56 \; \hat{\alpha}_2 = .23 \; \hat{\alpha} = .24$$

Thus we can now say that population density has about double the impact of either the dropout rate or income on the crime rate. In practice, one rarely has to run a separate standardized estimation since most regression packages will, upon request, provide standardized coefficients with the regression coefficients.

The *interpretation* of a standardized coefficient is in standard deviation terms; it provides the average standard deviation change in the dependent variable resulting from a unit standard deviation change in the independent variable. For example, the standardized coefficient $\hat{\alpha}_1$ is interpreted to mean that a 1 standard deviation change in population density will lead to a .56 standard deviation change in the crime rate.

Recall that the regression coefficient provides the average change in the dependent variable resulting from a unit change in the independent variable, and is thus in the dependent variable's unit of measurement. The relatively more awkward interpretation of standardized coefficients has limited their use. Nevertheless, standardized regressions are common in some fields, like psychology, where the variables may not have a ''natural'' unit of measurement.

IV. CLASSICAL ASSUMPTIONS

The use of Ordinary Least Squares (OLS) as the best estimation method for regression models is based on the regression model satisfying a set of assumptions. If these assumptions are *not* satisfied, we may have to consider an alternative estimating technique. These assumptions are called the ''classical assumptions''. For convenience, I have grouped them into four assump-

tions. First, I will simply state the assumptions, then for each, we will discuss how to identify assumption violations, their effects, and available remedies.

I. The dependent variable, Y, is a linear function of a specific set of independent variables, X, plus an error term, i.e.:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$$

This assumption implies that:

- the relationship between Y and X is linear
- no relevant independent variables have been excluded
- no irrelevant independent variables have been included

II. The observations on the independent variables can be considered fixed in repeated sampling.
This assumption implies that the observed values of the Xs are determined *outside* the model and thus independently of the values of the error term, i.e., the Xs and the ϵs are uncorrelated, or $COV(X\epsilon) = 0$.

III. The error term, the random variable ϵ, is assumed to satisfy the following four conditions:
   A. The error term has a zero *population* mean, or $E(\epsilon_i) = 0$. A violation of this assumption yields a biased intercept.
   B. The error term has constant variance for all observations, or $VAR(\epsilon_i) = \sigma^2$ for all observations *i*. Such an error term is called *homoskedastic*. If, alternatively, its variance is changing across observations, the error term is said to be *heteroskedastic*.
   C. The error term for one observation (ϵ_i) is not systematically correlated with the error term for another observation (ϵ_m), or $COV(\epsilon_i \epsilon_m) = 0$ for all observations $i \neq m$. In other words, a random shock to one observation does not affect the error term in another observation. The violation of this assumption is called *autocorrelation*.
   D. The error term is normally distributed.

IV. The number of observations exceeds the number of independent variables, and there are no exact linear relationships among the independent variables.
Perfect collinearity occurs if two variables are the same except for a scale factor or the addition of a constant, e.g.: $X_1 = aX_2 + b$. In this case, movements in one variable would be the same as movements in another, except for the scale factor. Therefore, one could not differentiate their effects on the dependent variable.

Assumptions I–IV are the classical assumptions. Their importance derives from the *Gauss-Markov Theorem*, which states that *if* all the above assumptions hold,[5] then the OLS estimates of β are the *best linear unbiased estimators of* β. Consider what this means.

The processes of sampling and estimation are used to gain information about a population by estimating its parameters from sample data. These sample estimators have characteristics with respect to the population parameters. Two of the most important are bias and the extent of dispersion.

If the sample estimator has an expected value equal to the population parameter it is estimating, it is said to be *unbiased*, i.e., $E(\hat{\beta}) = \beta$. This means that if we draw repeated samples from the population and calculate the sample estimator, their average will be equal to the true population parameter. Obviously, we would like our estimator to have this characteristic.

Another important characteristic is the extent of dispersion of the sample estimator around the population parameter. The larger the variance, the more likely it is that a given sample will produce a value for the sample estimator that is very different from the population parameter. Consequently, we would like to minimize this dispersion.

When a sample estimator is unbiased and has the smallest variance among the set of unbiased estimators, it is said to be *efficient*. The Gauss-Markov Theorem assures us that if the classical assumptions hold, the OLS estimator is efficient and thus the best estimator of the population parameters, i.e., the best measure of the relationship between the independent and dependent variables of interest to us.

Therefore, the first step in estimating a regression model is to determine if the classical assumptions are satisfied, and thus whether or not OLS is the appropriate estimating method. We begin by considering the *normality assumption*. Although the error term need not be distributed according to a normal distribution in order for OLS estimates to be efficient, the use of significance tests requires it.

In most cases, we just assume that the error term has a normal distribution. The Central Limit Theorem states that the distribution of the sum of independent variables (e.g., the error term) approaches the normal distribution as the sample size increases, regardless of the individual distributions of these random variables. Therefore, with large samples, the Central Limit Theorem assures us that the error term is normally distributed. With small samples, we are less confident. Cassidy (1981) demonstrates that the tendency toward a normal distribution occurs at sample sizes as small as 10 observations. Nevertheless, one should be skeptical about the reliability of significance tests with very small samples.

What about the other assumptions? The implications of and solutions for assumption violations vary according to the violation, so we discuss each separately. In Section V through VIII we discuss Assumptions I through IV respectively. Table 3 summarizes some of the important relationships to which we will refer in these discussions.

V. MODEL SPECIFICATION

Assumption I states that the dependent variable can be expressed as a *linear* function of a *specific* set of independent variables, plus an error term. This assumption is critical. The theoretical model must be correct, otherwise no inferences can be made.

The theory that underlies our understanding of the dependent variable should determine both the nature of the relationship (e.g., linear or not) and the selection of explanatory variables. As noted earlier, the linear specification has proven to be very robust in specifying a broad range of phenomena of interest to public sector scholars and practitioners. We will thus focus our attention on the set of independent variables. We begin by considering the consequences of selecting the wrong set of independent variables—by excluding relevant variables, or by including irrelevant variables, followed by an evaluation of a frequently used selection method—stepwise regression. We then consider two other specifications issues—the measurement level of the variables, and nonlinear relationships that can be handled within the regression framework.

A. Excluded Relevant Variables

Consider first the case where a variable we expect to influence the dependent variable cannot be included in the regression model (e.g., data are unavailable). What are the implications for

TABLE 3 Important Relationships

Bias	
$E(\hat{\beta})$ and β	Bias present if $COV(X\varepsilon) \neq 0$
	Bias present if $E(\varepsilon) \neq 0$
	Bias decreases as VAR(X) increases
$E(s^2_{\hat{\beta}})$ and $VAR(\hat{\beta})$	Bias present if $VAR(\varepsilon)$ not constant
	Bias present if $COV(\varepsilon_i\varepsilon_j) \neq 0$
$E(s_2)$ and σ^2_ε	Bias present if $VAR(\varepsilon)$ not constant
	Bias present if $COV(\varepsilon_i\varepsilon_j) \neq 0$
Dispersion	
$VAR(\hat{\beta}_i)$	Increases as $VAR(\varepsilon)$ increases
	Increases as $COV(X_iX_j)$ increases
	Decreases as $VAR(X_i)$ increases
$s^2_{\hat{\beta}}$	Decreases with increases in *df*
	Increases as residual variance increases
	Decreases as VAR(X) increases
	Increases as $COV(X_iX_j)$ increases
Hypothesis testing	
t-test	Invalid if $E(s^2_{\hat{\beta}}) \neq VAR(\hat{\beta})$
	Invalid $E(\hat{\beta}) \neq \beta$
t	Decreases as $VAR(\hat{\beta})$ increases

the regression model? More precisely, assume the model we want to estimate (ignoring the intercept) is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{20}$$

Instead, X_2 is omitted from the model and we estimate the following model (where μ is an error term):

$$Y = \beta_1 X_1 + \mu \tag{21}$$

Obviously the latter model will explain less of the variation in Y, but what about the information we obtain? Is the estimate of β_1 affected?

Recall that β_j represents the change in the expected value of the dependent variable given a unit change in X_j holding other independent variables constant. If a relevant independent variable is not included in the model, it is *not* being held constant for the interpretation of β_j—thus the estimate of β_j may be biased.

Whether or not $\hat{\beta}_j$ is biased depends on whether the *omitted* variable is correlated with the set of *included* independent variables. More precisely, in our example, the expected value of $\hat{\beta}_1$ can be shown to be:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2\, COV(X_1X_2)/VAR(X_1) \tag{22}$$

In order for $\hat{\beta}_1$ to be unbiased, i.e., for $E(\hat{\beta}_1)$ to equal β_1, it must be the case that *either* $\beta_2 = 0$ (X_2 is *not* a relevant variable) *or* $COV(X_1X_2) = 0$ (X_1 and X_2 are uncorrelated). Otherwise, the second term on the left side of Equation 22 is not zero, and the estimate of β_1 will be biased.

Intuitively, when X_1 and X_2 are correlated and X_2 is not included in the estimation, $\hat{\beta}_1$ will pick up the effect of X_2 on Y as well as the effect of X_1 on Y, hence the bias. If X_1 and X_2 are uncorrelated, $\hat{\beta}_1$ doesn't incorporate the impact of X_2 (it would be reflected in the error term) and no bias is introduced. For example, assume we believe that salary differences in an

organization can be explained by experience, education, and motivation. Unfortunately, we have no measure of motivation. If we estimate the model without motivation *and* motivation is unrelated to experience and education, then the sample estimates of β will be good indicators of the extent to which education and experience explain salary differences. If, however, highly motivated individuals are, for example, likely to be more educated (or more educated individuals are more motivated) then the effect of education on salaries will be overestimated, as it will include both the effect of education and the effect of motivation on salaries.

Unfortunately, for many public policy and administration applications, it is unlikely that two determinants of a variable are completely uncorrelated, so bias must always be considered when one excludes a relevant variable. Moreover, the probable presence of bias with an omitted relevant variable makes t-tests invalid. The use of t-tests requires that the parameter estimate be unbiased.

In summary, excluding a relevant variable from a regression model is quite serious. It is likely to yield biased parameter estimates (if the excluded variable is correlated with included independent variables) and invalid *t*-tests. Given this undesirable outcome, researchers may be tempted to include any variables that might be relevant. Consider then what happens if one includes irrelevant variables.

B. Included Irrelevant Variables

Assume that the dependent variable of interest is only determined by X_1, i.e., the true model (ignoring the intercept and with μ as the error term) is:

$$Y = \beta_1 X_1 + \mu \tag{23}$$

We, however, include X_2 in the model in error and estimate:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{24}$$

According to Equation 22, $E(\hat{\beta}_1) = \beta_1$ because $\beta_2 = 0$. Therefore, the parameter estimate of the relevant variable is *unbiased*, even if one includes an irrelevant variable in the model.

The variance of $\hat{\beta}_1$, however, *increases* unless $\mathrm{COV}(X_1 X_2) = 0$. To see this, consider the variance of $\hat{\beta}_1$ for the estimated model (Equation 24):

$$VAR(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n} (X_{i1} - \overline{X})^2 (1 - r_{12}^2)} \tag{25}$$

r_{12} is the Pearson correlation coefficient between X_1 and X_2, and captures the covariation between the two variables. According to Equation 25, if $r_{12} \neq 0$, $\mathrm{VAR}(\hat{\beta})$ increases, which means that the OLS estimates no longer have the smallest variance. In addition, the increased variance means that the t-statistic associated with X_1 is lowered (s_β is its denominator), making it harder to reject the null hypothesis.

So, how do we determine the variables to be included in a regression model? A well developed theory of the determinants of the dependent variable is required. If theory is uncertain, or data are not available on some variables, there are important costs. To summarize them:

- Excluding a relevant variable usually leads to biased parameter estimates, which means that the parameter estimate may be substantially incorrect in both sign and magnitude, *and* t-tests are invalid.
- Including an irrelevant variable yields inefficient parameter estimates, and underestimates the value of *t*, making it harder to pass a significance test.

If one is uncertain about whether or not a measurable variable should be included in a model, it is usually better to include it since the consequences of excluding a relevant variable are more dire than those of including an irrelevant one.

C. Stepwise Regression

Consider now a common and problematic strategy for selecting variables for inclusion in the estimating equation, stepwise regression. This procedure adds explanatory variables to a model based on their marginal contribution to R^2 or based on their t-statistic value. There are four problems with this approach:

1. *Invalid population inferences*. Stepwise regression ignores the theoretical aspects of estimation, which can lead one to make incorrect population inferences from the sample characteristics. For example, the sample may contain characteristics that are not important in the population (or vice versa).
2. *Invalid t-tests*. T-statistics no longer have the t-distribution because the selection procedure makes it more likely that a variable has a large *t*.
3. *Arbitrary model*. The final equation is arbitrary if there is correlation among the independent variables because the order in which they are considered will affect whether or not a variable is included in the model.
4. *Biased parameter estimates*. Parameter estimates *may* be biased since the procedure makes it easy to exclude relevant variables.

One should thus not rely on this technique to select independent variables for a regression analysis. Rather, one should rely on the prevailing understanding of the phenomenon under study.

There is, however, a legitimate use for stepwise regression. If there are no existing theories that explain the phenomenon of interest, then a stepwise technique could be used for hypothesis generation. Identifying the variables that have a large impact on the dependent variable in a particular sample may suggest possible hypotheses that can be explored using *another* data set.

D. Measurement Level of Variables

Regression analysis requires the use of *interval data*, variables with values that are ordered and scaled, like, for example, income. Many variables of interest to public administration, however, are noninterval. There are two types of noninterval variables—nominal and ordinal.

Ordinal variables can be ordered, but the distance between the values cannot be measured. For example, political interest can be categorized as ''not interested'', ''somewhat interested'', or ''very interested''. The order of the values is clear; on a scale of political interest, ''very interested'' is greater than ''somewhat interested'' and ''somewhat interested'' is greater than ''not interested''. The distance between the values, however, is not quantified; for example, one cannot say how much more political interest is represented by ''very interested'' as compared to ''somewhat interested.'' *Nominal* variables cannot be ordered, e.g., religious affiliation, gender. Values are grouped into named categories, but there is no order to the groupings.

Whether one should use regression analysis with noninterval data depends on which variables are noninterval. If the *dependent* variable is noninterval, one should use a qualitative dependent variable model (like logit or probit), not regression analysis. Noninterval *independent* variables, however, can be included in a regression model through the use of dummy variables.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1. Dummy Variables

Dummy Variables are variables that take on only two values, for example, X = 1 if an individual is a male, 0 if a female. A dummy variable can be included in a regression model as an independent variable and the OLS estimate of the coefficient remains efficient. The interpretation of the coefficient, however, changes.

Consider an example of a simple regression model of income as a function of gender:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{26}$$

where Y = income and X = 1 if male, 0 if female.

Note that for women, the expected value of Y is β_0 (X = 0 and $E(\epsilon) = 0$). β_0 is thus the average level of income for women. Similarly, for men, the expected value of Y is $\beta_0 + \beta_1$ (X = 1 and $E(\epsilon) = 0$). So, $\beta_0 + \beta_1$ is the average level of income for men.

The coefficient on the dummy variable, β_1, is thus interpreted as the *average change in income resulting from being male rather than female*. Within a more complete model of income determination that includes the key determinants of income differences, this parameter estimate would provide a measure of the average income differences that result from gender discrimination. Note how this interpretation compares to that of an interval variable, where the coefficient represents the average change in the dependent variable resulting from a unit change in the independent variable.

Noninterval independent variables with more than two categories can also be included in the regression model using dummy variables. As an example, consider Graddy and Nichol's 1990 study of the effects of different organizational structures on the rate of disciplinary actions by occupational licensing boards.

We modeled disciplinary actions per licensee as a function of several board and profession-specific variables, as well as the degree to which the board functions as an independent agency. Organizational structure was defined as an ordinal variable, with boards categorized as either independent, sharing power with a centralized agency, or subsumed within a centralized agency. This qualitative indicator of organizational structure can be used in a regression model by creating the following 3 dummy variables:

I = 1 if the board is independent, and 0 otherwise
S = 1 if the board shares power, and 0 otherwise
C = 1 if the board is controlled by a centralized agency, 0 otherwise

The estimated model can only include two of these dummy variables. Including all 3 would create perfect collinearity (since $I + S + C = 1$ for all observations) and the model could not be estimated.[6] Therefore, estimation requires that one category be omitted.[7] The omitted category serves as a base group with which to compare the others. For example, consider the independent boards as the base case. The model to be estimated is:

$$Y = \beta_1 S + \beta_2 C + \beta X + \epsilon \tag{27}$$

where Y is the rate of disciplinary actions, and **X** denotes all other included determinants of disciplinary actions.

β_1 represents the *difference* in the average number of disciplinary actions of boards that share power *compared* to independent boards. β_2 represents the difference in the average number of disciplinary actions of centralized boards compared to independent boards.

Estimation of this model revealed negative and significant coefficients on both S and C. This implies that both boards that share power with a centralized agency and those that are controlled completely by a centralized agency produce fewer disciplinary actions than indepen-

dent licensing boards. Note that these results are always interpreted relative to the base (omitted) category.

What if we want to compare the performance of centralized and shared-power boards? This requires that we re-estimate the model with centralized boards as the omitted category. In this estimation, the coefficient on S was *not* significantly different from zero, which implies there is no significant difference between the disciplinary performance of boards that share power and those that are fully centralized.

Finally, some researchers, to conserve on degrees of freedom, convert nominal variables to ordinal variables. For example, we could view the organizational structure of licensing boards in terms of their degree of centralization, and construct a variable Z that equals 1 if the board is independent; 2 if the board shares power; and 3 if the board is centralized.

Estimation of Equation 27 with Z substituted for S and C produced a negative and significant coefficient. The coefficient on an ordinal variable should be interpreted as one would an interval variable, e.g., the average decrease in disciplinary actions associated with a unit increase in centralization.

There is nothing wrong with this approach *if* the underlying scale (e.g., centralization) makes sense. But, this estimation strategy yields less precise information. In this case, for example, the estimations using dummy variables revealed that shared-power boards and centralized boards behave about the same with respect to disciplinary actions, but significantly different from independent boards. The estimation using the ordinal variable suggested that centralized boards produce fewer disciplinary actions than those that share power—which is not the case.

E. Alternative Functional Forms

Thus far we have specified a linear relationship between the dependent and independent variables, but the regression model actually allows more latitude. The regression model in fact requires only that the model be *linear in its parameters*, which means that some nonlinear relationships can be used. Several specifications that capture nonlinear relationships within the regression framework can be found in the literature. We consider here two of the more common—interaction effects and quadratic relationships.

1. Interaction Effects

What if the expected effect of an independent variable depends on the level of another variable? Consider, for example, the determinants of volunteering. According to Sundeen (1990), marital status and family size are important determinants of volunteering activity. In particular, he argues that it is unlikely that *single parents* have the time to volunteer, although single individuals without children may. This suggests the need for an interaction term. Consider, for example, the model:

$$Y = \beta_1(X_1X_2) + \beta X + \epsilon$$

where

Y = the number of hours volunteered
X_1 = 1 if an individual is single, 0 otherwise
X_2 = the number of children
X denotes other determinants of volunteering

The interaction variable is the product of ''singleness'' and the number of children. If one is not single or has no children this variable becomes zero and drops out of the model. The

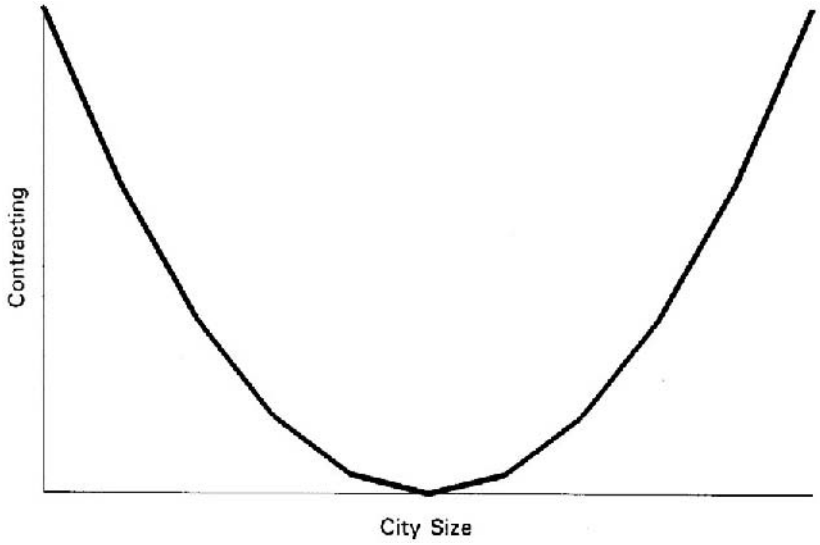


FIGURE 5 U-shaped relationship.

interaction variable will thus capture the impact of being a single parent on volunteering behavior. The estimate of β_1 is expected to be negative.

Note, that if theory supports it, we may also include one or both variables separately in the equation. For example, if we believe that more children limit volunteering time for married parents as well as single parents, we could add X_2 to the model as a separate variable. One would then interpret the impact of an additional child on volunteering as β_1 for single parents plus the coefficient on X_2 for married parents.

2. *Quadratic Relationships*

Ferris and Graddy (1988) argue that the relationship between the contracting decision of cities and their size is not linear. Small cities may want to contract out services to gain the advantages of scale economies, but large cities may contract out more than smaller cities because they have a wider selection of available suppliers. This suggests the u-shaped relationship between the probability of contracting and city size depicted in Figure 5.

This u-shape represents a quadratic relationship, which can be captured in a regression model for a variable X by including the following specification in the model: $\beta_1 X + \beta_2 X^2$. For the contracting example, the model is:

$$Y = \beta_1 X_1 + \beta_2 X_1^2 + \beta X + \epsilon$$

where

- Y denotes the incidence of contracting
- X_1 denotes city size
- **X** denotes other determinants of contracting

The interpretation of the coefficients depends on the signs of $\hat{\beta}_1$ and $\hat{\beta}_2$. If $\hat{\beta}_1$ is negative and significant this supports the economies-of-scale hypothesis; contracting decreases with increasing city size. If β_2 is positive and significant, this supports the importance of available suppliers; after some size, contracting increases with increasing city size. The u-shaped quadratic

hypothesis is only supported if both $\hat{\beta}_1$ and $\hat{\beta}_2$ are statistically significant and the appropriate sign. One can, of course, hypothesize an inverted u-shaped relationship between variables with the sign expectation on the coefficient estimates being reversed.

VI. ASSUMPTION II VIOLATIONS

According to Assumption II, in order for OLS to provide the best estimates of β, it must be the case that the independent variables are uncorrelated with the error term. To see the rationale for this assumption, consider (ignoring the intercept) the simple regression model: $Y_i = \beta X_i + \epsilon_i$. It can be shown that:[8]

$$E(\hat{\beta}) = \beta + \frac{\sum_{i=1}^{n} (X_i - \bar{X})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{30}$$

If, as is stated in Assumption II, the Xs are fixed in repeated samples, then the correlation between X and ϵ (the numerator of the second term) is zero, and $\hat{\beta}$ is unbiased. If, however, X and ϵ are correlated, the second term is not zero, and the expected value of $\hat{\beta}$ will not be β, i.e., $\hat{\beta}$ will be a biased estimator.

There are two common situations that violate Assumption II—an independent variable measured with error; and an independent variable determined in part by the dependent variable. The former situation is called ''errors in variables''; the latter ''simultaneous causality.'' We consider each in turn.

A. Errors in Variables

We have thus far assumed that all variables used in regression analysis are measured without error. In practice, there is often measurement error. What then are the implications for our coefficient estimates? Consider two cases—the dependent variable measured with error, and an independent variable measured with error.

1. Dependent Variable Measured with Error

Assume the true dependent variable is Y, but its measured value is Y*. The measurement error can be specified as:

$$Y^* = Y + w \tag{31}$$

where w is an error term that satisfies Assumption III.

The true model (ignoring the intercept) is: $Y = \beta X + \epsilon$. According to Equation 31, $Y = Y^* - w$, so the estimated equation is:

$$Y^* = \beta X + \epsilon + w \tag{32}$$

The error term in this regression of Y* on X is $\epsilon + w$. As long as X is uncorrelated with this error term (and there is no reason to assume otherwise), then the parameter estimates are *unbiased*.

The only effect of measurement error in the *dependent* variable is increased error variance. The measurement error will be reflected in the residuals, increasing them and the estimated

error variance,[9] which in turn inflates the coefficient variance. The practical implication of the increased coefficient variance is a lower *t* value (recall the related discussion in Section V.B), making it more difficult to pass a significance test.

2. *Independent Variable Measured with Error*

Let X be the true independent variable and X* be the measured variable, then measurement error can be represented as:

$$X^* = X + v \tag{33}$$

where v is an error term that satisfies Assumption III.

The true model (ignoring the intercept) is: $Y = \beta X + \epsilon$. According to Equation 33, $X = X^* - v$. Thus, the estimated equation is:

$$Y = \beta X^* + \epsilon - \beta v \tag{34}$$

The error term in this regression of Y on X* is $\epsilon - \beta v$. But, according to Equation 33, v and X* are correlated. Thus, the independent variable in Equation 34 is correlated with the error term, violating Assumption II. In this situation, OLS will produce biased estimates of β. Moreover, the t-statistic is biased and significance tests are invalid.

Thus measurement error in *independent* variables can be quite problematic. Such measurement error can be ignored if it is assumed to be *too small and random* to affect the parameter estimates. If, however, one cannot make that assumption, then an alternative estimation strategy, instrumental variable estimation, is needed.[10]

B. Simultaneous Causality

In some situations, the dependent variable being modeled influences one or more of the independent variables. The process is thus characterized by simultaneous causality. For example, consider the following simple model of national income determination:

$$C_t = \beta N_t + \epsilon_t \ (\beta > 0) \tag{35}$$

where C denotes aggregate consumption and N denotes national income in year t.

Assume, in addition, that national income itself is just the sum of consumption, investment, and government spending, or:

$$N_t = C_t + I_t + G_t \tag{36}$$

where I denotes aggregate investment and G denotes government spending in year t.

Consider a random shock that increases ϵ. According to Equation 35, an increase in ϵ implies that C goes up, and an increase in C, according to Equation 36, will cause an increase in N. But, N is also in Equation 35, and if N goes up then C increases too.

If only Equation 35 is estimated, OLS attributes *both* of the increases in consumption to the increase in income, not just the latter (because ϵ and N are correlated). Therefore, the OLS estimate of β is biased—in this case overestimated. More precisely, recall Equation 30, reproduced here with national income (N) as the independent variable:

$$\hat{\beta} = \beta + \frac{\sum_{t=1}^{n} (N_t - \bar{N})\,(\epsilon_t - \bar{\epsilon})}{\sum_{i=1}^{n} (N_t - \bar{N})^2} \tag{37}$$

N and ϵ are correlated because N is a function of C (see Equation 36) and C is a function of ϵ (see Equation 35). The amount of the bias is the second term in Equation (37).

In general, whenever an independent variable is a function of the dependent variable, OLS will produce biased parameter estimates. The problem is usually identified by theory, in that one must recognize that the relationship is simultaneous. There are two alternative estimation procedures. One involves estimating the single equation of interest using a special case of instrumental-variable estimation, Two-Stage Least Squares. The second requires estimation of all the equations—multi-equation estimation. Both approaches are beyond the scope of this chapter.[11]

VII. ERROR TERM ASSUMPTIONS

Assumption III refers to the assumed distribution of the error term. Specifically, ϵ is assumed to be normally distributed with a zero population mean, constant variance, and no correlation in the error terms across observations. Each aspect of this assumption has implications for estimation. We have already addressed the assumption of normality. Now, the assumption of a zero population mean, correlation across observations, and a constant variance are considered in turn.

A. Biased Intercept

Assumption IIIa states that the error term has a population mean of zero. This means that we believe we have included all important non-random determinants of the dependent variable in our model. However, even if our model is correct, it is possible for this assumption to be violated. There could, for example, be *systematic* positive or negative measurement errors in calculating the dependent variable. The consequences are serious. If $E(\epsilon) \neq 0$, OLS estimation yields biased parameter estimates, i.e., $E(\hat{\beta}) \neq \beta$.

The problem is most easily addressed by forcing the error term to have a zero mean by adding or subtracting the sample error mean to the intercept. For example, if the error term mean for a particular sample is some number d, one could subtract d from ϵ and add d to $\hat{\beta}_0$:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$Y = \beta_0 + d + \beta_1 X + \epsilon - d$$

The two equations are equal since only a constant is added and subtracted, but the latter error term (ϵ-d) has a zero mean. This transformation is exactly how the OLS procedure corrects a non-zero sample error mean. If an intercept (β_0) is included in the estimation, OLS will force the mean of the error term for the *sample* to be zero in its estimation of the intercept. For example, for the simple regression model, $Y = \beta_0 + \beta_1 X + \epsilon$, OLS estimates β_0 as: $Y - \beta_1 X$.

This approach assures that Assumption IIIa is satisfied, and that the estimates of the slope parameters (β_1 through β_k) are unaffected. The estimate of the intercept, however, is affected. The OLS correction produces an unbiased estimate of the *new* intercept ($\beta_0 + d$), but a *biased* estimate of the original intercept (β_0) (if the sample error term mean is in fact non-zero). Thus, we sacrifice a biased intercept estimate for unbiased slope estimates—a trade we are usually quite willing to make since the intercept estimate is usually unimportant theoretically.

The correction has two implications for estimation. First, we cannot rely on the estimate of β_0. In fact, the constant (intercept) can usefully be thought of as a ''garbage-can'' term, as it will contain any systematic sample anomalies. Second, we should always include a constant in a regression model. If it is omitted *and* there is a non-zero error mean in the sample, then

all the parameter estimates will be biased. Non-zero sample error means are particularly likely in small samples.

B. Autocorrelation

Autocorrelation, a violation of Assumption IIIc, occurs when the error terms associated with two or more observations are correlated. This is a common problem in *time-series* models. Consider, for example, a model of the size of Southern California's economy in year t (E_t):

$$E_t = \beta X_t + \epsilon_t$$

The Northridge earthquake was a random shock that caused ϵ_{94} to be negative. Some of this negative effect carried over and affected ϵ_{95} and beyond. This phenomenon of a random shock to one period that carries over into future periods is called autocorrelation.

Observed autocorrelation can result from two possible sources—a random shock or a missing relevant variable. The former, the type described above, is called *pure autocorrelation*. A missing relevant variable can also generate *systematic* patterns in the error term over time. Autocorrelation resulting from this source is really a specification error and can be corrected by including the relevant variable in the model.

Pure autocorrelation is much less likely with *cross-sectional* data since the effects of random shocks to one family or firm do not normally carry over to another. Therefore, in this section we consider only time-series models.

1. Consequences

What effect does autocorrelation have on OLS estimation? Applying OLS to a model that satisfies all the classical assumptions *except* the absence of autocorrelation yields:

- *Unbiased parameter estimates*. The expected value of the parameter estimate is the true β *if* the independent variables do not include a lagged dependent variable (i.e., Y_t does not depend on Y_{t-1}).
- *Inefficient parameter estimates*. The variance of the parameter estimates is inflated, making it less likely that a particular estimate obtained in practice will be close to the true β. VAR($\hat{\beta}$) increases because autocorrelation inflates VAR(ϵ), to which it is positively related.
- *Invalid t-tests*. The t-statistic is incorrect; thus hypothesis testing is invalid. With autocorrelation, s^2 is no longer an unbiased estimator of VAR(ϵ), which leads to bias in $s^2_{\hat{\beta}}$ as an estimator of VAR($\hat{\beta}$). Therefore, the denominator of the t-statistic, $s^2_{\hat{\beta}}$, is biased. The direction of the bias cannot be determined, so hypothesis testing is invalid.

Autocorrelation is sufficiently common and these problems are sufficiently serious that a solution is needed. The usual approach is to use an alternative estimation strategy, Generalized Least Squares (GLS).

2. Generalized Least Squares

GLS is an estimation procedure that allows a general variance-covariance structure, i.e., one that does not require constant variance and the absence of covariation across error terms. To use GLS to correct for autocorrelation, we must specify the nature of the correlation among the error terms.

The most common specification, *first-order autocorrelation*, assumes the error term depends on last period's error term as follows:

$$\epsilon_t = \rho\epsilon_{t-1} + v_t$$

where ρ (rho) = coefficient of autocorrelation ($-1 < \rho < 1$) and v is an error term that satisfies Assumption III.

The degree and type of autocorrelation is indicated by the magnitude and sign of rho:

$\rho = 0$ indicates that $\epsilon_t = v_t$ and the absence of autocorrelation.
$\rho > 0$ indicates positive autocorrelation and is consistent with pure autocorrelation.
$\rho < 0$ indicates negative autocorrelation and suggests a specification error, since an error term switching signs in subsequent observations is inconsistent with pure autocorrelation.

Given this specification of the autocorrelation process, GLS estimation is straight-forward. The logic behind the procedure can be revealed by manipulating the following simple regression model with first-order autocorrelation:

$$Y_t = \beta X_t + \epsilon_t \quad with\ \epsilon_t = \rho\epsilon_{t-1} + v_t \tag{41}$$

Recall, we can always multiply both sides of an equation by a constant without changing the equality. Multiplying Equation 41 by $-\rho$ and rewriting it for period $t - 1$ yields:

$$-\rho y_{t-1} = -\beta\rho x_{t-1} - \rho\epsilon_{t-1} \tag{42}$$

Adding Equations 41 and 42 yields:

$$y_t - \rho y_{t-1} = \beta x_t - \beta\rho x_{t-1} + \epsilon_t - \rho\epsilon_{t-1}$$

Substituting $\rho\epsilon_{t-1} + v_t$ for ϵ_t (from Equation 41) yields:

$$y_t - \rho y_{t-1} = \beta[x_t - \rho x_{t-1}] + v_t \tag{44}$$

If we *know the value of rho*, then we can create two new variables: $y_t^* = y_t - \rho y_{t-1}$ and $x_t^* = x_t - \rho x_{t-1}$. Substituting these new variables into Equation (43) yields:

$$y_t^* = \beta x_t^* + v_t \tag{45}$$

Since v_t satisfies Assumption III and β is the same coefficient as in our original model (Equation 41), OLS estimation of Equation 45 yields unbiased, minimum-variance estimates of β.

If, as is far more likely, we *do not know rho*, then we must estimate it. This is GLS estimation. In practice, if we request a first-order autocorrelation correction from a regression software package, the software will estimate ρ, and then use this estimate to create y_t^* and x_t^* and estimate β from the equivalent of Equation 45. For large samples, the GLS estimates of β (based on an *estimated* rho) are unbiased, and have lower variance than the original OLS estimates.

3. *Testing for Autocorrelation*

Autocorrelation is so common in time-series models, that one should always test for it. The most widely used test is the Durbin-Watson test. It relies on the following DW statistic, which is computed by most regression packages:

$$DW \approx 2(1 - r_e) \tag{46}$$

where r_e is the correlation coefficient between e_t and e_{t-1}.

The value of DW indicates the presence and probable source of any sample autocorrelation as follows:

- If $r_e = 0$, e_t is not correlated with e_{t-1} (presumably because ϵ_t is uncorrelated with ϵ_{t-1}), then DW = 2. Therefore, a DW value of approximately 2 indicates the absence of autocorrelation.
- If $r_e > 0$, which is most likely with positive autocorrelation, then DW is less than 2. Therefore, a DW value between 0 (its minimum value) and 2 *may* indicate positive autocorrelation.
- If $r_e < 0$, which is most likely with negative autocorrelation, then DW is greater than 2. Therefore, a DW value between 2 and 4 (its maximum) *may* indicate negative autocorrelation

Since negative autocorrelation usually indicates a specification error, we test for *positive* autocorrelation. The decision rule involves two critical statistics, d_L and d_U, which are found in collections of statistical tables and all econometrics textbooks. These numbers vary with the number of independent variables, the number of observations, and the level of significance. The decision rule is:

If $DW < d_L$, positive autocorrelation is indicated ($\rho > 0$)
If $DW > d_U$, positive autocorrelation is NOT indicated ($\rho \leq 0$)
If $d_L \leq DW \leq d_U$, the test is inconclusive

Consider an example. Assume we have a time-series model with 3 independent variables and 25 observations. A Durbin-Watson table reveals that for these values, $d_L = 1.12$ and $d_U = 1.66$. If the computed DW (from a regression package) is:

- less than 1.12, assume that positive autocorrelation exists
- greater than 1.66, assume there is *no* positive autocorrelation
- between 1.12 and 1.66, the test is inconclusive.

The weakness of the Durbin-Watson test is this inconclusive range. There is disagreement about how to treat this result. Some investigators choose to treat this as support for the absence of autocorrelation. But, given we do not know how to interpret this information, it seems more appropriate to assume that we may have positive autocorrelation if we are in the inconclusive range.

When the Durbin-Watson test indicates positive autocorrelation, one should apply the GLS correction discussed earlier. Finally, note that a simple correction for some autocorrelation is, if possible, to increase the time unit of the data. Obviously, it is more likely for random shocks to extend to future periods with daily, weekly, or monthly data, than with annual data.

C. Heteroskedasticity

Heteroskedasticity occurs when the variance of the error term is not constant over all observations ($E(\epsilon_i^2) \neq \sigma^2$ over all i), and is a violation of Assumption IIIb. The concept is best understood with an example. Consider a model of family food expenditures as a function of family income. Homoskedasticity implies that the *variation* in food expenditures is the same at different income levels; but, we expect less variation in consumption for low income families than for high income families. At *low* income levels, average consumption is low with little variation. Food expenditures can't be much below average, or the family may starve; expenditures can't be much above average due to income restrictions. High income families, on the other hand, have fewer restrictions, and thus can have more variation. Heteroskedasticity is indicated a priori.

Heteroskedasticity is much more likely to appear with *cross-sectional* data than with time-series data because the range in values of the variables is usually much larger in cross-sectional data—e.g., the range in food expenditures or city size.[12] Thus, heteroskedasticity is discussed in the context of cross-sectional data.

With OLS estimation, the consequences of heteroskedasticity are the same as with autocorrelation—coefficient estimates are unbiased, but their variance is inflated, and t-tests are invalid. The correction is also similar. But first, consider how we detect heteroskedasticity.

1. Testing for Heteroskedasticity

The most common test for heteroskedasticity is a visual inspection of the residuals. The residuals $(Y_i - \hat{Y}_i)$ are plotted on a graph against the independent variable that is suspected of causing heteroskedasticity. If the absolute magnitude of the residuals appear on average to be the same regardless of the value of the independent variable, then there probably is *no* heteroskedasticity. If their magnitude seems *related* to the value of the independent variable, then a more formal test is indicated. For example, a residual plot of our food expenditure example should indicate more widely scattered residuals at high income levels than at low income levels. Most regression software packages will provide residual plots upon request against any variable selected by the user.

There is obviously a fair amount of subjectivity in interpreting residual patterns; unfortunately subjectivity is a characteristic of the more formal tests as well. Unlike the widely accepted use of the Durbin-Watson test to detect autocorrelation, there are several competing tests for heteroskedasticity (e.g., the Goldfeld-Quandt, Breusch and Pagan, and White tests). Here, we briefly discuss the White test.

White (1980) proposes that we regress the squared residuals against all the explanatory variables suspected of causing heteroskedasticity *and* their squares and cross products. For example, if two variables X_1 and X_2 are suspected, then one would run the following regression:

$$(Y - \hat{Y})^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1^2 + \alpha_4 X_2^2 + \alpha_5 X_1 X_2 + \mu$$

The R^2 associated with this regression provides a test for heteroskedasticity. If the error term is homoskedastic, then nR^2 is distributed as a chi-square with $k(k + 1)/2$ degrees of freedom. If this hypothesis is rejected, either heteroskedasticity or a misspecification is indicated.

2. Correction

As with autocorrelation, a GLS correction is possible for heteroskedasticity if one can specify the error variance. For example, consider the following simple regression model with a heteroskedastic error term:

$$Y_i = \beta X_i + \epsilon_i \quad \textit{with Var } (\epsilon_i) = Z_i^2 \sigma^2 \tag{48}$$

where Z_i denotes an exogenous variable that varies across observations (if Z_i is a constant then ϵ is homoskedastic).

The GLS approach in this case is to transform the original equation into one that meets the homoskedastic assumption. For example, dividing both sides of Equation 48 by Z_i yields:

$$Y_i/Z_i = \beta X_i/Z_i + \epsilon_i/Z_i \tag{49}$$

The variance of the new error term (ϵ_i/Z_i) can be shown to equal σ^2. Thus, the transformed equation has a homoskedastic error.

We need only create 3 new variables: $Y^* = Y/Z$, $X^* = X/Z$, and $\epsilon^* = \epsilon/Z$, and substitute them into Equation 49, which yields:

$$Y^* = \beta X^* + \epsilon^* \tag{50}$$

Estimating Equation (50) using OLS produces unbiased and minimum-variance estimates of β.

The problem, of course, is the specification of Z. It is unusual to know the specification of the error variance. The most common approach is to specify Z as a function of the independent variable suspected of causing the heteroskedasticity problem (e.g., 1/X). If, however, the specification of Z is arbitrary, then it is uncertain whether GLS estimates are better or worse than the OLS estimates of the original equation.

A more direct solution to heteroskedasticity is to redefine the variable in question—if it makes sense theoretically. For example, assume we want to explain city expenditures for services (E) as a function of the income of its citizens (l), i.e.,:

$$E = \beta_0 + \beta_1 I + \epsilon$$

Obviously, *large* cities have more income, more expenditures, and presumably more variation in those expenditures, than small cities, which suggests heteroskedasticity. But, the model may in fact be misspecified. Is the relationship of interest between *levels* of expenditure and income or between *per capita* expenditures and income? Reformulating the model in per capita terms ($PE = \beta_0 + \beta_1 PI$) removes size and its spurious effect and thus eliminates the heteroskedasticity.

Heteroskedasticity may thus indicate a specification problem, either from an incorrect formulation or from an omitted variable. In these cases the solution is straight-forward—reformulation or inclusion of the omitted variable. In general, one is *less* likely to apply GLS to correct heteroskedasticity than to correct autocorrelation—because, with heteroskedasticity, we have less confidence in the validity of the GLS specification of the error term.

VIII. MULTICOLLINEARITY

Assumption IV states that there are at least as many observations as independent variables, and that there are no exact linear relationships between the independent variables. This assumption (unlike the others) can easily be checked for any specific model. In fact, if either part of this assumption is violated, it is impossible to compute OLS estimates.

The first part of Assumption IV refers to the need to have more than *k* pieces of information (observations) to estimate *k* parameters. It requires only that the sample size be larger than the number of parameters to be estimated—the number of independent variables plus one. In fact, researchers usually seek the largest available and cost effective sample, since increasing the degrees of freedom (the number of observations minus the number of parameters to be estimated) reduces the variance in the estimates.

The second part of Assumption IV, the absence of an exact linear relationship among the independent variables, is unlikely to be an issue in naturally occurring data. It can, however, occur in data that have been constructed by the researcher. The problems created are easily illustrated. Assume that we want to estimate the following model:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon \tag{51}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

But,

$$X_2 = 5 + 2X_1 \tag{52}$$

Substituting Equation 52 into Equation 51 and then rearranging yields:

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 X_1 + \alpha_2(5 + 2x_1) + \epsilon \\ Y &= \alpha_0 + 5\alpha_2 + X_1(\alpha_1 + 2\alpha_2) + \epsilon \end{aligned} \tag{53}$$

If we could run OLS on this last equation, the parameter estimates would be:

$$a_0 = \alpha_0 + 5\alpha_2$$

$$a_1 = \alpha_1 + 2\alpha_2$$

But, we have 2 equations and 3 unknowns (α_0, α_1, α_2), and therefore cannot recover the model parameters.

Since the computer software will reject the regression run, it is easy to detect perfect collinearity. The error is usually one of variable construction, and can be avoided by exercising care.

It is quite possible, however, to have an *approximate* linear relationship among independent variables. The phenomenon of two or more variables that are highly, but not perfectly, correlated is called *multicollinearity*. High multicollinearity implies that, with a given change in one variable, the observations on the variable with which it is highly correlated are likely to change predictably. Unfortunately, such relationships are relatively common among social science variables.

A. Consequences

Multicollinearity causes problems similar to autocorrelation and heteroskedasticity. OLS parameter estimates are unbiased, but their variance is affected, yielding two undesirable consequences:

Large Parameter Variances. The variances of the parameter estimates of collinear variables are very large. Intuitively, one doesn't have enough unique information on a collinear variable to produce precise estimates of its effect on the dependent variable. To see this, consider the variance of $\hat{\beta}_1$ in the two-regressor model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$:

$$VAR(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (X_{i1} - \overline{X}_1)^2 (1 - r_{12}^2)} \tag{55}$$

r_{12}, the correlation coefficient between X_1 and X_2, captures the collinearity between X_1 and X_2. As r increases, the variance of the estimates of both β_1 and β_2 increase.

Biased t-statistics. The denominator of a t-statistic, the *estimated* standard deviation of the parameter estimate ($s_{\hat{\beta}}^2$), reflects the same dependence on the collinearity between the independent variables as the true variances.

$$s_{\hat{\beta}_1}^2 = \frac{s^2}{\sum_{i=1}^{n} (X_{i1} - \overline{X}_1)^2 (1 - r_{12}^2)} \tag{56}$$

where s^2 is an unbiased estimator of the unobserved variance of the error term, σ_ϵ^2.[13]

As r increases, $s_{\hat{\beta}}^2$ increases, which decreases the associated t-statistic. Multicollinearity thus causes t-statistics to be biased *downward*, making it more difficult to achieve statistical significance.

B. Diagnosis

Although multicollinearity is often suspected *a priori*, there are estimation results that indicate a problem. Two phenomena are particularly indicative of high multicollinearity. First, what if the estimation reveals a high R^2, but *most* of the parameter estimates are statistically insignificant? These are inconsistent results, the high R^2 indicates that the model has good explanatory power; the lack of significant variables indicates that most of the independent variables have no effect on the dependent variable. What then is explaining the variance in the dependent variable? One explanation is that something is deflating the t-statistics, e.g., high multicollinearity.

Second, what if the parameter estimates in the model change greatly in value when an independent variable is added or dropped from the equation? For example, we are uncertain about whether to include a particular independent variable and estimate the model twice, omitting the variable in question the second time. We find that the parameter estimates associated with one or more of the other independent variables change significantly. This indicates a high correlation with the omitted variable.

These symptoms indicate a potential multicollinearity problem; diagnosis requires examining the intercorrelation among the independent variables using auxiliary regressions. *Auxiliary Regressions* are *descriptive* regressions of each independent variable as a function of all the other independent variables. For example, for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, the associated auxiliary regressions are:

$$X_1 = a_0 + a_1 X_2 + a_2 X_3 + u_1$$
$$X_2 = b_0 + b_1 X_1 + b_2 X_3 + u_2$$
$$X_3 = c_0 + c_1 X_1 + c_2 X_2 + u_3$$

If multicollinearity is suspected, one should run the auxiliary regressions and examine the R^2 associated with each. If any are close to 1 in value, indicating that most of the variance in one of the independent variables is explained by the other independent variables, there is high multicollinearity.

C. Solutions

The appropriate solution to multicollinearity depends on whether the espoused model is in fact valid. Sometimes the presence of multicollinearity alerts us to a specification error. For example, two measures of the same phenomenon could have been included in the model. In that case, only the measure of the independent variable that most closely captures the theoretical concept should be retained.

If, however, the model is correct, then multicollinearity indicates a sample problem—two or more independent variables that have separate theoretical influences cannot be distinguished due to covariation in the sample. The most straight-forward solution is to *increase the sample size*. A richer data set may resolve the problem.

Unfortunately, one cannot always obtain more data. This creates a more serious situation. One solution, if it makes sense, is to *combine* the collinear independent variables. For example, consider a model that explains voting behavior as a function of socio-economic variables, like

income and race, and indicators of media exposure, like the number of hours spent watching television and the number of hours spent reading newspapers. If the two media variables are highly collinear in the sample, they could be combined into a single measure or index of media exposure. Note that the resulting parameter estimate will not allow us to differentiate the effects of different media sources, i.e., the role of television compared to newspapers. But, we will have information on the role of the media compared to the other independent variables.

If the highly collinear variables *cannot* be combined, OLS cannot separate the effects of the individual collinear variables. The model can still be used for *prediction* (i.e., one can predict Y given all the Xs), but the separate effects of the collinear Xs on Y cannot be identified.

Finally, note that it is *not* a good idea to just discard the offending variable. If the original model is in fact correct, one is trading the consequences of multicollinearity for the more serious consequences of excluding a relevant variable. Consider, for example, the following two variable model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Unfortunately, X_1 and X_2 are highly correlated, so we drop X_2 and estimate: $Y = \beta_0 + \beta_1 X_1 + \mu$. If X_2 in fact affects Y, the parameter estimate of β_1 will be *biased*. The estimate will incorporate the effect of X_2, as it does the effect of any omitted relevant variable with which it is correlated.

IX. CONCLUSION

Regression analysis is an extremely powerful methodology. It offers public sector scholars an easy to use and readily understandable way to summarize multivariate relationships. Its ease of use, however, makes it vulnerable to misuse and its results to misinterpretation. The legitimate use of regression analysis requires an understanding of the assumptions that underlie its use as a tool of inferential statistics. Understanding the implications of these assumptions will enable the user to either appropriately qualify his or her conclusions or to select a more appropriate estimation strategy.

To conclude our discussion of regression analysis, let's recap the general issues involved with determining an appropriate estimation strategy. It is useful to view estimation strategy in a decision analytic framework. The first decision level considers the nature of the dependent variable. The second level considers the classical assumptions.

I. *Is the dependent variable an interval-level variable*?
   - If Yes, then Regression Analysis is appropriate.
   - If No, a Qualitative Dependent Variable Model is indicated

Within the Regression Model context, the Gauss-Markov theorem tells us that if the classical assumptions are satisfied, then Ordinary Least Squares (OLS) estimation yields the best linear unbiased estimates of the underlying parameters, β. Thus we need only consider an alternative technique if one of these assumptions is violated.

lla. Is the model correctly specified?
   - If Yes, consider the next assumption.
   - If No, there is a specification error. Excluding relevant variables usually yields biased parameter estimates and invalid t-tests. Including irrelevant variables yields inefficient parameter estimates and lowered t-statistics. The solution is to correctly specify the model.

IIb. Are the observations of the independent variables determined outside the model and thus independently of the error term?

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

- If Yes, then OLS estimation is appropriate.
- If No, OLS parameter estimates are biased. The preferred estimation strategy is Instrumental-Variable Estimation.

IIc. Do the random components of the model (the error terms) have a constant variance and are they uncorrelated across observations?
- If Yes, then OLS estimation is appropriate.
- If No, OLS parameter estimates are not efficient, and t-tests are invalid. Generalized Least Squares (GLS) estimation or a re-specification of the model should be considered.

IId. Do the number of observations exceed the number of parameters to be estimated, and are there no exact linear relationships among the independent variables?
- If Yes, regression analysis is appropriate.
- If No, the regression equation cannot be estimated. If high, rather than perfect, multicollinearity exists, OLS parameter estimates are inefficient and t-tests are biased; additional data or a new specification is needed.

NOTES

1. Throughout this chapter, observations are denoted with the subscript *i* and range from 1 to *n*, while variables are denoted with the subscript *j* and range from 1 to *k*. The observation subscript, however, will usually be suppressed for ease of exposition.
2. Data source: Nelson A. Rockefeller Institute of Government (1983). *1983–84 New York State Statistical Yearbook*, 10th edition, Albany, NY: Nelson A. Rockefeller Institute of Government, State University of New York.
3. These observations can be either cross-sectional data (data collected on different units—e.g., firms, families, cities—at a single point in time) or time series data (data collected on the same unit over time).
4. Greene, William H. *LIMDEP*, Version 6.1. Econometric Software, Inc. New York, N.Y.
5. Assumption IIID is not necessary for the Gauss-Markov result, but it is important for the use of significance tests.
6. This is a violation of Assumption IV and will be discussed in Section VIII.
7. In general, if a noninterval variable has *g* mutually exclusive and exhaustive categories, then *g*-1 dummy variables can be used to represent it in the regression model.
8. For an accessible derivation, see Johnson, Johnson and Buse (1987), ch. 15.
9. The estimated error variance, s^2, equals $\sum e_i^2/(n\text{-}k\text{-}1)$.
10. *Instrumental-variable estimation* is a general estimation procedure applicable to situations in which the independent variable is *not* independent of the error term. The procedure involves finding an ''instrument'' to replace X that is both uncorrelated with the error term and highly correlated with X. A discussion of this procedure is beyond the scope of this chapter, but can be found in many econometrics textbooks (e.g., Maddala, 1992, ch. 11).
11. The interested reader is referred to Pindyck and Rubinfeld (1991), who do a good job of developing both.
12. With time-series data, changes in the variables over time are likely to be similar orders of magnitudes.
13. The positive square root of s^2 is called the *standard error of the equation*.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

REFERENCES

Cassidy, H.J. (1981). *Using Econometrics: A Beginner's Guide*. Reston, VA: Reston Publishing.

Devaney, B., L. Bilheimer, and J. Schore (1992). ''Medicaid Costs and Birth Outcomes: The Effects of Prenatal WIC Participation and the Use of Prenatal Care,'' *Journal of Policy Analysis and Management*, *11*: 4, 573–592.

Ferris, J. (1988). ''The Public Spending and Employment Effects of Local Service Contracting,'' *National Tax Journal*, *41*: 2(June), 209–217.

Ferris, J. and E. Graddy (1988). ''Production Choices for Local Government Services,'' *Journal of Urban Affairs*, *10*: 3, 273–289.

Graddy, E. (1994).''Tort Reform and Manufacturer Payout—An Early Look at the California Experience,'' *Law & Policy*, *16*: 1 (January), 49–61.

Graddy, E. and M. Nichol (1990). ''Structural Reforms and Licensing Board Performance,'' *American Politics Quarterly*, *18*: 3 (July), 376–400.

Greenwald, H.P., M.L. Peterson, L.P. Garrison, L.G. Hart, I.S. Moscovice, T.L. Hall, and E.B. Perrin (1984). ''Interspecialty Variation in Office-Based Care,'' *Medical Care*, *22*: 1, 14–29.

Johnson, A.C., Jr., M.B. Johnson, and R.C. Buse (1987). *Econometrics: Basic and Applied, New York*: MacMillan Publishing.

Maddala, G.S. (1992). *Introduction to Econometrics*, 2nd edition, New York: MacMillan Publishing.

Mann, J., G. Melnick, A. Bamezai, and J. Zwanziger (1995). ''Managing the Safety Net: Hospital Provision of Uncompensated Care in Response to Managed Care,'' In *Advances in Health Economics and Health Research*, Volume 15, JAI Press, 49–77.

May, P.J. and R.J. Burby (1996). ''Coercive Versus Cooperative Policies: Comparing Intergovernmental Mandate Performance,'' *Journal of Policy Analysis and Management*, *15*: 2, 171–201.

Pindyck, R.S. and D.L. Rubinfeld (1991). *Econometric Models and Economic Forecasts*, 3rd edition, New York: McGraw-Hill.

Robertson, P.J. (1995). ''Involvement in Boundary-Spanning Activity: Mitigating the Relationship between Work Setting and Behavior,'' *Journal of Public Administration Research and Theory*, *5*: 1, 73–98.

Sundeen, R.A. (1988). ''Explaining Participation in Coproduction: A Study of Volunteers,'' *Social Science Quarterly*, *69*, 547–568.

Sundeen, R.A. (1990). ''Family Life Course Status and Volunteer Behavior,'' *Sociological Perspectives*, *33*: 4, 483–500.

White (1980). ''A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,'' *Econometrica*, *48*, 817–838.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

18

Multivariate Techniques for Dichotomous Dependent Variables

Mack C. Shelley II
Iowa State University, Ames, Iowa

I. INTRODUCTION: WHY DICHOTOMIZE?

A. Scenarios and Solutions

Research applications in public administration, like many other aspects of social science research more broadly, has been influenced heavily in recent years by the spread of methods that are adapted to realistic situations in which the outcomes that are of interest fall into two discrete possible categories. Here are three scenarios explaining examples of situations in which there are dichotomous outcomes, for which we would be interested in knowing what other variables help us to predict with accuracy into which of the two categories an observation actually falls.

1. Scenario 1

In a study of forms of urban administration, we might want to know whether there is any systematic pattern that separates strong mayor forms of government from other forms (weak mayor, commission, or council-manager could all be classified as ''other than strong mayor''). We could do this by looking at patterns of variation for both types of municipal administration in variables that identify city traits (size, rate of population growth, ethnic mix, median income or the equity of income distribution within the city, annexation activity, and other predictor variables). Really, predictor variables here should be thought of as ''classification variables,'' because the real task is to maximize the likelihood that we guess correctly into which of the two types of administration any particular city falls.

2. Scenario 2

In a study of inter-city migration, we might decide that the most interesting thing to know is what things need to be known about cities in general that would distinguish between those that have experienced net gains in population and those that have lost population. As predictor (or ''classifier'') variables, we could use weather or climate measures such as mean annual temperature (to help operationalize differences between ''Frostbelt'' and ''Sunbelt'' cities, for example), rates of taxation, crime rates, or levels of unionization among the workforce.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

3. Scenario 3

Unlike the first two examples, in which we have had to ''help'' the analysis along a bit by forcing a dichotomy to exist for variables that otherwise could be measured as a polytomy (with perhaps four different categories) in the first example, or finding a convenient and fairly natural ''cut point'' for what otherwise is a continuous dependent variable in the second example, in another context we may have a natural dichotomy that needs no assistance. Such naturally-occurring dichotomous situations would arise, for instance, if we were predicting the outcome of a referendum to deprive gay men and lesbian women of their constitutional rights; the referendum either succeeds or fails, and the margin of passage or defeat would be of only minimal interest. Another case of a natural dichotomy would arise if we were using topographic and construction characteristics to predict whether private wells are contaminated by coliform bacteria (or atrazine, nitrates, *E. coli*, or some other biohazard).

B. The Trouble With Tuples

Under any and all of these sets of circumstances, we would be faced with a very different research problem than what can be handled with least squares tools such as multiple regression, multiple and partial correlation, analysis of variance, analysis of covariance, and related methods. For least-squares-based methods to ''work,'' both in terms of the practical interpretations that can be drawn from their results and in terms of the statistical properties that are required under what frequently is referred to as ''normal theory,'' at three things must be true:

a. the dependent variable must be distributed following a bell-shaped curve with specific proportions of observations occurring within specified standard deviation intervals below and above the mean of the variable;
b. the dependent variable has a constant variance over all values of the independent variables; and
c. that the observations from which the data were gathered are mutually independent.

In addition, usually we assume that the values of the independent variables (the X's, in the standard notation) are fixed in repeated samples, and that what we happen to observe in our one and only one sample is a random realization of all possible such samples, for different times or for different sets of observations. Furthermore, least squares analysis is usually accompanied by the assumption that all the data were collected from a simple random sample, which means that all samples were equally likely to be chosen and that each individual observation (a city, a state, or a voter, for example) has a known and possibly equal chance of having been selected. For another, we prefer it to be the case that the independent variables are not functions or near-functions of each other (that is, that the condition of multicollinearity does not exist, or at least that it is not serious).

Pretty much by the very definition of a dependent variable as dichotomous, there is no realistic way to believe that a normal distribution could exist for the dependent variable in any of our three scenarios. For much the same reason, constant error variance is unreasonable; in fact, we are virtually guaranteed to introduce a severe condition of heteroscedasticity (or, non-constant error variance) by virtue of the fact that only two possible outcome values can be ''predicted.'' So, assumptions (a) and (b) pretty much automatically go by the board when we consider how to deal with dichotomous dependent variables that are of interest for a research project. However, assumption (c), of independence, still may be true, although whether that

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

assumption is correct also would need to be investigated by thinking through the process that was used to generate the data set.

II. THE MAXIMUM LIKELIHOOD CURE, AND A GOOD WORD FOR LEAST SQUARES

Clearly, we will need a different set of mechanics than what works for the more comfortable least squares situation, together with a quite different form of interpretative logic that will allow us to reason through the implications of what our models can tell us. Generally, these methods are known as *maximum likelihood techniques*, because their goal is to maximize the chances that we can correctly (that is successfully, or accurately) classify an observation in one or the other outcome category based on the background variables that we happen to know about each data value. These maximum likelihood principles can be extended rather easily to address ordered or nonordered polytomies, with at least two outcome categories. For instance, we might be interested in developing a model that would tell us whether we can predict successfully the level of severity of the injuries sustained (none, minor, major, or fatal) in head-on collisions involving passenger cars, based on knowledge of where in the car each person was sitting, as well as the person's gender, and age, and what kind of safety device, if any, the person was using when the impact occurred.

Although we have just dismissed least squares-based methods, such as ordinary least squares (OLS, for short) multiple regression and its immediate offshoots like analysis of variance and analysis of covariance, for analyzing dichotomous-outcome situations, there is at least one least-squares-related technique that has been used with some success in predicting two-valued categorical outcomes: This method is known as discriminant analysis, which, like least squares, depends critically on the three assumptions listed above—particularly, the assumption of multivariate normality (that is, a jointly normal distribution of the predictor variables in the model), which is unlikely to be a correct, or at least reasonable, assumption in most practical applications. We begin our discussion below with discriminant analysis, largely because it retains much of the ''look and feel'' of least-squares methods and thereby provides a rather familiar and comfortable bridge, leading to a much more fully developed discussion of maximum likelihood-based techniques.

We will focus chiefly on logistic regression as the usual weapon of choice for doing combat with dichotomous outcome dependent variables, although logistic regression is in fact only part of an even broader set of methods that could be labeled as generalized linear (and nonlinear) models. The entire discussion is built around the unifying theme that, whether the specific tools of analysis are least-squares-based or predicated on the more flexible and more easily generalizable principles of maximum likelihood, the operative logic always is that of attempting to classify an observation correctly according to its known background characteristics. There are other alternative approaches to dealing with dichotomous dependent variables apart from logistic regression models. One such alternative employs a probit transformation, which involves re-expressing the probabilities of either of the two possible outcomes occurring as values of the cumulative normal distribution. However, probit models are less desirable, and less useful, than logistic models, because they are difficult to generalize beyond a single predictor variable and because statistical inference becomes more difficult. Generally, probit and logit models produce comparable results except in the neighborhood of the extremes (zero and one). Another related method that produces generally comparable results is the complementary log-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

log transformation, $\log_e(-\log_e(1-\pi))$, which produces a range of possible values from negative infinity to positive infinity, rather than from 0 to 1.

A. A Starting Point: Crosstabular Analysis

The simplest, but still effective, method for analyzing dichotomous dependent variables is to formulate the analysis as a two-way crosstabulation, in which the ''predictor'' variable also is a dichotomy or polytomy. In this configuration, the information of interest is whether the proportion of observations on the dependent dichotomy varies across the categories of the ''predictor'' variable.

As an example of what is entailed in this mode of analysis, consider the flood waters that inundated much of the Midwest during the summer of 1993. For a time, aerial infrared maps of the earth's surface revealed what came to be referred to as the ''sixth Great Lake'' running from the upper Midwest of Minnesota and Iowa down the Mississippi River and its tributaries to about St. Louis, Missouri.

In the aftermath of this massive flooding, public health agencies feared that groundwater may have been contaminated, endangering the drinking water supplies of particularly those living in rural areas. A regional study was undertaken by the Centers for Disease Control and Prevention (CDC), in Atlanta, Georgia, to attempt to find out how serious this problem might have become. In Iowa, that state's Department of Public Health (IDPH) supplemental data also were gathered to provide a stronger basis for comparison of three different types of rural well construction, to see which type of well design—buried slab, other large-diameter, and small-diameter—minimized the risk of contaminants seeping into the well water supply. Laboratory tests were conducted to attempt to detect measurable levels of coliform bacteria, *E. coli* bacteria, nitrates, and atrazine.

Table 1 summarizes the results from a two-way crosstabulation of the three well construction types (WELLGRP) by a dichotomy (ZNIT) that distinguishes between wells that have no more than the detectable threshold of nitrate contamination ($\leq$1 part per million), coded ''0,'' and wells that attain or exceed the detectable threshold ($>$1 part per million), coded ''1.'' The analysis of these data was conducted by the Statistical Analysis System (SAS) statistical software package, using its FREQ procedure.

From the marginal column frequencies of Table 1, it can be stated that, of the total of 1061 rural Iowa wells for which information is available on both variables, the test results for a bit less than 44.5% (472 of 1061) indicated no serious risk of contamination with nitrates, and the remaining 55.5% (589 of 1061) had unsafe levels of nitrate contamination.

The row marginals also show that, of the 1061 total wells, about 24.8% (263 of 1061) were constructed according to the buried slab design, while about 29.4% (312 of 1061) were other large-diameter designs and the remaining 45.8% (486 of 1061) were of small-diameter construction. All else equal, because they provide less of an opening for contaminants to get directly into their water contents, small-diameter wells would be expected to do better than most large-diameter wells.

Tables 1 and 2 below were generated by the FREQ procedure in SAS, using the following commands:

```
proc freq data=bbb.nnset1;
 tables wellgrp*znit/chisq;
run;
```

The question of particular interest here is whether there is an appreciable difference between the more sophisticated buried slab construction type and other forms of large-diameter

TABLE 1 Initial Crosstabulation Analysis of Well Contamination Data

Table of WELLGRP by ZNIT

WELLGRP	ZNIT		
Frequency percent			
Row pct.	Safe	Unsafe	
Col. pct.	0	1	Total
Buried slab	89	174	263
Wells	8.39	16.40	24.79
	33.84	66.16	
	18.86	29.54	
Other large	37	275	312
Diameter	3.49	25.92	29.41
Wells	11.86	88.14	
	7.84	46.69	
Small	346	140	486
Diameter	32.61	13.20	45.81
Wells	71.19	28.81	
	73.31	23.77	
Total	472	589	1061
	44.49	55.51	100.00
Frequency missing = 88			

wells. The nature of the conclusions that could be drawn from these data are important for the study of public policy and public administration, because they can inform policymakers and others about how best to go about protecting the water that is needed for human and animal consumption.

Whether, in fact, there is a statistically meaningful difference between the proportions of each of the three well types that have serious levels of nitrate contamination, may be determined by the use of different versions of the chi-square test statistic and other measures based on chi-square that serve as correlation coefficients. Here, we only scratch the surface of what turns out

TABLE 2 χ^2-Based Summary Statistics for the Initial Crosstabulation Analysis of Well Contamination Data

Statistics for table of WELLGRP by ZNIT			
Statistic	DF	Value	Prob
Chi-Square	2	286.927	0.001
Likelihood Ratio Chi-Square	2	310.504	0.001
Mantel-Haenszel Chi-Square	1	143.462	0.001
Phi Coefficient		0.520	
Contingency Coefficient		0.461	
Cramer's V		0.520	
Effective Sample Size = 1061			
Frequency Missing = 88			

to be a very large and diverse family of nonparametric measures of association. Readers interested in further information on other measures for relationships among discrete data are referred to Garson (1971) and to Bishop et al. (1975). Our interest here is focused on the family of statistical measures based on the chi-square statistic.

The first thing that should be noticed from the results in Table 2 is that four separate, but related, sets of information are provided. First, there are several different labels presented under the heading *Statistic*: Chi-Square, Likelihood Ratio Chi-Square, Mantel-Haenszel Chi-Square, Phi Coefficient, Contingency Coefficient, and Cramer's V.

Of these six statistics, the first three are variations on a common theme. The statistic labeled simply ''Chi-Square'' is known more formally as the Pearson Chi-Square, named after Karl Pearson, who invented and popularized much of what we do today in parametric statistical applications. This statistic is appropriate for all variables and can detect any kind of association, but is less powerful for detecting a linear association because its power is dispersed over a greater number of degrees of freedom than the Mantel-Haenszel version of chi-square. The Mantel-Haenszel chi-square statistic requires an ordinal scale for both variables, and is designed to detect a linear association. In contrast, the Pearson chi-square is appropriate for all variables, and can detect any kind of association. However, the Pearson statistic has less statistical power than the Mantel-Haenszel statistic, which means that it is less well able to detect a linear association, because it requires a larger number of degrees of freedom unless the crosstabulation table has two rows and two columns (when the Pearson degrees of freedom equal one).

Here are the formulas for each of these statistics, and some explanation of what each one does. For a crosstabulation table with rows labeled by the values X_i, $i = 1, 2, \ldots, r$, and with columns labeled by the values Y_j, $j = 1, 2, \ldots, c$, the crosstabulation table has the number of observations in each column denoted by $n_{.j} = \sum_i n_{ij}$, the number of observations in each row denoted by $n_{i.} = \sum_j n_{ij}$, and the overall total number of observations denoted by $n = \sum_i \sum_j n_{ij}$. The Pearson chi-square statistic operates off the expected number of observations within each cell under the null hypothesis of no association between the row and column variables, denoted by $m_{ij} = n_{i.} n_{.j}/n$. The alternative hypothesis is that there is a pattern of general association relating the two variables to each other. With $(r - 1) \times (c - 1)$ degrees of freedom, the Pearson chi-square (χ^2_p) is defined by

$$\chi^2_p = \sum_i \sum_j (n_{ij} - m_{ij})^2/m_{ij}.$$

The likelihood-ratio chi-square statistic, which also is known as the G^2 statistic, is computed by ratios between the observed and expected frequencies, with the alternative hypothesis being general association between the two variables. With $(r - 1) \times (c - 1)$ degrees of freedom,

$$G^2 = 2 \sum_i \sum_j n_{ij} \ln(n_{ij}/m_{ij}).$$

The Mantel-Haenszel chi-square statistic (χ^2_{MH}) tests the alternative hypothesis of a linear association between the row and column variables, with 1 degree of freedom, and is defined as

$$\chi^2_{MH} = (n - 1)r^2,$$

where r is the value of the Pearson product-moment correlation between the two variables. This calculation is valid only when both variables are measured on ordinal scales. Therefore, for the example we examine here of nitrate contamination and well construction type, the Mantel-Haenszel result is not a valid measure of association, and it should not be used to interpret the relationship between these two variables.

Of these chi-square statistics, the most useful for our purposes is the Likelihood Ratio Chi-Square, because this is closely related to comparable statistics used in more advanced methods for analyzing dichotomous dependent variables. Here, the estimated Likelihood Ratio Chi-Square value of 310.504 would be compared against an appropriate critical value of chi-square with 2 degrees of freedom:

$$\chi^2_{2,.05} = 5.991; \chi^2_{2,.01} = 9.210$$

The computed value of Likelihood Ratio Chi-Square value of 310.504 far exceeds these critical values, and the attained level of significance (or, the area in the right tail of the chi-square distribution that is beyond 310.584 for a chi-square statistic with 2 degrees of freedom) is rounded to .001 (it actually is much smaller than this, but, like most ''canned'' statistical programs, SAS has an internal limitation that forces it to print a value of .001 for any chi-square statistic calculated by its FREQ procedure that is actually much farther off in the right tail of the distribution.

The remaining statistics computed for the crosstabulation table (shown in Table 2) generated by this analysis all are functions of the Pearson chi-square value. Unlike the Pearson version and the other two types of chi-square statistics shown here, however, the three other statistics reported by SAS operate as nonparametric (that is, distribution-free) measures of correlation. Except when the crosstabulation table has just two rows and two columns, each of these measures are always nonnegative (the Phi Coefficient can be negative for a crosstabulation table with two rows and two columns) and are structured such that larger values (closer to 1) imply a stronger correlation between the row and column variables, and smaller values (closer to zero) imply a weaker correlation between the two variables.

The Phi coefficient (ϕ) is calculated by

$$\phi = [\chi^2_p/n)]^{0.5} = [286.927/1061]^{0.5} = 0.520$$

which is just the square root of the Pearson Chi-Square value (286.927) divided by the total number of observations in the crosstabulation table (1061). The Phi Coefficient ranges in value between 0 and 1 (although the upper limit actually may be less than one, depending on the distribution of the marginal values).

The Contingency Coefficient also falls within the range between zero and 1, although, like the Phi Coefficient, the maximum possible value that it can attain may be less than 1, depending on the marginal distribution. The Contingency Coefficient (CC) is calculated by

$$CC = [\chi^2_p/(\chi^2_p + n)]^{0.5} = [286.927/(286.927 + 1061)]^{0.5} = 0.461$$

which is an adjusted version of ϕ that controls for the magnitude of the Pearson chi-square statistic, which increases with the number of cells in the crosstabulation table. This is a more conservative statistic than ϕ; it must be smaller than ϕ because the denominator is larger by the value of χ^2_p.

Finally, the upper bound of Cramer's V, also based on the Pearson chi-square, always equals one, unlike ϕ and the Contingency Coefficient. This statistic is defined as

$$V = [(\chi^2_p/n)/\min(r - 1)(c - 1)]^{0.5} = [(286.927/1061)/1]^{0.5} = 0.520$$

which is equal to the ϕ statistic here because one of the dimensions of the crosstabulation table equals two (there are two columns in the crosstabulation table).

III. THE MAXIMUM LIKELIHOOD APPROACH: LOGISTIC REGRESSION AND ITS VARIANTS

A. What's So Different About Logistic Regression?

Logistic regression is an application of a broader class of nonlinear regression models, defined by the implicit form

$$Y_i = f(X_i, \gamma) + \epsilon_i,$$

where each observation Y_i of the dependent, or response, variable Y is the sum of a mean response $f(X^i, \gamma)$, which is determined by the nonlinear response function f (X, γ) and the error term ϵ_i. There are one or more (q, in general) X variables, and one or more (p, in general) regression coefficients:

$$\underset{q\times 1}{X_i} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iq} \end{bmatrix} \quad \underset{p\times 1}{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix}$$

Although we will treat nonlinear regression as involving a completely different approach than the more familiar approach of linear regression, both nonlinear and linear models may be thought of as belonging to the family of generalized linear models (McCullagh and Nelder, 1989).

Logistic regression analysis usually proceeds under the assumptions that the errors are distributed identically and independent of each other, following a normal distribution with mean zero and constant variance. That is,

$$\epsilon \sim N(0, \sigma^2 I)$$

where $E(\epsilon) = 0$, $E([\epsilon - E(\epsilon)^2] = \sigma^2 I$, and $E(\epsilon_i, \epsilon_j) = 0$.

The logistic regression model, assuming its simplest version of a single predictor variable and normally distributed errors is

$$Y_i = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \epsilon_i$$

and the fitted, or response, function is given by

$$f(X, \gamma) = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X)}$$

Like linear regression models, parameter estimation for nonlinear regression models, such as logistic regression models, may be obtained by either least squares or maximum likelihood methods. Again just as with linear regression, least squares and maximum likelihood methods produce identical parameter estimates when the error terms for the nonlinear regression model are mutually independent, normally distributed, and have constant variance.

However, there are a number of important differences between estimation approaches in linear regression and in logistic regression. First, it is crucial to note that the parameters of the logistic response function—γ_0, γ_1, and γ_2—are not linear, which means that their interpretations will have to be undertaken with different goals in mind than what we are used to for ''regular'' regression results. Also, the number of predictor X variables in the nonlinear regression model

(q) is not necessarily the same as the number of regression parameters in the response function (p), unlike ordinary least squares regression. Another difference from linear regression is that with nonlinear regression methods such as logistic usually it is not possible to derive analytical expressions for the least squares or maximum likelihood estimators. Instead, numerical search procedures that frequently require considerably greater amounts of computational time must be used for either least squares or maximum likelihood estimation, using appropriate computer software. This is in contrast to linear regression, where hand calculations may be done rather easily to derive the relevant estimators.

B. What Does the Response Function Mean When the Outcome is Binary?

It would be possible to estimate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the dependent variable Y_i takes on the binary values of either 0 or 1. In this case, the expected response is

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i,$$

where π_i is the probability that $Y_i = 1$ when the level of the predictor variable is X_i, because $E(\epsilon_i) = 0$. When the dependent variable is a binary 0,1 indicator variable, the mean response always estimates the probability that $Y = 1$ for the given level(s) of the predictor variable(s).

C. Does This Create a Problem?

Well, yes, actually there are three particular problems that arise with binary dependent variable linear regression models. Each of these makes the use of linear regression inappropriate and signifies the need for an alternative approach. The discussion that follows emphasizes the central arguments surrounding the use of logistic regression, with technical details relegated to an appendix.

1. The first problem is that the error terms can't be distributed normally. This happens because each error term can assume only two possible values. Consequently, the assumption of normally distributed errors cannot be appropriate.
2. Second, the error terms do not have equal variances when the response variable is an indicator variable taking on values of zero and one. This happens because the variance of Y_i depends on the value of X_i; consequently, the error variances are different for different levels of X. As a result, ordinary least squares is no longer optimal.
3. Finally, the response function represent the set of probabilities when the outcome variable is equal to either zero or one. The mean responses from the response function thus are constrained within the limits of zero and one, because

$$0 < E(Y) = \pi < 1.$$

Consequently, linear response functions would not be appropriate because they well may produce predicted values that are either negative or greater than one; also, it would be unduly constraining in the linear case for all outcomes to have either zero or one probabilities when in fact most outcomes will fall between those extremes. What we need instead is for these extremes to be approached slowly—that is, asymptotically—as values of X become either very small or very large, and probabilities should decrease or increase toward those extremes nonlinearly.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

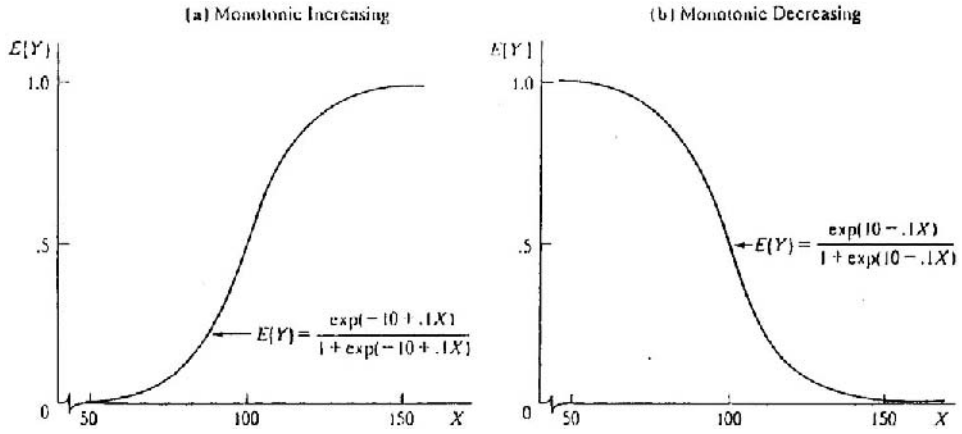


FIGURE 1 Examples of logistic response functions.

D. The Simple Logistic Response Function

Thus, for binary response variables the response function typically will be curvilinear, following an S-shaped pattern that fits the previously-discussed constraints that are imposed on E(Y) when the outcome variable is binary. As shown by the accompanying diagrams in Figure 1, this curvilinear relationship may follow either one of two general forms: (a) monotonically increasing at a varying rate from an asymptotic expected value for Y of zero for smaller values of X at the lower-left, toward an asymptotic expected value for Y of one for larger values of X at the upper-right, when $\beta_1 > 0$, or (b) monotonically decreasing at a varying rate from an asymptotic expected value of Y of one for smaller values of X at the upper-left, toward an asymptotic expected value for Y of zero for larger values of X at the lower-right, when $\beta_1 < 0$. In either version, it is useful to note that the middle part of the logistic curve is more or less linear, between values for E(Y) or about .2 and .8, and that only the ends vary dramatically from that pattern.

The fact that this logistic representation is not completely disconnected from linear regression logic is demonstrated by the ability to transform the logistic response function back into linear form. This is done rather easily by performing a logit transformation of E(Y) as the logarithm of the ratio of the probability of a success (defined here as π) and the probability of a failure $(1 - \pi)$:

$$\pi' = \log_e \left[\frac{\pi}{1 - \pi} \right]$$

which becomes

$$\pi' = \beta_0 + \beta_1 X.$$

The ratio of probabilities, $\pi/(1 - \pi)$, is known as the odds ratio; the transformed response function, $\pi' = \beta_0 + \beta_1 X$, is called the logit response function, or the logarithm of the odds ratio; and the value of π' is referred to as the logit mean response, which can vary from negative infinity to positive infinity as X varies over the same range.

E. The Simple Logistic Regression Model

When the response variable takes on values of only 1 (with probability π) and 0 (with probability $1 - \pi$), the simple logistic regression model takes the form

$$Y_i = E(Y_i) + \epsilon_i$$

where the error term ϵ_i follows the binomial (Bernoulli) distribution of Y_i with expected values $E(Y_i) = \pi_i$.

The likelihood function of the parameters to be estimated in the logistic regression model, given the sample observations, is expressed as $\log_e L(\beta_0, \beta_1)$, which is the logarithm of the likelihood function (or the log-likelihood function). The maximum likelihood estimates of β_0 and β_1 are the values of those parameters that maximize the log-likelihood function, which must be found by computer algorithms using search procedures that converge on the estimated values. After these values have been found, they are substituted into the response function to generate the fitted, or estimated, logistic response function

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

In simple logistic regression, the interpretation of b_1, the estimate of the ''slope'' parameter, β_1, differs from the usual interpretation of the corresponding parameter estimate in ordinary least squares models. This difference comes about because the measured effect of a one-unit increase in the predictor variable, X, is different in a simple logistic model depending on the location of the starting point on the scale of the X variable. In contrast, in ordinary least squares simple regression, the slope parameter represents a constant value at all points on the regression line. So, our interpretation of the effect on Y of a one-unit change in X will need to be expressed in terms of the value of b_1, which measures the proportional relative percentage change in Y in response to a change of one unit in X.

F. Multiple Logistic Regression

A direct extension of the simple logistic regression model to handle more than one independent variable results in the expression

$$E(Y) = \beta_0 + \beta_1 + \ldots + \beta_{p-1} X_{p-1}$$

The predictor variables may include the full range of options for regression models, including higher-order (quadratic or beyond) polynomials, interaction effects, continuous quantitative variables, or qualitative (indicator) variables. For the special case of a multiple logistic regression model that contains only qualitative variables, this specification often is referred to as a log-linear model.

Then, the multiple logistic response function is:

$$E(Y) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)} = [1 + \exp(-\beta' X)]^{-1},$$

where B is a vector containing the p model parameters and X is the matrix of data values for the predictor variables.

G. Stepwise Multiple Logistic Regression

Just as with linear regression models, it often is the case that not all predictor variables contribute equally well to explaining variation in the dependent variable. In such circumstances, it may be beneficial to attempt to find an ''optimal'' subset of predictors that contains only the most important independent variables. However, unlike linear regression approaches that may be used to generate all possible models (as can be done using PROC RSQUARE in SAS, for example),

the heavy computational demands required to estimate maximum likelihood solutions to logistic regression models generally require an alternative approach—stepwise logistic regression—that is very close to the stepwise methods used with linear regression.

Backward elimination logistic regression model building proceeds from a ''full-model'' containing all relevant predictor variables, and then eliminates one at a time at each successive step the predictor variable that is the least helpful for classifying the dichotomous outcomes correctly. Forward selection logistic regression begins by generating a simple logistic model using the single predictor variable that is the most helpful for correct classifications of the response variable, and then proceeds by adding more predictor variables that add progressively smaller increments to correct classifications. Both processes cease according to stopping rules, defined by relative changes in significance level, in proportion of correctly classified outcome values, or in other measures of model adequacy.

H. Logistic Regression: An Example With a Categoric Predictor Variable

Major statistical packages, such as SAS (the Statistical Analysis System) or SPSS (the Statistical Package for the Social Sciences), provide different ways to estimate logistic regression models. In SAS, this form of analysis can be conducted using either PROC CATMOD or PROC LOGISTIC. We will illustrate the CATMOD procedure first, in part because this approach is closer to the spirit of the crosstabulation results we have just examined. More important, however, is the fact that CATMOD is the most appropriate procedure when there is a qualitative independent variable. In general, the explanatory variables in logistic regression may be either categorical or continuous, and both types may be incorporated together into a single model. Logistic regression constitutes a research method for data analysis, model estimation, and statistical inference that has found broad applications beyond its initial extensive uses in medical and epidemiological studies to social science research in political science, public administration, sociology, economics, education, business, and other related areas of investigation. One of the primary advantages of logistic regression is that the estimated parameters of the model may be interpreted as the ratio of the odds of one outcome occurring relative to odds of the other outcome occurring. These odds ratios are functions directly of the parameters of the logistic model. This situation is in sharp distinction to least squares regression models, in which the parameter estimates are expressed in terms of the mean of the continuous dependent variable relative to the mean of continuous predictor variables or relative to different means for the dependent variable across categories of a discrete main effect (as in analysis of variance or analysis of covariance).

The structure of the SAS commands for this analysis is fairly simple, shown here for the analysis of the well contamination data using the single predictor variable of well construction type:

```
proc catmod data=bbb.nnset1;
 model znit=wellgrp;
run;
```

A separate group, or ''sample,'' is constructed by CATMOD for each variable or combination of explanatory variables. In this initial example, the response variable is ZNIT, which has two levels (nitrate concentrations less than or equal to 1, and greater than 1); there are three populations, referring to the three different categories of well construction type in the predictor variable (buried slab, other large diameter, and small diameter); the overall sample size is 1061; the row marginal frequencies from the previous crosstabulation table (Table 1) are reproduced as the ''Sample Size'' values here for each of the ''POPULATION PROFILES.'' The response

TABLE 3 Structural Information About the PROC-CATMOD Logistic Regression Model for Well Contamination as a Function of Type of Well Construction

CATMOD PROCEDURE

Response: ZNIT	Response levels	(R) =	2
Weight variable: none	Populations	(S) =	3
Data set: NNSET1	Total Frequency	(N) =	1061
Frequency missing: 88	Observations	(Obs) =	1061

POPULATION PROFILES

Sample	WELLGRP	Sample size
1	Buried slab	263
2	Other large	312
3	Small diameter	486

RESPONSE PROFILES

Response	ZNIT
1	0
2	1

variable and its values are presented in the ''RESPONSE PROFILES'' information in Table 3; the value of ZNIT equal to 0 (nondetectable levels of nitrate contamination) is the first response category, and ZNIT equal to 1 (detectable levels of nitrate contamination) is the second response category. What is important to note about this ordering of the response profiles is that CATMOD, unlike most other procedures for analyzing regression-type models, bases its model fit on (that is, the model is ''normed on'') the first category limited. Thus, resulting model parameters constitute contrasts against the probability that a rural well will not be contaminated.

Estimation of the logistic regression model is undertaken using maximum likelihood methods. In this instance, convergence to acceptable parameter estimates is attained in just five iterations. The maximum likelihood analysis of variance table (included in Table 4) presents Wald chi-square statistics measuring model effects, much like a partial F test in least squares analysis. The maximum likelihood results show that both the intercept and the effect of differences in well construction types are statistically significant. With the paramaterization process followed by the CATMOD procedure, the intercept provides a reference, or baseline, comparison level of the predictor variable, and the other parameters then estimate incremental changes compared to that baseline level for the other categories of the predictor variable. Such models are called reference cell models, or deviation from the mean models. Summary results from the application of PROC CATMOD to this model are shown in Table 4.

Here, the intercept (Parameter 1) may be interpreted as the estimated average log of the odds across the three well types of a well not being contaminated. Parameter 2 is the estimated change in log odds of a well not being contaminated when the construction type is buried slab, compared to the average across all three well types. The logic behind this model specification differs from what happens typically in conventional analysis of variance linear models, in which the omitted category (that is, the category for which no separate dummy variable is constructed, to avoid exact collinearity among the set of indicator variables distinguishing among mean levels of a continuous dependent variable) becomes the reference category. The important difference to note here is that CATMOD normalizes its model structure by contrasting the odds of a specific

TABLE 4 Summary of Results of Maximum Likelihood Estimation of the PROC CATMOD Logistic Regression Model for Well Contamination as a Function of Type of Well Construction

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub-iteration	−2 Log likelihood	Convergence criterion	Parameter Estimates 1	2	3
0	0	1470.8583	1.0000	0.0000	0.0000	0.0000
1	0	1156.2759	0.2139	−0.4414	−0.2050	1.0842
2	0	1147.5931	0.007509	−0.5670	−0.1033	−1.3678
3	0	1147.4258	0.000146	−0.5899	−0.0805	−1.4141
4	0	1147.4257	9.8712E-8	−0.5905	−0.0799	−1.4154
5	0	1147.4257	4.974E-14	−0.5905	−0.0799	−1.4154

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-square	Prob
INTERCEPT	1	54.41	0.0000
WELLGRP	2	238.14	0.0000
LIKELIHOOD RATIO	0	—	—

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard error	Chi-square	Prob
INTERCEPT	1	−0.5905	0.0801	54.41	0.0000
WELLGRP	2	−0.0799	0.1099	0.53	0.4669
	3	−1.4154	0.1290	120.46	0.0000

category of the predictor variable (well type) being at the lowest-valued category of the outcome variable (not contaminated) against the average odds of all three well types not being contaminated.

Parameter 3 is the estimated change in log odds of a well not being contaminated when the construction type is other large diameter, compared to the average for all wells. The statistically significant negative estimated effect for the final parameter translates into concluding that other large diameter wells are much less likely to be safe from nitrate contamination than are all three types of wells on average.

These conclusions are supported by looking at the ''Row Pct'' values in the crosstabulation table above (Table 1), which compare the percentages of wells of each of the three types that are either safe (ZNIT = 0) or unsafe (ZNIT = 1), and by comparing these results with the results on average, which are given by the marginal column percentages (44.49% safe, and 55.51% unsafe). Of the buried slab wells, 33.84% are safe and 66.16% unsafe, in terms of having measurable levels of nitrate contamination. In contrast, just 11.86% of other large diameter wells do not have detectable levels of nitrates, while 88.14% are unsafe by this measure. Finally, 71.19% of small diameter wells are categorized as safe, versus 28.82% that are unsafe. Thus, both buried slab and other large diameter wells are much more likely to be contaminated with nitrates than are small diameter wells, and it is not particularly surprising to find through the CATMOD results that the small diameter wells are significantly less likely to be contaminated that the average well of any construction type. Similarly, because both buried slab and other large diameter wells are more likely than the average well to be contaminated with nitrates, it is not terribly surprising that Parameter 2 does not show a significant difference between buried slab wells and the baseline average contamination rates for all three types of well construction.

I. Logistic Regression: An Example Using Continuous Predictor Variables

It is more efficient computationally, and more elegant in terms of what is shown on a printout, to run a different version of logistic regression when the predictor variable is continuous. This is so because CATMOD creates a separate group (''population'') for each combination of numeric values that the explanatory variables assume. (As an example, if we were to use CATMOD to analyze the effect of well depth and well age, both continuous explanatory variables, together with the three-way distinction among well construction types on nitrate contamination in our sample of well data, there would be 711 different ''populations,'' one for each unique combination of the depth to which wells are drilled, how old each well is, and the construction type of each well. Owing to missing data for the well depth and well age variables, only 823 observations are available for this analysis.) Accordingly, we will use the more appropriate procedure in SAS: PROC LOGISTIC.

To run this analysis using PROC LOGISTIC, it is necessary to create two dummy variables as part of the model-building process. Note that these dummy variables are created automatically by CATMOD for the case of a categoric variable addressed immediately above. The program that produced the results shown below was:

```
data bbb.nnset1;
  set bbb.nnset1;
  if nit le 1 then znit=0;
  if nit gt 1 then znit=1;
  if nit=. then znit=.;
  buriedsb=(WELLGRP='Buried Slab');
  otherlgd=(WELLGRP='Other Large');
  smalldia=(WELLGRP='Small D');
run;
/* ZNIT=0 : NITRATE LEVEL <= 1
   ZNIT=1 : NITRATE LEVEL > 1*/
proc logistic data=bbb.nnset1;
  model znit=otherlgd smalldia wdepth70 age70/ctable influence iplots corrb
itprint;
run;
```

The results from this SAS data analysis using PROC LOGISTIC are presented below. Although this program statement appears to be considerably more elaborate than the program used for the original CATMOD analysis using only one categoric predictor, in part this discrepancy is due to the desire to make clear the process of forming the dummy variables that were used in the model. Here, ZNIT is the same variable encountered previously in the CATMOD analysis to distinguish between detectable and nondetectable levels of nitrate contamination. In the SAS model statement, ZNIT (which may be expressed in either capitals or lower-case letters) is predicted by two dummy variables distinguishing other large-diameter wells from buried slab wells (**otherlgd**) and distinguishing small diameter wells from buried slab wells (**smalldia**), the depth to which the well is dug (**wdepth70**), and the age of the well (**age70**). Both **wdepth70** and **age70** are continuous variables. The model statement contains a number of optional commands requesting a classification table that evaluates the success with which the model correctly categorizes observations (**ctable**), various diagnostic tolls to evaluation model adequacy (**influence**, **iplots**, and **corrb**), and details of the iterations required for the maximum likelihood estimation process (**itprint**).

The results of this model estimation follow. Although the CATMOD and LOGISTIC procedures in SAS are alike in using maximum likelihood procedures to estimate model param-

TABLE 5 Structural Information About the PROC LOGISTIC Multiple Logistic Regression Model for Well Contamination as a Function of Type of Well Construction, Well Depth, and Well Age

The LOGISTIC Procedure

Data Set: BBB.NNSET1
Response Variable: ZNIT
Response Levels: 2
Number of Observations: 823
Link Function: Logit

Response Profile

Ordered value	ZNIT	Count
1	0	392
2	1	431

Note: 326 observation(s) were deleted due to missing values for the response or explanatory variables.

eters, PROC LOGISTIC employs what is known as the Fisher scoring method, developed by Sir R. A. Fisher (also of F-statistic fame, among his many other contributions to contemporary statistical practice), whereas PROC CATMOD (and a related procedure, PROC GENMOD, which is used for the analysis of generalized linear models) use Newton-Raphson algorithms for purposes of parameter estimation. With the estimation process followed by PROC LOGISTIC, the odds ratios that tell the researcher which effects in the model are important substantively are calculated through incremental effects parameterization, in which any categorical predictor variables are expressed as a series of dummy variables with values of either zero or one for each observation. To find the relevant odds ratios, it is necessary simply to exponentiate the parameter estimates; the resulting odds ratios are adjusted for the effects of any continuous predictor variables or any other categoric variables included in the model.

Table 5 summarizes the structure of the multiple logistic regression model for the well contamination data. The first thing to note about the results of this multiple logistic model is the Response Profile, which provides a convenient outline of how the model has been structured and of what it therefore will estimate. In particular, the Response Profile information tells us that the model is based on the probability that a randomly selected well contains safe water, that is, water in which there was not a detectable level of possible nitrate contamination. This is indicated by the Ordered Value of 1 (the first, and thus lower, ordered value). The model is estimated using data for 392 wells that are nitrate-free (ZNIT=0, and Ordered Value=1) and 431 that contain detectable levels of nitrate contamination (ZNIT=1, and Ordered Value=2). A total of 326 observations from the original data set of 1,149 have been omitted because one or more of the variables that appear in the SAS model command above contain missing values for those 326 wells. Note that this implies a rather heavy rate of censoring of data that may make these results somewhat unstable compared to what might be found in a parallel analysis in which less missing data would be present. However, the researcher generally has to take what she or he can get; in this case, there simply is a great deal of missing information on at least one of the variables of interest for this analysis.

The maximum likelihood estimation process followed in this multiple logistic regression

TABLE 6 Summary of Results of Maximum Likelihood Estimation of the PROC LOGISTIC Multiple Logistic Regression Model for Well Contamination as a Function of Type of Well Construction, Well Depth, and Well Age

Maximum likelihood iterative phase						
Iter Step	−2 Log L	INTERCPT	OTHERLGD	SMALLDIA	WDEPTH70	AGE70
0 INITIAL	1139.071450	−0.094846	0.000000	0.000000	0.000000	0.000000
1 IRLS	857.058348	−0.870225	−0.734660	1.363865	0.003275	−0.000268
2 IRLS	843.757687	−1.074117	−0.998767	1.309189	0.005969	−0.000511
3 IRLS	843.130296	−1.135322	−1.025882	1.274347	0.006891	−0.000651
4 IRLS	843.127827	−1.139468	−1.025468	1.272477	0.006955	−0.000663
5 IRLS	843.127827	−1.139487	−1.025463	1.272469	0.006955	−0.000663
Last Change in −2 Log L: 4.798062E-8						
Last evaluation of gradient						
INTERCPT		OTHERLGD		SMALLDIA	WDEPTH70	AGE70
0.000690799		−0.000037789		0.000725699	0.2385169008	0.015626784

analysis is summarized in Table 6. The estimates converge quickly, but in more elaborate models, a much larger number of iterations,. and hence possibly substantially more computing time, may be required to attain convergence. It also is possible that convergence may not be attained at all, or at least not within a reasonable number of iterations. If necessary, the user can set a higher number of iterations, or may choose to establish a less stringent convergence criterion, at the possible cost of producing less accurate, and hence less useful, estimates. Estimates are computed using a process called iteratively reweighted least squares (IRLS in the table) (SAS Institute, Inc., 1989, p. 1088).

The criteria that are used for evaluating how well the specified model fits the data are presented in Table 7. Both the −2 Log L value (−2 log likelihood, or G^2) and the score statistic use the chi-square distribution to test whether the explanatory variables jointly are significant predictors of the outcome variable. AIC (the Akaike Information Criterion) and SC (Schwartz's Bayesian Criterion) serve much the same role, while adjusting for the number of explanatory variables in the model and for the sample size. The value of −2 Log Likelihood is found by calculating

$$-2 \text{ Log L} = -2 \sum_j w_j \log(\hat{\pi}_j),$$

TABLE 7 Evaluation of the Adequacy of the PROC LOGISTIC Multiple Logistic Regression Model for Well Contamination as a Function of Type of Well Construction, Well Depth, and Well Age

The LOGISTIC procedure Testing global null hypothesis: BETA = 0			
Criterion	Intercept only	Intercept and covariates	Chi-square for covariates
AIC	1141.071	853.128	—
SC	1145.784	876.693	—
−2 LOG L	1139.071	843.128	295.944 with 4 DF (p = 0.0001)
Score	—	—	262.656 with 4 DF (p = 0.0001)

where w_j is the weight of the jth observation. The Akaike Information Criterion value is found from

$$\text{AIC} = -2 \text{ Log L} = 2(k + s),$$

where k is the number of ordered values of the response variable (here, $k = 2$) and s is the number of explanatory variables (here, $s = 4$). When comparing alternative models, a smaller value of AIC indicates a more desirable result. The Schwartz Criterion calculation is computed from

$$\text{SC} = -2 \text{ Log L} + (k + s)\log(n),$$

where n is the total number of observations. Smaller values of SC indicate better-fitted models. Some additional details about how these three statistics are estimated are presented in the appendix.

In Table 7, the −2 Log L and Score results under the heading of *Chi-Square for Covariates* provide an analogue to the least squares regression model (or explained) sum of squares, and both are highly significant ($p < .0001$) with 4 degrees of freedom. ''Covariates'' is alternative terminology for predictor variables.

A crude calculation may be performed of a ''pseudo-R^2'' statistic, using the *Chi-Square for Covariates* value as the numerator and the *Intercept Only* value (which is the sum of *Intercept and Covariates* and *Chi-Square for Covariates*) as the denominator parallel to the least squares concept of total sum of squares. The *Intercept and Covariates* value serves a function analogous to that of the least squares error sum of squares. Here

$$\text{pseudo-}R^2 = 295.944/1139.071 = .2598$$

which indicates a moderate degree of explanatory power on the interval (0, 1).

With it now an established fact that the estimated logistic regression model for well nitrate contamination is statistically significant overall, it remains to determine which of the estimated parameters within that model are significant, and which may be of relatively greater importance than others. This evaluation also requires a substantive interpretation of what the significant parameters convey about the physical processes underlying what the estimated effects suggest. (A highly useful guide to interpretation of logistic regression model parameters, and a source of excellent advice on alternative logistic modeling strategies, is Stokes et al. (1995).

The values under the heading *Parameter Estimate* are the estimated coefficients of the fitted logistic regression model. They may be used to write out the estimated logistic regression equation explicitly:

$$\begin{aligned}\text{logit}(\hat{\pi}) = {} & -1.1395 - 1.0255 \text{ OTHERLGD} + 1.2725 \text{ SMALLDIA} \\ & + 0.00696 \text{ WDEPTH70} - 0.00066 \text{ AGE70}\end{aligned}$$

The intercept value of −1.1395 estimates the log odds of a randomly selected shallow and newly-dug buried slab well being safe from measurable nitrate contamination, which translates to the odds ratio

$$e^{-1.1395} = .320$$

The expected change in the log odds of a well being nitrate-free if the well is other large diameter is given by the parameter estimate for OTHERLGD. Since this value (−1.0255) is negative, this indicates that, controlling for well depth and age of the well, other large-diameter wells are

$$e^{-1.0255} = 0.359$$

or about 36% less likely to be safe from nitrate contamination than are buried-slab wells.

The estimated slope for SMALLDIA demonstrates that, controlling for well depth and age, small-diameter wells are

$$e^{1.2725} = 3.570$$

or more than three and a half times more likely to be free of detectable nitrate contamination than are buried-slab wells.

Interpretation of the parameter estimates for the two continuous predictor variables in the model leads to the conclusion that the log odds ratio of the probability of a well being safe increases by 0.00696 for each additional foot of depth to which the well is dug and decreases by 0.00066 for each additional year that has elapsed since the well was dug. The respective odds ratios of a well being free of detectable nitrates are thus

$$e^{0.00696} = 1.007$$

for well depth and

$$e^{-0.00066} = 0.999$$

for well age. Both of these results are very nearly equal to one, which would indicate no difference in the relative probabilities of deeper/shallower and older/younger wells being free from detectable levels of nitrates. However, the standard error for the well depth variable does permit the effect of depth to be statistically significant. A substantive interpretation would be to suggest that each additional foot of well depth increases the odds of the well being nitrate-free by about seven-tenths of one percent.

Standard errors (s(β)) for these model parameter estimates are estimated from the square root of a quadratic form of the covariances among the parameter estimates:

$$s(\beta) = [(1, \mathbf{x}')\mathbf{V_b}(1, \mathbf{x}')]^{0.5},$$

where V_b is the estimated covariance matrix of the parameter estimates (not shown here). The statistical significance of each parameter estimate in the model is conducted using the *Wald Chi-Square* statistic, which is the square of a parameter estimate (β) divided by its standard error. Denoting the Wald chi-square as χ^2_w, we find these test statistics from

$$\chi^2_w = \left[\frac{\beta_i}{s(\beta_i)}\right]^2$$

The Wald test statistic is analogous to a squared t-statistic, or a partial F-test, in ordinary least squares regression terms. The p values (**Pr > Chi-Square**) are calibrated by comparing them against the tabulated ''critical'' value of the theoretical χ^2 distribution.

The *Standardized Estimate* values shown in Table 8 permit a more generally valid comparison among the parameter estimates, which otherwise vary drastically from each other not because of their relative importance to the model, but rather due to their units of measurement. The standardized estimates (z) are calculated from the expression

$$z = \frac{\beta_i}{[\pi/(3)^{0.5}]/s(\beta)}$$

where $[\pi/(3)^{0.5}]$ is the standard deviation of the underlying logistic distribution. The standardized estimates of the intercept parameters are set to missing.

To summarize, based on the results reported in Table 8, compared to the buried slab baseline, a well is significantly less likely to be safe from detectable nitrate contamination if it

TABLE 8 Interpretation of the PROC LOGISTIC Multiple Logistic Regression Parameter Estimates for the Model of Well Contamination as a Function of Type of Well Construction, Well Depth, and Well Age

		Analysis of maximum likelihood estimates					
Variable	DF	Parameter estimate	Standard error	Wald chi-square	Pr > chi-square	Standardized estimate	Odds ratio
INTERCPT	1	−1.1395	0.1737	43.0388	0.0001	—	—
OTHERLGD	1	−1.0255	0.2712	14.3027	0.0002	−0.246236	0.359
SMALLDIA	1	1.2725	0.2269	31.4551	0.0001	0.348335	3.570
WDEPTH70	1	0.00696	0.00143	23.7262	0.0001	0.419118	1.007
AGE70	1	−0.00066	0.00363	0.0333	0.8552	−0.009693	0.999

Association of predicted probabilities and observed responses

Concordant	= 82.8%	Somers' D	= 0.660
Discordant	= 16.9%	Gamma	= 0.662
Tied	= 0.3%	Tau-a	= 0.329
(168952 pairs)		c	= 0.830

is of other-large-diameter construction, and is significantly more likely to be nitrate-safe if it is of small-diameter construction and if the well is dug more deeply. The effect of well age is not significant ($p = 0.8552$).

Table 8 also contains the results from applying some commonly-used nonparametric measures of association that are helpful in assessing the validity of the fitted model, under the heading *Association of Predicted Probabilities and Observed Responses*. This information is divided into two columns. The left-hand column provides some basic calculations about trends among the data values, and the right-hand column makes use of that information to estimate four rank correlation indexes. We will look at the left-hand column first.

For all pairs of observations with different values on the response variable, a pair is concordant if the observation with the larger ordered value of the response (that would be a well that has a detectable level of nitrate contamination) also has a lower predicted probability of the event (that is, being free of contamination) than an observation having a smaller ordered value of the response (which would be a well that is free of detectable levels of nitrates). A pair of observations are discordant if the observation with the larger ordered value of the response has a higher probability of the predicted event compared to an observation having the smaller ordered value of the response (SAS Institute, 1995, p. 22). Pairs that are neither concordant nor discordant are referred to as ties. Here, far more pairs of wells are concordant (82.8%) than are discordant (16.0%), and there are very few ties (just 0.3%).

Where n is the total number of observations in the sample, C is the number of concordant pairs, D is the number of discordant pairs, and t is the total number of pairs of observations having different response values, the results shown in the right-hand column of Table 8 are computed easily from:

$$\text{Somer's D} = (C - D)/t$$

$$\text{Gamma} = (C - D)/(C + D)$$

$$\text{Tau-a} = (C - D)/.5n(n - 1)$$

and

$$c = (C = .5)(t - C - D))/t.$$

The value of c, which is the area under a receiver operating characteristic (ROC) curve, a graphical aid used often in epidemiological, medical, and similar research (Bamber, 1975; Hanley and McNeil, 1982), is related to Somer's D, by the formula:

$$\text{Somer's } D = 2(c - .5).$$

Another way to assess the validity of the logistic regression model is to evaluate the rate of success with which the fitted model correctly classifies each of the two outcomes—here, of contaminated and uncontaminated wells. The results of such an assessment are presented in Table 9.

The predicted values are computed from the expression

$$\hat{\pi} = \frac{1}{1 + \exp(\hat{\alpha} - \beta' x)}$$

where $\hat{\alpha}$ is the estimated intercept, β is the vector of estimated slope parameters, and x is the vector of explanatory variables for that observation. In general, the predicted probability from a binary logistic regression model is the estimated probability that an observation is an event (here, a ''safe'' well, with no detectable level of nitrate contamination). As an example, for an other-large-diameter well (otherlgd=1, smalldia=0) that is dug to the depth of 64 feet and that is 4 years old, the estimated probability of such a well being safe from nitrate contamination is

$$\begin{aligned}\hat{\pi} &= 1/[1 + \exp(1.1395 + 1.0255\ (1) - 0.00696\ (64) + 0.00066\ (4))] \\ &= 1/[1 + \exp(1.7222)] \\ &= 1/[1 + 5.603] \\ &= 0.15\end{aligned}$$

A word of caution is in order here. Using the same data to test the predictive adequacy of the model that were employed to estimate the model in the first place imparts a bias to the results of prediction and classification efforts. There are two standard ways around this difficulty: (1) You could use a new set of observations to test the predictive validity of the model. However, the practical difficulties associated with having only a single sample available in most application situations make it unlikely that more than one data set will be sitting around to be evaluated; instead, it is common to split the sample, estimating the model with one half and then evaluating predictive validity with the other half. This assumes, of course, that there are enough observations in the sample for this to be feasible. (2) Alternatively, a jackknife procedure could be employed, in which a single unified sample is used. With jackknifing, one observation is omitted each time, and that observation then is classified into either of the two dichotomous outcomes based on the model that has been estimated without the observation that is being classified. The results shown in Table 9 are the default output from PROC LOGISTIC.

The results in Table 9 require some explanation. This is the default SAS listing of classifications for probabilities ranging from the smallest estimated probability rounded down to the nearest 0.02 (.100), to the highest estimated probability rounded up to the nearest 0.02 (1.00), with increments of 0.02. The columns labeled *Correct* and *Incorrect* show the frequency with which observations are, respectively, correctly and incorrectly classified as events (safe from nitrate contamination) or nonevents (not safe from nitrate contamination), for each probability

TABLE 9 Predicted Values and Classification Table of the PROC LOGISTIC Multiple Logistic Regression Parameter Estimates for the Model of Well Contamination as a Function of Type of Well Construction, Well Depth, and Well Age

	The LOGISTIC procedure classification table								
	Correct		Incorrect		Percentages				
Prob. level	Event	Non-event	Event	Non-event	Correct	Sensitivity	Specificity	False POS	False NEG
0.100	392	0	431	0	47.6	100.0	0.0	52.4	—
0.120	388	47	384	4	52.9	99.0	10.9	49.7	7.8
0.140	373	137	294	19	62.0	95.2	31.8	44.1	12.2
0.160	368	163	268	24	64.5	93.9	37.8	42.1	12.8
0.180	367	174	257	25	65.7	93.6	40.4	41.2	12.6
0.200	366	179	252	26	66.2	93.4	41.5	40.8	12.7
0.220	364	181	250	28	66.2	92.9	42.0	40.7	13.4
0.240	364	181	250	28	66.2	92.9	42.0	40.7	13.4
0.260	364	182	249	28	66.3	92.9	42.2	40.6	13.3
0.280	357	193	238	35	66.8	91.1	44.8	40.0	15.4
0.300	343	233	198	49	70.0	87.5	54.1	36.6	17.4
0.320	328	272	159	64	72.9	83.7	63.1	32.6	19.0
0.340	317	296	135	75	74.5	80.9	68.7	29.9	20.2
0.360	309	316	115	83	75.9	78.8	73.3	27.1	20.8
0.380	295	330	101	97	75.9	75.3	76.6	25.5	22.7
0.400	291	335	96	101	76.1	74.2	77.7	24.8	23.2
0.420	287	340	91	105	76.2	73.2	78.9	24.1	23.6
0.440	284	344	87	108	76.3	72.4	79.8	23.5	23.9
0.460	283	344	87	109	76.2	72.2	79.8	23.5	24.1
0.480	283	345	86	109	76.3	72.2	80.0	23.3	24.0
0.500	282	346	85	110	76.3	71.9	80.3	23.2	24.1
0.520	282	348	83	110	76.5	71.9	80.7	22.7	24.0
0.540	281	348	83	111	76.4	71.7	80.7	22.8	24.2
0.560	281	349	82	111	76.5	71.7	81.0	22.6	24.1
0.580	277	355	76	115	76.8	70.7	82.4	21.5	24.5
0.600	271	358	73	121	76.4	69.1	83.1	21.2	25.3
0.620	269	362	69	123	76.7	68.6	84.0	20.4	25.4
0.640	259	372	59	133	76.7	66.1	86.3	18.6	26.3
0.660	248	374	57	144	75.6	63.3	86.8	18.7	27.8
0.680	227	378	53	165	73.5	57.9	87.7	18.9	30.4
0.700	212	385	46	180	72.5	54.1	89.3	17.8	31.9
0.720	198	390	41	194	71.4	50.5	90.5	17.2	33.2
0.740	186	392	39	206	70.2	47.4	91.0	17.3	34.4
0.760	171	395	36	221	68.8	43.6	91.6	17.4	35.9
0.780	149	405	26	243	67.3	38.0	94.0	14.9	37.5
0.800	118	408	23	274	63.9	30.1	94.7	16.3	40.2
0.820	96	411	20	296	61.6	24.5	95.4	17.2	41.9
0.840	85	415	16	307	60.8	21.7	96.3	15.8	42.5
0.860	73	419	12	319	59.8	18.6	97.2	14.1	43.2
0.880	58	424	7	334	58.6	14.8	98.4	10.8	44.1
0.900	40	425	6	352	56.5	10.2	98.6	13.0	45.3
0.920	32	428	3	360	55.9	8.2	99.3	8.6	45.7
0.940	26	428	3	366	55.2	6.6	99.3	10.3	46.1
0.960	17	430	1	375	54.3	4.3	99.8	5.6	46.6
0.980	9	430	1	383	53.3	2.3	99.8	10.0	47.1
1.000	0	431	0	392	52.4	0.0	100.0	—	47.6

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

cutpoint. As an example, for the cutpoint probability of 0.400, 291 events (that is, uncontaminated wells) and 335 nonevents (that is, contaminated wells) are classified correctly, while 96 uncontaminated wells and 101 contaminated wells were classified incorrectly as the opposite outcome.

The five columns under the heading *Percentage* provide alternative ways to assess the predictive accuracy of the model. The *Correct* column shows the probability that the model correctly classifies the sample data for each probability cutpoint. Without specially specified prior probabilities, the percentage correct is simply the number of correctly classified observations divided by the total number of observations, multiplied by 100. For the 0.400 cutpoint, this value is found from $[(291 + 335)/823] \times 100 = 76.1\%$.

The value labeled *sensitivity* is the proportion of events (that is, safe wells) out of all events, multiplied by 100. For the cutpoint of 0.400, the sensitivity equals $(291/392) \times 100 = 74.2\%$. The *specificity* value is the ratio of correctly classified nonevents (unsafe wells) to the total number of nonevents, times 100; here, for the 0.400 cutpoint, the value of specificity is $(335/431) \times 100 = 77.7\%$.

The column labeled *False POS* (false positives) is the ratio of the number of nonevents (that is, unsafe wells) incorrectly classified as events to the total number of events, multiplied by 100. At the 0.400 cutpoint, this value equals $[(96/(291 + 96)] \times 100 = 24.8$. Finally, *False NEG* refers to the false negative rate, which is found as the ratio of the number of events (that is, safe wells) classified incorrectly as nonevents (unsafe wells) to the total number of observations that were classified as nonevents, multiplied by 100. Here, for the cutpoint of 0.400, the false negative rate is $[101/(335 + 101)] \times 100 = 23.2\%$.

IV. THE LEAST-SQUARES APPROACH: TWO-GROUP DISCRIMINANT ANALYSIS

Discriminant analysis operates with three fundamental goals:

1. It is designed to determine which variables do the best job of differentiating between observations that belong in one group and those that belong in the other group.

2. The variables that are thus identified as optimal discriminators are then used to develop a parsimonious prediction equation that will ''boil down'' the set of potential discriminators to those that are the most effective ones for distinguishing one possible outcome from the other. The objective is to separate the means of the two groups as far from each other as possible, that is, to maximize the between (or explained) sum of squares, and to produce groups that are as homogeneous internally as possible, which is equivalent to minimizing the within-group (the error, or unexplained) sum of squares. Another way to express this second objective is to say that discriminant analysis is designed to maximize the ratio of between-groups sum of squares to within-groups sum of squares, to provide the maximum possible separation, or discrimination, between the two groups.

3. The model or models that provide the best results are then used to classify future observations into one or the other of the two groups, and may be used to inform decisionmakers about policy choices among alternative outcomes. Although the classification operation is not necessarily related to discriminant analysis proper, usually it is a goal that is facilitated and informed usefully by the method of discriminant analysis. The linear combination of predictor variables that maximizes the ratio of between to within sum of squares is known as the linear discriminant function. The values of the new variable (Z) that is formed from the linear discriminant function to provide this optimal degree of group separation are referred to as discriminant scores, and these scores are used to classify future observations. Classification occurs based on

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

a cutoff value that minimizes the number of errors of misclassification or minimizes the costs of misclassification of observations into two mutually exclusive and collectively exhaustive regions. Various forms of such misclassification costs can be investigated.

Two-group discriminant models can be extended rather easily to cases of more than two outcome categories.

A. The Linear Discriminant Model

Fisher's linear discriminant function, or the linear combination of predictor variables that forms the new variable Z, is defined as

$$Z = w_1X_1 + w_2X_2 + \ldots + w_{p-1}X_{p-1} = \boldsymbol{\beta}'\mathbf{X}$$

where Z is the discriminant function defining the dependent variable, the X's represent the independent, or classification, variables, and w_1 and w_2 are the weights of the discriminant function that maximize the value of

$$\lambda = \frac{\text{between-group sum of squares}}{\text{within-group sum of squares}}$$

In the discriminant analysis model, the significance of each discriminating variable is determined by the outcome of testing the null hypothesis that the two groups have the same mean

$$H_0: \mu_1 = \mu_2$$

against the alternative hypothesis that the group means are not different (which is the same as asserting that the discriminating variables do not help distinguish one group from the other)

$$H_a: \mu_1 \neq \mu_2$$

This test can be conducted using an independent two-sample t-statistic, but it is more common in applications of discriminant analysis methods to employ Wilks' lambda for this purpose:

$$\boldsymbol{\Lambda} = \frac{SS_w}{SS_t}$$

where SS_w is the sum of squares within and SS_t is that total sum of squares. The smaller the value of $\boldsymbol{\Lambda}$, the greater the probability that the null hypothesis will be rejected and the greater is the evidence that the discriminating function contributes to separating the two groups successfully.

B. An Example of Two-Group Discriminant Analysis: Population Changes in SMSAs

The author of this chapter has been a participant in recent research projects that have dealt with population growth and human migration at the municipal and state levels in the United States (Koven and Shelley, 1989; Shelley and Koven, 1993). The following example is an extension of that research stream.

This example is drawn from population characteristics for 150 Standard Metropolitan Statistical Areas (SMSAs) from the 1980 United States Census of Population (Characteristics of Population). The dependent variable—that is, the variable for which observations are to be classified between two different categories of outcomes—in this analysis is the percentage change in each SMSA's population between 1970 and 1980. In the original data set, this variable

is continuous, ranging in value from −8.6 (a loss in population of 8.6%) to 69.5 (an increase in population equal to 69.5%).

For purposes of this analysis, and in particular because the population change variable did not show any evidence of following a normal distribution, a decision was made to find a fairly arbitrary cutpoint that would not bias the outcome of the model building and classification process. The cutpoint—15, or 15.0%—was chosen because it divides the data values into two almost equal groupings. This is an important consideration, because discriminant analysis, and other classification methods, tend to be dominated by whichever group of outcome cases has the most observations. Using this cutpoint, 76 cities fall into the lower (<15) category, and 74 fall into the higher category (≥15).

The predictor, or discriminating, variables include the following:

area—land area, in square miles
popsqmi—total population per square mile
pctblk—percentage of population African American in central city of SMSA
pctspan—percentage of population of Spanish heritage in central city of SMSA
medage—median age in central city of SMSA
ov25hsg—percentage of high school graduates among persons 25 years old and older
pctmlab—percentage of males 16 years old and older in the labor force
pctflab—percentage of females 16 years old and older in the labor force
pctunem—percentage unemployed of the civilian labor force
pctman—percentage of employed persons 16 years old and older in manufacturing industries
mfin79—median family income in 1979
pctfpov—percentage of families below the poverty level in 1979

The model examined here is analyzed in two different ways, using the SAS procedures DISCRIM to perform a standard analysis, and STEPDISC to produce a stepwise analysis using the technique of backward elimination. Backward elimination modeling proceeds from the starting point of a repetition of the full model that is evaluated by DISCRIM, and then deletes independent variables one at a time until default criteria for variable elimination or retention are reached. Here is the SAS code used to generate these results:

```
proc discrim pool=test wcov pcov list;
class change;
var area popsqmi pctblk pctspan medage ov25hsg pctmlab pctflab pctunem
  pctman mfin79 pctfpov;

proc stepdisc backward simple stdmean tcorr wcorr;
var area popsqmi pctblk pctspan medage ov25hsg pctmlab pctflab pctunem
  pctman mfin79 pctfpov;
class change;
```

The discrim command includes a requested option to test for the equality of the two covariance matrices for the categories of the dependent variable. It is essential to note that the use of discriminant analysis is not an indiscriminate process. Among the crucial distinctions that must be made correctly for the results of a discriminant analysis to be valid is whether the groups are being compared on an equal footing. The nature of this comparability is evaluated using a chi-squared test for homogeneity of the within-group covariance matrices. The pooled covariance matrix is used unless the test statistic is significant, in which case a different variation of discriminant modeling must be followed. A conclusion of equal covariance matrices leads

to the use of Fisher's linear discriminant function, which operates off the assumption that patterns of covariation among the relevant predictor variables are the same in both groups. Alternatively, if a significant chi-square statistic is found, the assumption of equal covariation within groups is rejected, and a quadratic discriminant function is estimated instead.

It also is important to note that conventional forms of discriminant analysis assume that the variables within each group follow a multivariate normal distribution. When multivariate normality is a reasonable assumption, the discriminant function (also known as the classification criterion) is a function of generalized squared distance (Rao, 1973). Nonparametric alternatives are available, when multivariate normality is not plausible, including kernel methods and the k-nearest-neighbor method (Rosenblatt, 1956; Parzen, 1962).

The probability of observations being classified by the vector of their values of the predictor variables ($\mathbf{x}$) into group t is determined by Bayes' theorem, as

$$p(t|\mathbf{x}) = q_t f_t(\mathbf{x})/f(\mathbf{x})$$

where $p(t|\mathbf{x})$ is the posterior probability of an observation x belonging to group t, q_t is the prior probability of membership in group t, $f_t(\mathbf{x})$ is the group-specific density estimate at $\mathbf{x}$ from group t, and $f(\mathbf{x}) = \sum_t q_t f_t(x)$ is the estimated unconditional density at $\mathbf{x}$. The discriminant analysis partitions a vector space containing p dimensions, where p is the number of predictor variables into regions, R_t, which is the subspace containing all p-dimensional vectors $\boldsymbol{\gamma}$ that maximize $p(t|\boldsymbol{\gamma})$ among all groups. Any observation that lies in region R_t is classified as coming from group t, on grounds that it has the smallest generalized squared distance.

The squared distance from $\mathbf{x}$ to group t is

$$d_t^2(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_t)' V_t^{-1} (\mathbf{x} - \mathbf{m}_t)$$

where $\mathbf{V}_t = \mathbf{S}_t$ (the covariance matrix within group t) if the within-group covariance matrices are used, or $\mathbf{V}_t = \mathbf{S}$ (the pooled covariance matrix) if the pooled covariance matrix is used, $\mathbf{m}_t$ is the p-dimensional vector containing the means of the independent variables in group t. Then, the generalized squared distance from $\mathbf{x}$ to group t, which is used to classify each observation, is

$$D_t^2 = d^2{}_t(\mathbf{x}) + g_1(t) + g_2(t),$$

where $g_1(t)$ equals either $\log_e|\mathbf{S}_t|$ if the within-group covariance matrices are used, or zero if the pooled covariance matrix is used, and $g_2(t)$ equals either $-2 \log_e(q_t)$ if the prior probabilities are not all equal, or zero if the prior probabilities are all equal.

The posterior probability of a single observation, x, belonging to group t then equals

$$p(t|x) = \frac{\exp(-0.5\ D_t^2(x))}{\sum_{ij} \exp(-0.5\ D_{ij}^2(x))}$$

An observation then is classified into a particular group, u, if setting t = u produces the largest value of p(t|x) or the smallest value of $D_t^2(x)$). Different thresholds can be specified to define the cutoff probability that must be attained before an observation is classified.

C. Estimation of the Discriminant Analysis Model

The composition of the data analyzed in this example is summarized in Table 10. The table shows that the sample size is 150, there are 12 predictor variables, the outcome variable has two categories (classes), the number of degrees of freedom for the model equals 149 if a linear discriminant function is employed or 148 if a quadratic discriminant function is used, and 1

TABLE 10 Data Structure for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis				
150 Observations	149 DF total			
12 Variables	148 DF within classes			
2 Classes	1 DF between classes			
Class level information				
Change	Frequency	Weight	Proportion	Prior probability
0	74	74.0000	0.493333	0.500000
1	76	76.0000	0.506667	0.500000

degree of freedom is needed to estimate the difference between the two categories (in general, the Between Classes number of degrees of freedom is the number of categories minus one).

Separate within-group covariance matrices are shown in Table 11. *Change = 0* refers to SMSAs with lower (<15) percentage rates of population change, for which there are 74 − 1 = 73 degrees of freedom. Similarly, *Change = 1* refers to SMSAs experiencing higher (≥15) percentage rates of population growth, and there are 76 − 1 = 75 degrees of freedom for those observations. These are the covariance structures used if the null hypothesis that the separate covariance matrices are equal is not rejected. Table 12 shows the values of the pooled covariance matrix, upon which the discriminant analysis is based if the separate covariance matrices are found to be unequal.

The results shown in Table 13 are the values of the rank of the covariance matrix (here, this is just the number of variables in each matrix), and the natural logarithm of the determinant of the covariance matrix. The latter set of values is used in testing for the equality of the separate covariance matrices, and hence for determining whether a linear or quadratic discriminant function is required. Notice that there is some difference in these natural logarithms.

TABLE 11 Within-Class Covariance Matrices for Discriminant Analysis of Population Change in SMSAs

	Discriminant analysis			Within-class covariance matrices		
			Change = 0	DF = 73		
Variable	AREA	POPSQMI	PCTBLK	PCTSPAN	MEDAGE	OV25HSG
AREA	4529957	−89479	−6127	5102	−540	5670
POPSQMI	−89479	30627	213	165	136	72
PCTBLK	−6127	213	135	−30	−7	−55
PCTSPAN	5102	165	−30	86	−2	4
MEDAGE	−540	136	−7	−2	14	−1
OV25HSG	5670	72	−55	4	−1	67
PCTMLAB	3781	30	−11	−1	−10	12
PCTFLAB	2224	177	−2	−3	−6	10
PCTUNEM	−471	−116	−3	1	−1	−1
PCTMAN	−4824	528	11	−15	−1	−24
MFIN79	1499666	160405	−10360	−320	−180	13778
PCTFPOV	−2153	−168	28	4	−4	−21

TABLE 11 Continued

	Discriminant analysis			Within-class covariance matrices		
			Change = 0	DF = 73		
Variable	PCTMLAB	PCTFLAB	PCTUNEM	PCTMAN	MFIN79	PCTFPOV
AREA	3781	2224	−471	−4824	1499666	−2153
POPSQMI	30	177	−116	528	160405	−168
PCTBLK	−11	−2	−3	11	−10360	28
PCTSPAN	−1	−3	1	−15	−320	4
MEDAGE	−10	−6	−1	−1	−180	−4
OV25HSG	12	10	−1	−24	13778	−21
PCTMLAB	33	21	−3	6	8013	−9
PCTFLAB	21	28	−6	5	8098	−9
PCTUNEM	−3	−6	5	−2	−2185	3
PCTMAN	6	5	−2	55	−253	−1
MFIN79	8013	8098	−2185	−253	6618494	−7745
PCTFPOV	−9	−9	3	−1	−7745	15
			Change = 1	DF = 75		
Variable	AREA	POPSQMI	PCTBLK	PCTSPAN	MEDAGE	OV25HSG
AREA	1395069	−194916	2219	715	−754	1454
POPSQMI	−194916	1039111	1759	3215	897	−1208
PCTBLK	2219	1759	93	3	−1	−24
PCTSPAN	715	3215	3	35	2	−7
MEDAGE	−754	897	−1	2	4	−5
OV25HSG	1454	−1208	−24	−7	−5	65
PCTMLAB	602	−82	−6	−2	−1	13
PCTFLAB	434	149	−10	−2	−4	27
PCTUNEM	32	−125	2	−0	0	−7
PCTMAN	−3373	1268	−13	2	8	−29
MFIN79	132539	559701	−2729	−369	1080	16078
PCTFPOV	256	331	21	4	−1	−14
Variable	PCTMLAB	PCTFLAB	PCTUNEM	PCTMAN	MFIN79	PCTFPOV
AREA	602	434	32	−3373	132539	256
POPSQMI	−82	149	−125	1268	559701	331
PCTBLK	−6	−10	2	−13	−2729	21
PCTSPAN	−2	−2	−0	2	−369	4
MEDAGE	−1	−4	0	8	1080	−1
OV25HSG	13	27	−7	−29	16078	−14
PCTMLAB	12	12	−4	0	6639	−7
PCTFLAB	12	27	−6	−9	8680	−7
PCTUNEM	−4	−6	4	4	−2827	2
PCTMAN	0	−9	4	76	1236	−7
MFIN79	6639	8680	−2827	1236	9502969	−7089
PCTFPOV	−7	−7	2	−7	−7089	11

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 12 Pooled Covariance Matrices for Discriminant Analysis of Population Change in SMSAs

	Discriminant analysis					
	Within-class covariance matrix DF = 148					
Variable	AREA	POPSQMI	PCTBLK	PCTSPAN	MEDAGE	OV25HSG
AREA	2941331	−142910	−1898	2879	−648	3533
POPSQMI	−142910	541683	996	1711	521	−577
PCTBLK	−1898	996	114	−14	−4	−40
PCTSPAN	2879	1711	−14	60	0	−1
MEDAGE	−648	521	−4	0	9	−3
OV25HSG	3533	−577	−40	−1	−3	66
PCTMLAB	2170	−27	−9	−1	−5	13
PCTFLAB	1317	163	−6	−3	−5	18
PCTUNEM	−216	−121	−0	0	−0	−4
PCTMAN	−4088	903	−1	−6	4	−26
MFIN79	806865	362751	−6493	−345	459	14944
PCTFPOV	−932	85	25	4	−3	−18
Variable	PCTMLAB	PCTFLAB	PCTUNEM	PCTMAN	MFIN79	PCTFPOV
AREA	2170	1317	−216	−4088	806865	−932
POPSQMI	−27	163	−121	903	362751	85
PCTBLK	−9	−6	−0	−1	−6493	25
PCTSPAN	−1	−3	0	−6	−345	4
MEDAGE	−5	−5	−0	4	459	−3
OV25HSG	13	18	−4	−26	14944	−18
PCTMLAB	22	17	−3	3	7317	−8
PCTFLAB	17	27	−6	−2	8393	−8
PCTUNEM	−3	−6	4	1	−2510	2
PCTMAN	3	−2	1	66	502	−4
MFIN79	7317	8393	−2510	502	8080221	−7412
PCTFPOV	−8	−8	2	−4	−7412	13

TABLE 13 Within Covariance Matrix Information Used for Test of Equal Covariance Matrices for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis		Within covariance matrix information
CHANGE	Covariance matrix rank	Natural log of the determinant of the covariance matrix
0	12	65.77217
1	12	63.88594
Pooled	12	67.90455

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 14 Test for the Equality of the Two Group Covariance Matrices for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis	Test of homogeneity of within covariance matrices

Notation: K = Number of groups

P = Number of variables

N = Total number of observations—number of groups

N(i) = Number of observations in the ith group − 1

$$V = \frac{\prod |\text{Within SS matrix(i)}|^{N(i)/2}}{|\text{Pooled SS matrix}|^{N/2}}$$

$$RHO = 1.0 - \left[\text{SUM}\ \frac{1}{N(i)} - \frac{1}{N}\right] \frac{2P^2 + 3P - 1}{6(P + 1)(K - 1)}$$

$$DF = .5(K - 1)P(P + 1)$$

Under null hypothesis: $-2\ \text{RHO}\ \ln\left[\frac{N^{PN/2}\ V}{\prod N(i)^{PN(i)/2}}\right]$

is distributed approximately as chi-square (DF)
Test chi-square value = 418.684264
with 78 DF Prob > chi-sq = 0.0001
Since the chi-square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.

Source: Morrison, D.F. (1976). *Multivariate Statistical Methods* p. 252.

Table 14 shows the test for equality (homogeneity) of the within-covariance matrices, using SAS notation that appears directly on the output. The result of the chi-square test ($\chi^2 = 418.684264$), with 78 degrees of freedom, is to reject the null hypothesis of equal covariance matrices ($p < 0.0001$). Consequently, a quadratic discriminant function must be employed using separate covariance matrices. The pairwise generalized squared distances between groups are shown in Table 15. Observations are classified relative to their minimum generalized squared distance from each class.

The results of the quadratic discriminant function's classification of observations based on the 12 predictor variables employed in this analysis are shown in Table 16. The classification

TABLE 15 Pairwise Generalized Squared Distances Between Groups for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis

Pairwise generalized squared distances between groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)'\ COV_j^{-1}\ (\bar{X}_i - \bar{X}_j) + \ln |COV_j|$$

Generalized squared distance to CHANGE

From CHANGE	0	1
0	65.77217	66.17851
1	79.47668	63.88594

TABLE 16 Classification Outcomes for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis

Classification results for calibration data: WORK.SMSA
Resubstitution results using quadratic discriminant function
Generalized squared distance function:

$$D_j^2(X) = (X - \overline{X}_j)' \, COV_j^{-1}(X - \overline{X}_j) + \ln|COV_j|$$

Posterior probability of membership in each CHANGE:

$$Pfr(j|X) = \exp(-.5\, D_j^2(X))/\underset{k}{SUM} \exp(-.5\, D_k^2(X))$$

Posterior probability of membership in CHANGE:

Obs	From CHANGE	Classified into CHANGE	0	1
1	1	1	0.0398	0.9602
2	0	0	0.9994	0.0006
3	1	1	0.0000	1.0000
4	1	1	0.0004	0.9996
5	1	1	0.0703	0.9297
6	1	0*	0.6042	0.3958
7	1	1	0.3227	0.6773
8	1	1	0.0022	0.9978
9	1	1	0.4538	0.5462
10	1	1	0.0191	0.9809
11	1	1	0.0000	1.0000
12	1	1	0.0029	0.9971
13	1	1	0.0148	0.9852
14	1	1	0.0731	0.9269
15	1	0*	0.8847	0.1153
16	1	1	0.1437	0.8563
17	1	1	0.0354	0.9646
18	0	0	0.6397	0.3603
19	0	0	0.8551	0.1449
20	1	0*	0.5141	0.4859
21	0	1*	0.2249	0.7751
22	0	0	1.0000	0.0000
23	1	1	0.1316	0.8684
24	1	1	0.0952	0.9048
25	0	0	0.9824	0.0176
26	0	1*	0.4342	0.5658
27	0	1*	0.4559	0.5441
28	1	1	0.0006	0.9994
29	1	0*	0.5241	0.4759
30	0	0	0.7976	0.2024
31	0	0	0.7008	0.2992
32	0	0	0.9502	0.0498
33	1	0*	0.5589	0.4411
34	0	0	0.9985	0.0015
35	1	1	0.2141	0.7859
36	1	1	0.0000	1.0000
37	1	1	0.0000	1.0000

(*Table continues*)

TABLE 16 Continued

Obs	From CHANGE	Classified into CHANGE	0	1
38	1	1	0.4132	0.5868
39	1	1	0.2639	0.7361
40	1	0*	0.5930	0.4070
41	1	1	0.0000	1.0000
42	1	1	0.0004	0.9996
43	1	1	0.0786	0.9214
44	1	1	0.0000	1.0000
45	1	1	0.0000	1.0000
46	1	1	0.0000	1.0000
47	1	1	0.0000	1.0000
48	1	1	0.0521	0.9479
49	0	0	0.9915	0.0085
50	1	1	0.0225	0.9775
51	1	1	0.0016	0.9984
52	1	1	0.1681	0.8319
53	1	1	0.0000	1.0000
54	0	0	1.0000	0.0000
55	0	0	0.7674	0.2326
56	0	0	0.8109	0.1891
57	0	0	0.9985	0.0015
58	0	0	0.9999	0.0001
59	0	0	0.9669	0.0331
60	0	1*	0.1548	0.8452
61	1	0*	0.9563	0.0437
62	0	0	0.9994	0.0006
63	0	0	1.0000	0.0000
64	0	0	1.0000	0.0000
65	0	0	1.0000	0.0000
66	0	0	1.0000	0.0000
67	1	0*	1.0000	0.0000
68	0	0	1.0000	0.0000
69	0	0	1.0000	0.0000
70	0	0	1.0000	0.0000
71	1	1	0.4317	0.5683
72	1	1	0.0000	1.0000
73	1	1	0.0000	1.0000
74	1	1	0.0000	1.0000
75	0	1*	0.0051	0.9949
76	1	1	0.0000	1.0000
77	1	1	0.0334	0.9666
78	1	1	0.0000	1.0000
79	1	1	0.0001	0.9999
80	0	1*	0.0057	0.9943
81	1	1	0.0000	1.0000
82	1	1	0.0000	1.0000
83	1	1	0.0000	1.0000
84	1	1	0.0000	1.0000
85	1	1	0.0001	0.9999
86	1	1	0.0032	0.9968

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 16 Continued

Obs	From CHANGE	Classified into CHANGE	0	1
87	0	0	0.9432	0.0568
88	0	0	0.9550	0.0450
89	1	1	0.0000	1.0000
90	0	0	0.9992	0.0008
91	0	0	0.9975	0.0025
92	0	0	0.8177	0.1823
93	0	0	0.8629	0.1371
94	0	0	1.0000	0.0000
95	0	0	0.9957	0.0043
96	0	0	0.9831	0.0169
97	0	0	0.8257	0.1743
98	0	0	0.9803	0.0197
99	0	1*	0.0020	0.9980
100	0	0	0.9400	0.0600
101	1	0*	0.7209	0.2791
102	0	0	0.9964	0.0036
103	0	0	0.8078	0.1922
104	0	0	0.9902	0.0098
105	0	0	0.7791	0.2209
106	0	0	1.0000	0.0000
107	0	0	0.9718	0.0282
108	0	0	0.9981	0.0019
109	0	0	1.0000	0.0000
110	0	0	0.9761	0.0239
111	0	0	0.8493	0.1507
112	0	0	0.6760	0.3240
113	1	0*	0.7784	0.2216
114	0	0	1.0000	0.0000
115	0	0	0.9083	0.0917
116	0	0	0.9987	0.0013
117	0	0	1.0000	0.0000
118	1	0*	0.7791	0.2209
119	0	0	0.6266	0.3734
120	0	0	0.7568	0.2432
121	0	0	0.6508	0.3492
122	1	1	0.3443	0.6557
123	1	0*	0.5025	0.4975
124	1	0*	0.5054	0.4946
125	1	1	0.0012	0.9988
126	1	0*	0.6708	0.3292
127	1	0*	0.6439	0.3561
128	1	1	0.1484	0.8516
129	1	0*	0.8205	0.1795
130	1	1	0.3282	0.6718
131	1	1	0.0603	0.9397
132	1	1	0.1474	0.8526
133	1	1	0.1865	0.8135
134	1	1	0.4881	0.5119
135	0	0	0.6337	0.3663

(*Table continues*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 16 Continued

Obs	From CHANGE	Classified into CHANGE	0	1
136	0	0	0.5554	0.4446
137	0	0	0.9351	0.0649
138	0	0	1.0000	0.0000
139	1	1	0.0625	0.9375
140	0	0	1.0000	0.0000
141	0	0	0.9888	0.0112
142	0	0	0.9397	0.0603
143	0	0	0.9999	0.0001
144	0	0	0.9866	0.0134
145	1	1	0.0017	0.9983
146	0	0	0.7953	0.2047
147	1	0*	0.9366	0.0634
148	0	0	0.9980	0.0020
149	0	0	0.9995	0.0005
150	0	0	0.9891	0.0109

* Misclassified observation.

process is conducted using resubstitution methods, which are known otherwise as jackknifing. Resubstitution methods generally involve estimating the discriminant function without one observation, thus providing an unbiased estimate of the accuracy with which the discriminant function successfully classified the sample observations.

The column in Table 16 labeled *From CHANGE* shows the actual classification of a metropolitan area into either category 0 (lower rates of population growth) or category 1 (higher rates of population growth). The column labeled *Classified into CHANGE* denotes the category for which the higher posterior probability was produced by the quadratic discriminant function. Each observation that is classified incorrectly, according to its larger probability value, is so indicated by an asterisk (*). A quick glance through the classification outcomes shows that most observations are classified correctly into either category 0 or category 1, but that there are 24 incorrectly classified cases.

A convenient summary of the strength of this quadratic discriminant model is provided in Table 17, which also is referred to variously as a ''truth table'' or ''confusion table.'' The marginal values of the table show that 84 metropolitan areas were classified by the quadratic discriminant function into CHANGE category 0—that is, low population growth—and the remaining 66 observations were classified into the high-population-growth category (CHANGE = 1). This contrasts with the known distribution, of 74 metropolitan areas in category 0 and 76 in category 1.

A closer examination of the table shows that 67 of the 74 lower-population-change metropolitan areas (90.54%) are classified correctly into CHANGE = 0, while 59 of the 76 higher-population-change metropolitan areas (77.63%) were classified correctly into CHANGE = 1. Adding together the two main-diagonal values of the correctly-classified observations in the table and dividing by the total number of cases yields a total percentage of correct classifications equal to (67 + 59)/150 = 126/150 = .84. Thus, 84% of all 150 cases were correctly classified as either low-growth or high-growth metropolitan areas.

Two separate error rates are computed. The first, for outcome category 0, is the proportion

TABLE 17 Summary of Classification Outcomes for Discriminant Analysis of Population Change in SMSAs

Discriminant analysis

Classification summary for calibration data: WORK.SMSA
Resubstitution summary using quadratic discriminant function
Generalized squared distance function:

$$D_j^2(X) = (X - \overline{X}_j)' \, COV_j^{-1} (X - \overline{X}_j) + \ln|COV_j|$$

Posterior probability of membership in each CHANGE:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) / \sum_k \exp(-.5\, D_k^2(X))$$

Number of observations and percent classified into CHANGE:

From CHANGE	0	1	Total
0	67	7	74
	90.54	9.46	100.00
1	17	59	76
	22.37	77.63	100.00
Total	84	66	150
Percent	56.00	44.00	100.00
Priors	0.5000	0.5000	

Error count estimates for CHANGE:

	0	1	Total
Rate	0.0946	0.2237	0.1591
Priors	0.5000	0.5000	

of metropolitan areas actually characterized by low population growth (CHANGE = 0) that were misclassified as high-growth areas. It is calculated as 7/74 = .0946. The second error rate, for outcome category 1, is the proportion of metropolitan areas actually characterized by high population growth (CHANGE = 1) that were misclassified as low-growth. That value is found from computing 17/76 = .2237. The Priors values indicate that each category was assumed to have the same probability of occurring. Clearly, the quadratic discriminant function does a more satisfactory job of detecting metropolitan areas that were characterized by lower population growth rates than of classifying correctly those that experienced higher rates of growth.

D. You Say You Want More? Okay, Here's a Backward Elimination Stepwise Version

''Stepwise'' discriminant analysis is characterized by attempts to build more parsimonious, and perhaps more accurate, models than would be available from the model we have just investigated using all 12 predictor variables. One critically important distinction to draw with least squares regression methods is the fact that adding more variables to a discriminant function does not necessarily produce better results, defined as higher percentages of correct classifications or lower error rates for either or both outcomes. In contrast, ordinary least squares regression models always experience an increase in the coefficient of determination (R^2) whenever a new predictor variable is added—unless the new variable has a correlation equal to zero with the dependent variable—because doing so helps to minimize further the least-squares sum of squared errors function, which is reduced because the new information gets the observations on average closer to the regression hyperplane. Not necessarily so in discriminant analysis; in fact, adding new explanatory variables to a discriminant function well may *reduce* the percentage

TABLE 18 Structure for Backward Elimination Discriminant Analysis of Population Change in SMSAs

Stepwise discriminant analysis

150 Observations	12 Variable(s) in the analysis
2 Class levels	0 Variable(s) will be included

The method for selecting variables will be: BACKWARD
Significance level to stay = 0.1500

Class level information

CHANGE	Frequency	Weight	Proportion
0	74	74.0000	0.493333
1	76	76.0000	0.506667

of cases that are classified correctly, because the new variables may contain ''misleading'' information that leads the discriminant function to believe (falsely) that some observations share characteristics with observations in the opposite category. All the more reason to do what we can to minimize the number of variables that are to be included in the discriminant function, because fewer predictor variables in fact may produce a better fit to the data.

Table 18 presents basic information about the structure of this analysis. The method chosen to select variables to be retained in this process of variable reduction (and complexity reduction) is backward elimination. We assume that the variables within each class are multivariate normal with a common covariance matrix. Backward elimination begins with all predictor variables included in the model; then, at each successive step the variable contributing the least to the ability of the function to discriminate accurately (as measured by Wilks' lambda) is eliminated. This process continues until all remaining variables meet the criterion to stay in the model (which is set here to the default value of .1500).

The steps followed by this backward elimination discriminant analysis are shown in Table 19. Throughout these steps, the following points are important to know.

Tolerance represents one minus the squared multiple correlation of each variable as it enters the model with the other variables already in the model. For a variable already in the model, tolerance is one minus the squared multiple correlation of that variable with the entering variables and with the other variables already in the model. A threshold level can be established by default, or chosen by the researcher. Tolerance values are computed from the correlation matrix for the total sample (i.e., the pooled within-class correlation matrix).

Wilks' Lambda is shown, with its associated approximate F-value and p-value computed after the variable in question at that step has been removed. Wilks' lambda is the likelihood ratio statistic for testing the null hypothesis that the means of the classes on the selected variables are equal in the population. For any two groups that are well-separated, and thus distinct from each other, lambda is close to zero.

Pillai's Trace also is presented, with its p-value and approximate F-statistic. This is a multivariate statistic for testing the null hypothesis that the means of the classes on the selected variables are equal in the population.

The Average Squared Canonical Correlation (ASCC) is Pillai's trace divided by the number of groups minus one (here, of course, since the number of groups less one equals one, the values of Pillai's trace and ASCC will be identical. The value of ASCC is close to one when all groups are separated clearly and if all or most directions in the discriminant space show good separation for at least two groups.

Note that part of the procedure at Step 1 is to provide a complete model, with partial

TABLE 19 Steps in the Backward Elimination Discriminant Analysis of Population Change in SMSAs

Stepwise discriminant analysis

Backward elimination: Step 0
All variables have been entered

Multivariate Statistics

Wilks' Lambda = 0.65831281	F(12, 137) = 5.926	Prob > F = 0.0001
Pillai's Trace = 0.341687	F(12, 137) = 5.926	Prob > F = 0.0001

Average squared canonical correlation = 0.34168719

Backward elimination: Step 1
Statistics for removal, DF = 1, 137

Variable	Partial R**2	F	Prob > F
AREA	0.0220	3.075	0.0817
POPSQMI	0.0497	7.169	0.0083
PCTBLK	0.0111	1.535	0.2174
PCTSPAN	0.0271	3.809	0.0530
MEDAGE	0.0100	1.386	0.2412
OV25HSG	0.0312	4.414	0.0375
PCTMLAB	0.0028	0.380	0.5387
PCTFLAB	0.0082	1.133	0.2891
PCTUNEM	0.0040	0.556	0.4573
PCTMAN	0.0035	0.487	0.4866
MFIN79	0.0491	7.067	0.0088
PCTFPOV	0.0032	0.445	0.5057

Variable PCTMLAB will be removed
The following variable(s) have been removed: PCTMLAB

Multivariate statistics

Wilks' Lambda = 0.66013848	F(11, 138) = 6.459	Prob > F = 0.0001
Pillai's Trace = 0.339862	F(11, 138) = 6.459	Prob > F = 0.0001

Average Squared Canonical Correlation = 0.33986152

Backward Elimination: Step 2
Statistics for removal, DF = 1, 138

Variable	Partial R**2	F	Prob > F
AREA	0.0269	3.819	0.0527
POPSQMI	0.0488	7.084	0.0087
PCTBLK	0.0109	1.525	0.2189
PCTSPAN	0.0271	3.845	0.0519
MEDAGE	0.0073	1.013	0.3160
OV25HSG	0.0285	4.054	0.0460
PCTFLAB	0.0130	1.824	0.1790
PCTUNEM	0.0043	0.603	0.4389
PCTMAN	0.0035	0.479	0.4900
MFIN79	0.0464	6.722	0.0106
PCTFPOV	0.0020	0.272	0.6031

Variable PCTFPOV will be removed

Backward elimination: Step 2
The following variable(s) have been removed: PCTMLAB PCTFPOV

(*Table continues*)

TABLE 19 Continued

Multivariate statistics

Wilks' Lambda = 0.66143748 F(10, 139) = 7.115 Prob > F = 0.0001
Pillai's Trace = 0.338563 F(10, 139) = 7.115 Prob > F = 0.0001
Average squared canonical correlation = 0.33856252

Backward elimination: Step 3
Statistics for removal, DF = 1, 139

Variable	Partial R**2	F	Prob > F
AREA	0.0252	3.588	0.0603
POPSQMI	0.0472	6.891	0.0096
PCTBLK	0.0330	4.749	0.0310
PCTSPAN	0.0385	5.564	0.0197
MEDAGE	0.0054	0.759	0.3851
OV25HSG	0.0275	3.935	0.0493
PCTFLAB	0.0115	1.617	0.2056
PCTUNEM	0.0040	0.553	0.4583
PCTMAN	0.0052	0.732	0.3936
MFIN79	0.0817	12.365	0.0006

Variable PCTUNEM will be removed
The following variable(s) have been removed: PCTMLAB PCTUNEM PCTFPOV

Multivariate statistics

Wilks' lambda = 0.66406907 F(9, 140) = 7.869 Prob > F = 0.0001
Pillai's trace = 0.335931 F(9, 140) = 7.869 Prob > F = 0.0001
Average squared canonical correlation = 0.33593093

Backward elimination: Step 4
Statistics for removal, DF = 1, 140

Variable	Partial R**2	F	Prob > F
AREA	0.0245	3.516	0.0629
POPSQMI	0.0494	7.283	0.0078
PCTBLK	0.0375	5.449	0.0210
PCTSPAN	0.0404	5.901	0.0164
MEDAGE	0.0086	1.217	0.2718
OV25HSG	0.0279	4.013	0.0471
PCTFLAB	0.0228	3.263	0.0730
PCTMAN	0.0058	0.816	0.3678
MFIN79	0.0790	12.016	0.0007

Variable PCTMAN will be removed
The following variable(s) have been removed: PCTMLAB PCTUNEM PCTMAN PCTFPOV

Multivariate statistics

Wilks' lambda = 0.66794092 F(8, 141) = 8.762 Prob > F = 0.0001
Pillai's trace = 0.332059 F(8, 141) = 8.762 Prob > F = 0.0001
Average squared canonical correlation = 0.33205908

Backward elimination: Step 5
Statistics for removal, DF = 1, 141

Variable	Partial R**2	F	Prob > F
AREA	0.0307	4.471	0.0362
POPSQMI	0.0528	7.855	0.0058
PCTBLK	0.0564	8.435	0.0043
PCTSPAN	0.0479	7.100	0.0086
MEDAGE	0.0102	1.459	0.2291

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 19 Continued

Variable	Partial R**2	F	Prob > F
OV25HSG	0.0691	10.471	0.0015
PCTFLAB	0.0238	3.442	0.0656
MFIN79	0.1179	18.840	0.0001

Variable MEDAGE will be removed
The following variable(s) have been removed: MEDAGE PCTMLAB PCTUNEM PCTMAN PCTFPOV

Backward elimination: Step 5

Multivariate statistics

Wilks' lambda = 0.67485335	F(7, 142) = 9.774	Prob > F = 0.0001
Pillai's trace = 0.325147	F(7, 142) = 9.774	Prob > F = 0.0001

Average squared canonical correlation = 0.32514665
Backward elimination: Step 6
Statistics for removal, DF = 1, 142

Variable	Partial R**2	F	Prob > F
AREA	0.0302	4.417	0.0373
POPSQMI	0.0455	6.774	0.0102
PCTBLK	0.0483	7.200	0.0082
PCTSPAN	0.0434	6.435	0.0123
OV25HSG	0.0609	9.214	0.0029
PCTFLAB	0.0154	2.217	0.1387
MFIN79	0.1088	17.344	0.0001

No variables can be removed; No further steps are possible

F-ratios and their associated p-values. Among the predictor variables, PCTMLAB has the highest p-value (0.5387), and hence is deleted because it is the worst discriminator of the 12 that were entered originally. This process continues through the elimination of five predictors, until no other classifier variable meets the criteria for elimination.

Table 20 provides further perspective on the backward elimination model-simplifying process. It can be seen, from, for example, the increasing values of Wilks' Lambda (which indicate progressively poorer separation between the two outcome groups), that eliminating each variable in turn produces a small loss of model adequacy. However, such losses are offset by the greater model efficiency attained by the process of elimination.

V. CONCLUSION

This chapter has only scratched the proverbial surface of the analysis that can be conducted on dichotomous dependent variables. A large family of nonparametric correlational statistics for crosstabulation tables have been glossed over. More general classes of loglinear models also could be applied to such research problems. LISREL-type structural equation models also have not been covered here, largely because of a host of complexities that make such models beyond our ability to address adequately without the development of much more elaborate statistical and mathematical machinery. The interested reader is invited to investigate these and other related methods for dealing with dichotomous dependent variables in sources such as Neter et al., 1996; Sharma, 1996; Johnson and Wichern, 1991; Flury and Riedwyl, 1988; Agresti, 1990.

TABLE 20 Summary of the Backward Elimination Discriminant Analysis of Population Change in SMSAs

Backward elimination: Summary

Step	Variable removed	Number in	Partial R**2	F statistic	Prob > F	Wilks' lambda	Prob < lambda
0		12	—	—	—	0.65831281	0.0001
1	PCTMLAB	11	0.0028	0.380	0.5387	0.66013848	0.0001
2	PCTFPOV	10	0.0020	0.272	0.6031	0.66143748	0.0001
3	PCTUNEM	9	0.0040	0.553	0.4583	0.66406907	0.0001
4	PCTMAN	8	0.0058	0.816	0.3678	0.66794092	0.0001
5	MEDAGE	7	0.0102	1.459	0.2291	0.67485335	0.0001

Step	Variable removed	Number in	Average squared canonical correlation	Prob > ASCC
0		12	0.34168719	0.0001
1	PCTMLAB	11	0.33986152	0.0001
2	PCTFPOV	10	0.33856252	0.0001
3	PCTUNEM	9	0.33593093	0.0001
4	PCTMAN	8	0.33205908	0.0001
5	MEDAGE	7	0.32514665	0.0001

APPENDIX: SOME TECHNICAL DETAILS ABOUT LOGISTIC REGRESSION

The First Two Problems With Binary Dependent Variable Linear Regression Models

1. The error terms can't be distributed normally. This happens because each error term

$$\epsilon_i = Y_i - (\beta_0 + \beta X_i)$$

can assume only two possible values, equal to either

$$\epsilon_i = 1 - \beta_0 - \beta_1 X_i$$

when $Y_i = 1$, or

$$\epsilon_i = -\beta_0 - \beta_1 X_i$$

when $Y_i = 0$. Consequently, the assumption of normally distributed errors cannot be appropriate.

2. The error terms do not have equal variances when the response variable is a 0,1 indicator variable. This happens because the variance of Y_i is

$$\sigma^2(\epsilon_i) = \sigma^2(Y_i) = E\{(Y_i - E(Y_i)]^2\} = E\{Y_i\}(1 - E\{Y_i\})$$
$$= (\beta_0 + B_1 X_i)(1 - \beta_0 - \beta_1 X_i),$$

which means that $\sigma^2(\epsilon_i)$ depends on the value of X_i, and consequently that the error variances are different for different levels of X. As a result, ordinary least squares is no longer optimal.

Response Functions for Binary Dependent Variables

The logistic response functions that trace out the patterns shown in Figure 1 have the general form

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

which also can be expressed as

$$E(Y) = [1 + \exp(-\beta_0 - \beta_1 X)]^{-1}.$$

The fact that this logistic representation is not completely disconnected from linear regression logic is demonstrated by the ability to transform the logistic response function back into linear form. This is done rather easily by performing a logit transformation of E(Y) as the logarithm of the ratio of the probability of a success (defined here as π) and the probability of a failure $(1 - \pi)$:

$$\pi' = \log_e \left[\frac{\pi}{1 - \pi} \right]$$

which becomes

$$\pi' = \beta_0 + \beta_1 X.$$

The ratio of probabilities, $\pi/(1 - \pi)$, is known as the odds ratio; the transformed response function, $\pi' = \beta_0 + \beta_1 X$, is called the logit response function, or the logarithm of the odds ratio; and the value of π' is referred to as the logit mean response, which can vary from negative infinity to positive infinity as X varies over the same range.

The Simple Logistic Regression Model

When the response variable takes on values of only 1 (with probability π) and 0 (with probability $1 - \pi$), the simple logistic regression model takes the form

$$Y_i = E(Y_i) + \epsilon_i$$

where the error term ϵ_i follows the binomial (Bernoulli) distribution of Y_i with expected values $E(Y_i) = \pi_i$. The simple logistic model can be reexpressed more usefully as

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

where the observed values of X are assumed to be known constants. For X random, the values of $E(Y_i)$ become conditional means given the value of X_i.

The likelihood function of the parameters to be estimated in the logistic regression model, given the sample observations, is expressed as

$$\log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log_e[1 + \exp(\beta_0 + \beta_1 X_i)]$$

where $\log_e L(\beta_0, \beta_1)$ is the logarithm of the likelihood function (or the log-likelihood function). The maximum likelihood estimates of β_0 and β_1 are the values of those parameters that maximize the log-likelihood function, which must be found by computer algorithms using search procedures that converge on the estimated values. After these values have been found, they are substituted into the response function to generate the fitted, or estimated, logistic response function

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

The fitted logit response function, using the logit transformation, then can be expressed as

$$\hat{\pi}' = b_0 + b_1X = \log_e\left[\frac{\pi}{1-\pi}\right]$$

The exact relationship in the fitted logistic function is based on the fact that a one-unit increase in the value of X is related to the product of $\exp(b_1)$ and the estimated odds,

$$\hat{\pi}/(1-\hat{\pi}).$$

The value of the fitted logit response function evaluated at the arbitrary level of $X = X_j$ is

$$\hat{\pi}'(X_j) = b_0 + b_1X_j,$$

and at the level of $X = X_j + 1$,

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1).$$

Thus, the change in the two fitted values when X increases by one unit is

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = [b_0 + b_1(X_j + 1)] - [b_0 + b_1X_j] = b_1.$$

As shown before, $\hat{\pi}'(X_j)$ is the logarithm of the estimated odds when $X = Xj$, which could be rewritten as $\log_e(\text{odds}_j)$; also, $\hat{\pi}'(X_j + 1)$ is the logarithm of the estimated odds when $X = (X_j + 1$, which similarly could be rewritten as $\log_e(\text{odds}_{j+1})$. The difference between the two fitted logit response values then becomes

$$[\log_e(\text{odds}_j) - \log_e(\text{odds}_{j+1})] = \log_e\frac{\text{odds}_{j+1}}{\text{odds}_j} = b_1.$$

By taking antilogs of each side of this statement, the estimated odds ratio simply equals $\exp(b_1)$:

$$\text{estimated odds ratio} = \frac{\text{odds}_{j+1}}{\text{odds}_j} = \exp(b_1).$$

So, our interpretation of the effect on Y of a unit change in X will need to be expressed in terms of the exponentiated value of b_1, which translates into the proportional relative percentage change in Y in response to a change of one unit in X.

Multiple Logistic Regression

The multiple logistic regression model can be written, for a single observation, as

$$E(Y_i) = \beta_0 + \beta_1X_{i1} + \ldots + \beta_{p-1}X_{i,p-1},$$

In turn, this model can be written more compactly in matrix form as

$$E(Y_i) = \boldsymbol{\beta}'\mathbf{X}_i,$$

where

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \text{ and } \underset{p\times 1}{\mathbf{X_i}} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix}$$

Then, the multiple logistic regression model is:

$$E(Y_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{X_i})}{1 + \exp(\boldsymbol{\beta}'\mathbf{X_i})} = [1 + \exp(-\boldsymbol{\beta}'\mathbf{X_i})]^{-1}$$

and the logit transformation

$$\pi_i' = \log_e\left[\frac{\pi_i}{1 - \pi_i}\right]$$

produces the logit response function

$$\pi_i' = \boldsymbol{\beta}'\mathbf{X_i}.$$

Maximum likelihood methods are used to estimate the parameters of the multiple logistic response function, $\boldsymbol{\beta}$, by maximizing the log-likelihood function

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\mathbf{X_i}) - \sum_{i=1}^{n} \log_e[1 + \exp(\boldsymbol{\beta}'\mathbf{X_i})]$$

and finding the estimates

$$\underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

The fitted multiple logistic response function is

$$\hat{\pi} = \frac{\exp(\mathbf{b}'\mathbf{X})}{1 - \exp(\mathbf{b}'\mathbf{X})} = [1 + \exp(-\mathbf{b}'\mathbf{X})]^{-1}$$

where $\mathbf{b}'\mathbf{X} = b_0 + b_1X_1 + \ldots + b_{p-1}X_{p-1}$, and the fitted values become

$$\hat{\pi}_i = \frac{\exp(\mathbf{b}'\mathbf{X_i})}{1 + \exp(\mathbf{b}'\mathbf{X_i})} = [1 + \exp(-\mathbf{b}'\mathbf{X_i})]^{-1}$$

where $\mathbf{b}'\mathbf{X_i} = b_0 + b_1X_{i1} + \ldots + b_{p-1}X_{i,p-1}$.

Where Do the Results in Table 7 Come From?

The AIC, SC, and −2 LOG L statistics shown in Table 7 all evaluate the fit of the model for which y_j is the response value of the jth observation and estimates ($\hat{\pi}_j$) of $\pi_j = P(Y_j = y_j)$ are obtained by substituting into the model equation the maximum likehood estimates of the regression coefficients.

The score statistic operates off of the vector, $U(\boldsymbol{\gamma})$, of partial derivatives of the log likeli-

hood with respect to the vector of parameters, $\boldsymbol{\gamma}$, with dimension r. Denoting the matrix of the negative second partial derivatives of the log likelihood with respect to $\boldsymbol{\gamma}$, under the null hypothesis that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, the score statistic

$$\mathbf{U}'(\boldsymbol{\gamma}_0)\mathrm{I}^{-1}(\boldsymbol{\gamma}_0)\mathrm{U}(\boldsymbol{\gamma}_0)$$

has asymptotically a χ^2 distribution with r degrees of freedom.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*, New York, NY: John Wiley and Sons.

Bamber, D. (1975). ''The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph,'' *Journal of Mathematical Psychology, 12*: 387–415.

Bishop, Y., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*, London, England: Chapman and Hall.

Garson, G. D. (1971). *Handbook of Political Science Methods*. Boston, MA: Holbrook Press, Inc.

Hanley, J. A., and B. J. McNeil (1982). ''The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve,'' *Radiology, 143*: 29–36.

Hosmer, D. W. and S. Lemeshow (1989). *Applied Logistic Regression*, New York, NY: John Wiley and Sons.

Johnson, N. and D. Wichern (1991). *Applied Multivariate Statistical Analysis*, Third edition, Englewood Cliffs, NJ: Prentice-Hall.

Koven, S. G. and M. C., Shelley, II. (1989). ''Public Policy Effects on Net Urban Migration,'' *Policy Studies Journal, 17*(4): 705–718.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, Second edition, New York, NY: Chapman Hall.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models*, Fourth edition, Chicago, IL: Richard D. Irwin.

Parzen, E. (1962). ''On Estimation of a Probability Density Function and Mode,'' *Annals of Mathematical Statistics, 33*: 1065–1076.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edition, New York, NY: John Wiley & Sons, Inc.

Rosenblatt, M. (1956). ''Remarks on some Nonparametric Estimates of a Density Function,'' *Annals of Mathematical Statistics, 27*: 832–837.

SAS Institute, Inc. (1989). *SAS/STAT User's Guide, Version 6, Fourth edition, Volume 2*, Cary, NC: Author.

SAS Institute, Inc. (1995). *Logistic Regression Examples Using the SAS System, Version 6, First edition*, Cary, NC: Author.

Sharma, S. (1996). *Applied Multivariate Techniques*, New York, NY: John Wiley & Sons, Inc.

Shelley, M. C., II, and S. G. Koven (1993). ''Interstate Migration: A Test of Competing Interpretations,'' *Policy Studies Journal, 21*(2): 243–261.

Stokes, M. E., C. S. Davis, and G. G. Koch (1995). *Categorical data analysis using the SAS system*. Cary, NC: SAS Institute, Inc.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

19
Causal Modeling and Path Analysis

Evan M. Berman
University of Central Florida, Orlando, Florida

1. INTRODUCTION

Causal modeling enables researchers to examine relationships among variables that are complex, indirect and often multi-directional (i.e., with feedback loops). In comparison with multiple regression, causal modeling provides a more accurate and complex description of reality. It is uniquely applicable to many situations in social science, including those in public administration, that are characterized by complexity. Research that uses causal modeling is usually easily identified by its use of complex figures which illustrate many different relationships. Causal modeling is a methodological extension of regression analysis, but involves, among other things, some additional assumptions and estimation techniques.

Recent articles in public administration include several examples of causal modeling. For example, Berman and West discuss a causal model of municipal commitment to Total Quality Management (TQM) that examines complex relationships between TQM commitment, implementation strategies, organizational and HRM policies, and internal and external conditions (Berman and West, 1995). The utility of using causal modeling in this instance is that external stakeholder demands (e.g., client complaints) have negligible direct effects on doing TQM, but very significant indirect effects through other variables. If only multiple regression had been used as an analytic technique, the study would have wrongfully concluded that external stakeholder demands are irrelevant to implementing TQM. Other recent examples of causal modeling in public administration include an examination of factors affecting the effectiveness of the Senior Executive Service, and the adoption of local measures that limit development of hazardous areas (Perry and Miller, 1991; Burby and Dalton, 1994).

Causal modeling is theory-driven: it offers no panacea for theory-building. The complexity of relationships require that researchers provide some specification (i.e., a priori structure) of the relations that they examine. Many authors contend that the primary function of causal analysis is to test hypothesized relations (Lavee, 1988). Statistical software programs also require that researchers specify relationships among variables. Researchers do explore alternative specifications, but for the purpose of examining the robustness of different models. Different model specifications are also examined to identify those that are inconsistent with the data and, hence, implausible. Causal modeling is not an exploratory data activity, and researchers who use causal modeling in this manner must be mindful of violating the statistical properties of their tests through data mining and specification searches (Johnston, 1984).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

This chapter discusses the use of causal modeling in public administration research. The first section discusses the notion of causality in causal modeling. The second section discusses some heuristic, initial uses of causal modeling in case studies and meta-analysis. The third section examines the use of recursive models, i.e. models that do not contain feedback loops. Such models can be estimated with Ordinary Least Squares (OLS) techniques. The fourth section describes the application of non-recursive causal models which include feedback loops. An important development is the increased use of LISREL, a statistical software program that estimates models that contain feedback loops, as well as latent variables (i.e., unobserved constructs or factors that are indicated by observed variables). Finally, this chapter examines the outlook for using causal modeling in public administration research.

Although the potential for causal modeling applications is very broad in public administration, it is in fact one of the least often used techniques (Perry and Kraemer, 1994). Applications of causal modeling are more frequent in journals in other fields such as psychology, political science and sociology. This conclusion is consistent with other findings that public administration research is less quantitatively rigorous, and less oriented toward theory-testing, than research in other academic and practitioner-focused fields (Houston and Delevan, 1994). In this regard, causal modeling is mentioned as an approach that reflects methodological sophistication. It should be noted that although causal modeling is not widely used, the quality of causal modeling applications in public administration research is often as rigorous as that which is found in other fields.

Historically, causal analysis as a statistical tool was developed in various disciplines almost simultaneously during the late 1950s. Simon (1957) and Blalock (1957) developed recursive path analysis in sociology during the late 1950s and early 1960s (Asher, 1983). These models are estimable using OLS techniques, but produce biased estimates for nonrecursive models, and cannot be used for estimating these models (see further). The solution to this problem was advanced by econometricians who used instrumental variables and two-stage least squares solutions in the late 1960s to estimate supply and demand models (Johnston, 1972; Theil, 1971). Thus, recursive and nonrecursive models were increasingly used in the 1970s. Simultaneous advances in factor analysis during the 1960s, especially in psychometrics, were incorporated in 1972 in a program that estimated nonrecursive models with latent variables. This program, called LISREL (Linear Structural Relations), became popularized after significant improvements in 1981. Thus, techniques for doing causal modeling have been available for 15 to 30 years. They are disseminated through SPSS and other statistical software packages.

Finally, it should be noted that, as an extension of multiple regression, this chapter assumes basic understanding of the theory, practice and assumptions of multiple regression. Readers who require additional reading on multiple regression should consult chapter 15 of this Handbook. Increasingly, as in the case of LISREL, causal modeling also incorporates advances in factor analysis. Chapter 6 provides a discussion of that subject.

II. CAUSALITY

In recent years, a mainstream view of causality has come into existence in social science (McClendon, 1994; Goodwin, 1988; Babbie, 1994; Mulaik, 1987). This view takes into account a broad range of different perspectives, and suggests the following standards for the practice of causal modeling: (1) models and paths (i.e., relationships among variables) should be theory-based; (2) models should include all relevant paths; (3) assumptions and conditions of models should be clearly stated and tested; (4) a range of control variables should be identified and used; (5) the robustness of initial findings should be tested; (6) researchers should examine

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

temporal assumptions about the data (what-if analysis); and (7) when subjectivity of measures may yield biased measures, corroborating measures should be sought. These standards are discussed below. In addition, causal models must satisfy other standards, such as assumptions that are part of statistical estimation techniques, the reliability and validity of data, and strategies regarding research design, which are discussed in subsequent sections. The above standards apply only to the aspect of modeling.

Causal models allow the researcher to deal with complex relations among variables, and thus offer a more accurate portrayal of reality. However, causal models are not intended to emulate reality exactly. Causal models differ from reality in that they are based on the essential principles and variables that describe reality (Stokey and Zeckhauser, 1977). The selection of variables should be parsimonious but sufficient to obtain reasonably accurate descriptions and predictions of reality. In doing so, modeling enables researchers to set aside variables and relationships that while perhaps interesting, are of little practical importance. In instances when there are many fundamental relationships and variables, the use of causal modeling also allows researchers to focus piecemeal on these different relationships, providing a better understanding of the parts as well as the whole.

The utility of any model is partially determined by how well it conforms with reality (see also Chapter 21). However, empirical models cannot describe reality for values that variables do not assume, or for relationships and events that are not incorporated in models. All models make assumptions and have limitations and constraints. These aspects must be clearly communicated. Specifically, researchers should discuss paths (i.e., relationships) that are not addressed, control variables that are not included, and values of variables that are not assumed. Furthermore, they should also test the robustness of their findings by examining the model under different assumptions and constraints. Usually, there are several models that are equally theoretically plausible and compatible with the data. The range of such models should also be assessed.

Although the parsimony of variables and the consistency of findings are important standards in assessing causal models, evidence of causality involves a higher standard. According to the above authors, causality requires (1) temporal sequence, (2) covariance, and (3) nonspuriousness. The temporal condition requires that causes precede effects, and that causal mechanisms are specified. According to some, causes and effects cannot occur simultaneously. Nonexperimental data may satisfy this condition when survey items ask about prior conditions or by reasonably assuming that existing conditions existed in the recent past. For example, a city's current form of government can be said to temporally affect the level of professionalism, by assuming that the present form of government also existed in the recent past. However, causal mechanisms must be spelled out, because temporal sequence does not imply causation: for example, although night follows day, it is does not follow that night is caused by day. A theory is also needed of how, why and under what conditions variables cause each other. In the above example, a theory is needed to consider how form of government causes professionalism.

The second condition, covariance, means that if X causes Y, then a change in X should produce some predictable response in Y. That is, the two variables should correlate, i.e., covary. Covariance alone does not imply causality, but the opposite is true: causality implies covariance. This standard is straightforward in causal modeling, except that concerns are sometimes raised about the objectivity of observations. Specifically, researchers and informants (e.g., survey respondents) may cloud their observations by personal biases and experiences. For example, the level of trust in the workplace is apt to be differently assessed by supervisors and employees. To deal with the problem of skewed observations, social science relies on the standard of inter-subjectivity (or inter-rater agreement). The standard of inter-subjective agreement implies that observations should be corroborated by other, independent observations. Such external validation or ''criterion'' validity is a requirement for causality. Preferably, these other observations

should be based on ''hard'' (i.e., objective) data that is less likely to be affected by observer biases.

Finally, nonspuriousness means that observed relationships are not caused by some other variable. In experimental designs problems of intervening variables are resolved through randomization. This is not the case in nonexperimental research: the causal model must consider the full range of plausible control variables. There is increasing concern that the multiple regression assumption may be invalid that random distribution of the error term indicates a net zero effect of control variables that are not included in the model (Clogg and Haritou, 1993). Examination of the error term reveals the presence of problems such as heteroscedasticity, correlated error terms (time series), misspecification, etc., but the absence of this evidence does not imply that the full range of control variables is considered. The problem of missing control variables is much more serious in causal modeling than in regression analysis, because causal modeling involves more relationships. This increases the need for planning in causal modeling to ensure that adequate data are gathered concerning these variables.

III. INITIAL USES OF CAUSAL MODELING

Although the focus of this chapter is on quantitative applications of causal modeling, many researchers use causal models for theory-building and for setting the stage for subsequent data analysis that does not involve quantitative causal modeling techniques. The use of causal modeling in this manner is called initial or preliminary. Initial applications assist sound theory-building by focusing on key variables and clarifying relationships in complex situations.

The literature includes many examples of causal modelling that is used for theory-building. For example, Wilson and Durant (1994) develop a general, causal model to evaluate the outcomes of Total Quality Management. They argue that prior approaches to TQM evaluation have been a-theoretical, narrowly focused, prescriptive, and do not account for influences on outcomes that are independent of the effort to install TQM. Their contingency-based model of TQM outcomes focuses on the role of intervening variables from the organization (e.g., culture) and environment, as well as the influence of different implementation styles. They also take a broad perspective of TQM outcomes as creating an organizational quality culture. The causal model illustrates a multitude of different relationships, shown in Figure 1.

It should be noted that the graphical representation shows broad, theoretical concepts (e.g., culture). It does not show the variables that constitute these concepts. Although the authors do discuss specific variables that operationalize their concepts, it is not possible to depict all of these variables in one model, nor is there necessarily any unique set of variables that operationalize such concepts as culture. Researchers are apt to operationalize theoretical concepts in different ways. Another aspect is that Wilson and Durant develop causal sub-models. They are especially interested in processes of creating a quality culture, and the importance of team building in TQM. To explore these areas, causal (sub-) models are developed of these processes. This application demonstrates well the use of causal modeling for clarifying complex situations.

Because their objective is to aid theory-development, these authors conclude their article with a set of testable hypotheses and directions for future research. Other studies use causal modeling of theory as a basis for subsequent applications that do not involve quantitative causal modeling techniques. For example, Kravchuk (1993) uses a causal model to portray a vicious cycle of economic decline in Connecticut: however, his empirical work is a case study of the role of gubernatorial leadership in implementing administrative reforms that aid economic recovery. Schneider (1993) uses causal diagrams to show the linkage between AFDC and Medic-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

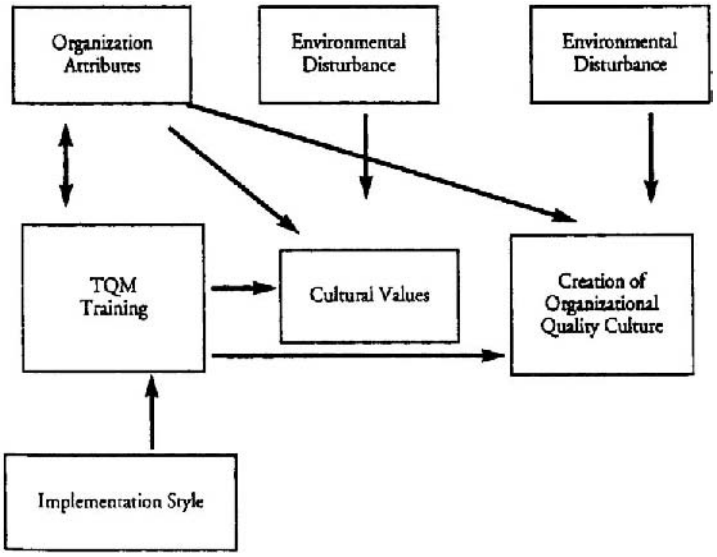


FIGURE 1 Generalized causal model adapted for TQM (adapted from Chen and Rossi, 1983).

aid, as well as influences on AFDC and Medicaid expenditures. Subsequent research uses multiple regression to discern the effects of these influences on AFDC and Medicaid expenditures. Similarly, Streib (1992) studies professionalism among directors of local government agencies and hypothesizes both direct and indirect effects of education and other variables on attitudes toward citizen participation in local governance. Further research uses multiple regression. Roberts (1995) develops a conceptual, causal model involving developmental performance appraisal, rater and ratee acceptance, and the effectiveness of information for decision-making. He, too, uses multiple regression for hypothesis testing. In many instances, these works might have used causal modeling to provide additional insights.

Initial uses of causal modeling are also found in meta-analysis. Robertson and Seneviratne (1995) compare studies of outcomes of planned organizational change to assess differences between the public and private sectors. A causal model is developed that encompasses the range of variables that are involved in these studies. It is theory-driven, based on a generalized understanding of planned change processes. Using this model, the authors organize and compare the results of 52 disparate studies using meta-analytic techniques. A similar approach is used by Hasenfeld and Brock (1991) to evaluate research in the field of social policy implementation. They develop a model of policy implementation with feedback loops that encompasses many variables of policy instruments, actors, driving forces, delivery systems and outputs. This model is a synthesis of diverse theories about implementation. Subsequently, they assess the extent that past research examines the variables mentioned in Figure 2.

Comparison with applications in leading journals of management, sociology and political science show very similar initial uses of causal modeling. Causal modeling is used to both simplify and depict complex theoretical relations, and also as a basis for further research. The models are theory-based. Some additional applications that were not found in public administration journals are causal models which are used as flow-charts, for example, showing the progression and stratification of research subjects over time (e.g., the chance of high school students dropping out) (Upchurch and McCarthy, 1990). Causal models can also show relationships

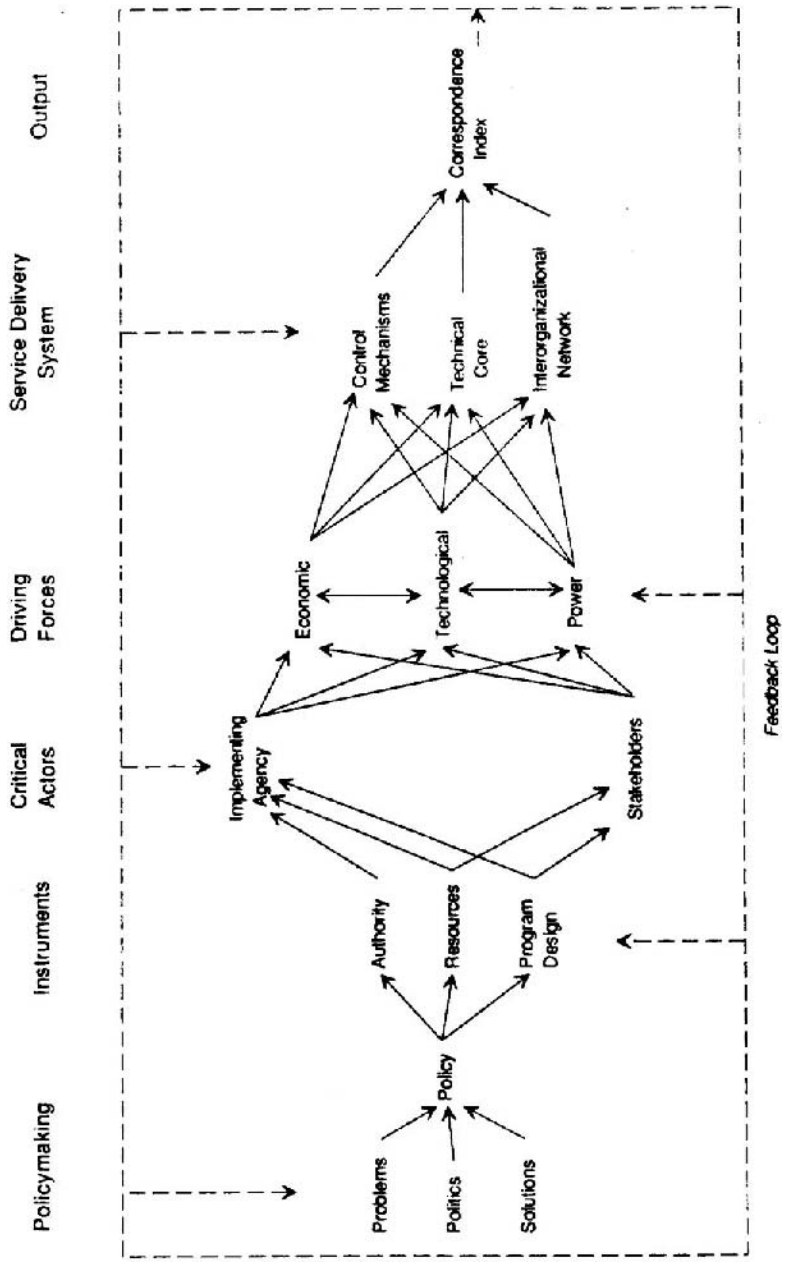


FIGURE 2 A political economy model of implementation.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

among different models, rather than relationships within single models (Kelley and Lewin, 1991). The following chapters focus on quantitative applications of causal modeling in data analysis, rather than for theory-building.

IV. RECURSIVE MODELS

Recursive models involve direct and indirect relationships among variables, but do not include feedback loops. An example is shown as Figure 3. The relevance of this restriction, which is defined in formal terms below, is that recursive models can be estimated with OLS techniques. This makes recursive models relatively easy to estimate.

The first step in causal modeling is theory-based specification of relationships among variables (i.e., paths). Figure 3 shows the hypothesized relations among four variables X_1 through X_4. The relationship between X_3 and X_4 is direct, the relationship between X_1 and X_4 is indirect (namely, it occurs through X_3), and the relationship between X_2 and X_4 is both direct and indirect. Note that there are no feedback loops. Causal modeling does not use the terminology of independent and dependent variables (as used in multiple regression), because it is often inconclusive: X_3 is both an ''independent'' variable (causing X_4), as well as well ''dependent'' variable (caused by X_1 and X_2). Rather, causal modeling distinguishes between exogenous (or predetermined) variables which are unaffected by other variables in the model (X_1 and X_2 and in Figure 3), and endogenous variables that are (at least somewhat) affected by other variables (i.e., X_3 and X_4). Lagged endogenous variables are considered to be exogenous variables when they influence other variables and are not affected by other variables in the model (Welch and Comer, 1988; Davis, 1985; Birnbaum, 1981).

The relationships among the variables in Figure 3 are estimated using the following regressions:

$$X_4 = a_1 + b_1X_3 + b_2X_2 + e_1 \quad (1)$$

$$X_3 = a_2 + b_3X_1 + b_4X_2 + e_2 \quad (2)$$

Each of these models is separately estimated using OLS. In the terminology of causal modeling, the regression coefficients are called structural coefficients, and the standardized regression coefficients (also known as beta coefficients) are called path coefficients. The effects of different variables are usually determined through path coefficients, because of their property of comparability. In this regard, the direct effect of X_2 on X_4 is defined as b_2, and the indirect effect is the product of effects along the path, hence, $b_1 * b_4$. The total effect is defined as the sum of direct and indirect effects. Thus, the effects on X_4 in Figure 3 are defined as:

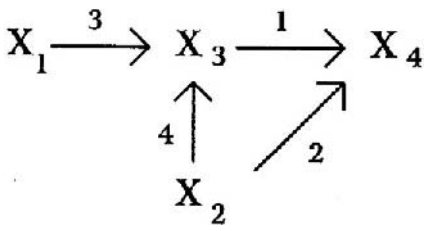


FIGURE 3 Simple path diagram.

	Effects		
	Direct	Indirect	Total
X_1	—	$b_3 * b_1$	$b_3 * b_1$
X_2	b_2	$b_1 * b_4$	$b_2 + b_1 * b_4$
X_3	b_1	—	b_1

For example, when b_2 is 0.44, b_1 is 0.38 and b_4 is 0.47, the total effect of X_2 on X_4 is [0.44 + 0.38 * 0.47 =] 0.62. Comparison of total effects across variables allows for making such statements as ''variable A has XX times more effect on variable B than variable C.'' Welch and Comer (1988) provide further examples of such calculations. Some researchers also calculate the difference between total effects and zero order effects (i.e., bivariate correlations), which are termed ''spurious'' effects (McClendon, 1994). It is customary to indicate the path coefficients along the paths, and $(1 - R^2)^{1/2}$ as the random variance affecting each endogenous variable. In the above Figure 3, it is assumed that all of the structural coefficients are statistically significant. This implies that reported models are often the product of considerable ''theory trimming,'' in which initial models are substantially modified. To indicate the ways in which the initial model has been modified, some researchers report their initial or preliminary causal model as a figure.

Extensions of the model in Figure 3 follow the same logic in terms of calculated direct and indirect effects: e.g., an additional variable X_5 that affects X_1 only (through new path 5) has an indirect effect on X_4 that is equal to $b_5 * b_3 * b_1$. This additional path does not affect any of the above calculated effects.

The condition that no feedback loops exist is stated formally that (1) no path passes through the same variable more than once; (2) no path goes backward against the direction of an arrow that it has gone through before; and (3) no path may pass through a double-headed arrow more than once (Wright, 1934). Double-headed arrows represent unanalyzed correlations between exogenous variables. No such unanalyzed relationships are indicated in Figure 3. An example of such a relationship would be the effect of X_3 on a new endogenous variable X_6 (through new path 6) that is not influenced by any other variable. In this instance, the relationship between X_4 and X_6 is unanalyzed and should be indicated through a double-headed arrow. These three conditions imply that the model is hierarchical, that is, that no feedback loops exist.

Regression estimations must satisfy standard OLS assumptions about the distribution of error terms. In addition, recursive models assume that each error term is uncorrelated with (1) *all* exogenous variables in the model (hence, of all regressions) and (2) with other error terms. It is assumed that e_1 and e_2 in Figure 3 do not violate these assumptions, either. Berry suggests that in practice these additional assumptions are not always upheld (Berry, 1984). For example, given that error terms represent variables not included in the model, and given the similarity of the subject matter among regressions, such variables are likely to be relevant to more than one regression. Thus, it is necessary to empirically assess these assumptions.

Recursive models are used in public administration research. Burby and Dalton (1994) examine the adoption of local measures that limit economic development in hazardous areas. Such measures aim to reduce the risk of natural disasters, and are an alternative to stricter building codes and efforts to reduce the hazards (e.g., flood control programs). Burby and Dalton develop a model in which the adoption of measures limiting development is affected by: (1) staff capacity, (2) demands from state planning mandates local political action, (3) the availability of resources, (4) local conditions of population density and the proportion of the population living

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

in hazardous areas and (5) the 'seriousness' of the problem as indicated by previous catastrophic and repetitive losses, and the demand for land in hazardous areas. These are direct effects.

Burby and Dalton also hypothesize that staff capacity is directly caused by items 2–5, above. These effects represent indirect effects on adoption of local measures (through the endogenous variable staff capacity). They also hypothesize that 1–5 cause the development of plan recommendations to limit economic development. This endogenous variable is not hypothesized to affect the adoption of measures that limit economic development, but their data analysis shows such effects exist. Thus, indirect effects also occur through the development of plan recommendations.

Their analysis shows that many of the hypothesized relations (and other relations that were not hypothesized through plan development) are statistically significant. Burby and Dalton do not separately report the indirect and direct effects on the adoption of limits. Their results show that 58% of total effects are accounted for by land use plan recommendations, local political action, the demand for land in hazardous areas, and state mandates. Each of these factors has about equal effect. Burby and Dalton also analyze their model separately for cities in states with and without state mandates. The theoretical basis for doing this is the hypothesis that causal adoption mechanisms differ in the presence of state mandates. Such separate estimation is similar to that in regression analysis, for example, when estimating the production of war and peacetime economies. War/peace cannot be treated as a control variable because the causal processes of production are very different during war and peace. Another example of recursive path analysis is provided by Bozeman and Loveless (1987), who examine environmental, organizational and other constraints on the productivity of research laboratories in the public and private sectors.

In a second application, Berman and West (1995) use path analysis to examine the strategies and conditions that are associated with municipal commitment to TQM. Their contingency-based theoretical model is similar to that of Durant and Wilson, discussed above. The Berman and West application differs from that of Burby and Dalton in that it uses multi-variate constructs of the concepts indicated by theory. By contrast, the Burby and Dalton model is based on single variables. The advantage of the Berman and West approach is that it may have greater content validity. For example, their measure of TQM commitment is based on an index of four constructs (the number of TQM applications in local government, and the range of training, resources and rewards provided in TQM efforts) that involve forty-two separate variables. They also use multi-variate constructs of transformational (11 variables), representational (9 variables) and transactional strategies (8 variables), as well as for HRM contributions to employee development and organizational culture (each 7 variables). Internal and external forces are also multi-variate constructs that are based on, respectively, 8 and 11 variables.

Their analysis of direct and indirect effects shows that whereas transformational and transactional strategies account for 64% of all direct effects, these strategies account for only 35% of total effects. The indirect effects of representational strategies, and combined effects of organizational policies and HR contributions to employee development are greater than the direct effects of transformational and transactional strategies. Interestingly, their analysis shows that there are no direct effects of internal and external driving forces on TQM commitment. Such effects occur indirectly through transformational, representational, and transactional strategies, hence, public administration matters (Figure 4).

Studies in fields other than public administration show very similar uses of recursive modeling. Examples include: links between citizen attitudes, policy incentives, and the siting of landfills (Bacot et al., 1994); factors that cause cities to adopt tax abatements (Varady, 1990); explanations for patterns in the residential proximity of Asians and blacks in residential neighborhoods (Fong, 1994); the effect of social integration and communication in top management

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

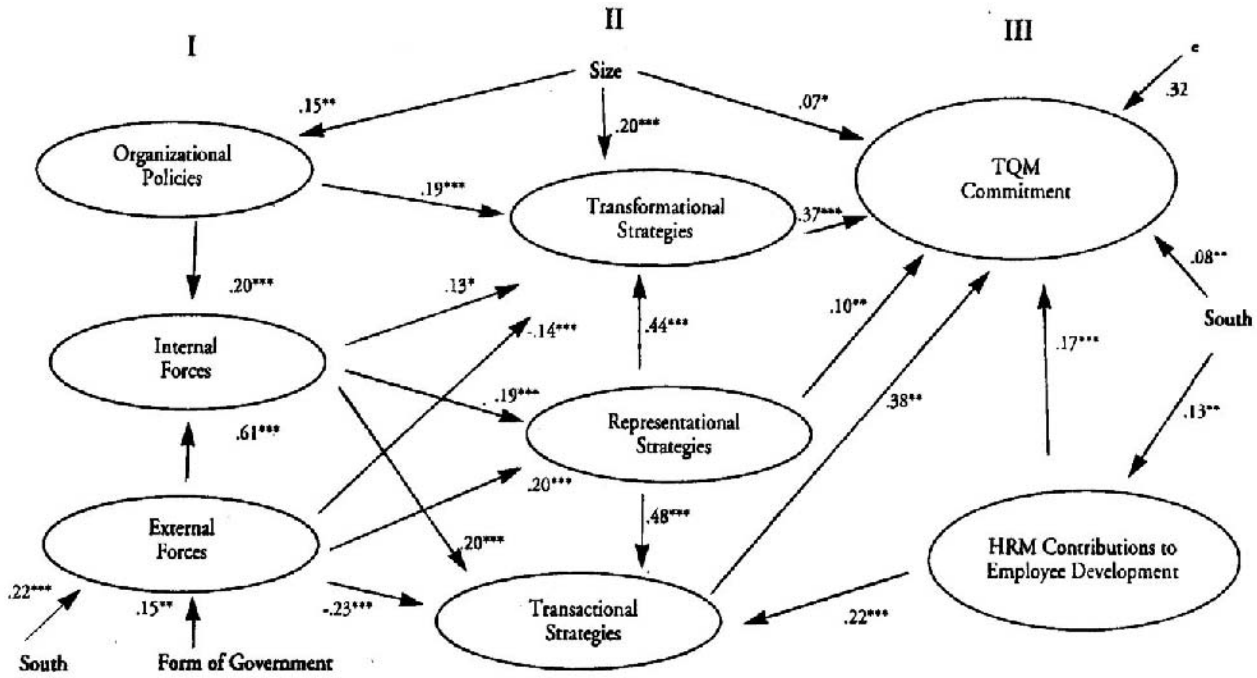


FIGURE 4 A path analysis of TQM commitment. ***1 percent significance; **5 percent significance; *10 percent significance [numbers are standardized regression coefficients(betas)].

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

teams (Smith et al., 1994); the relation between gender role attitudes, religiosity, dieting and bulimia in college women (Morgan et al., 1990); factors affecting individual and corporate dispute resolution (Lind et al., 1993); the determinants and outcomes of state legislative effectiveness (Weisert, 1991); forces affecting imprisonment rates across the states (Taggart and Winn, 1993); and an investigation of whether international dependency incites political violence in nations (Boswell and Dixon, 1990). These studies provide further examples of recursive modeling.

The above criteria (see ''Modeling Causality'') provide a useful matrix for evaluating efforts in this area. No formal, systematic assessment of recursive modeling in the social sciences exists, but some observations can be made about the quality of the above mentioned studies, or at least in their reporting. All studies develop theory-driven models, and the theory development (or justification) of initial models is extensive. Authors also discuss key control variables. However, the articles do not extensively discuss paths that are not hypothesized, but indicated by data. The impact of alternative path specifications is also seldom discussed, although some authors state that their findings are robust in this regard. The temporal condition is not often discussed, as are matters regarding the quality of data and limitations of models. Finally, it is not always evident that OLS assumptions have been carefully examined, nor the two additional OLS assumptions that have been stated above. Thus, there is room for improvement in reporting the results of causal models. Such improvements usually occur over time, as higher standards are achieved.

Although recursive models are easy to estimate using OLS, they are limited by the absence of feedback loops. Implications arising from feedback loops cannot be considered. Thus, a need exists for considering models with feedback loops (i.e., nonrecursive models), which are discussed below.

V. NONRECURSIVE MODELS

Nonrecursive causal models do not include restrictions about the ordering of effects among the endogenous variables. Models in which endogenous variables affect each other through feedback loops are called nonrecursive (or simultaneous equations). However, the ''cost'' of increased flexibility is heightened statistical sophistication. The following discussion starts with two problems that are common to all nonrecursive modeling: the need for (1) identification and (2) alternative estimation techniques. Overcoming these problems allows researchers to estimate nonrecursive models without latent variables. However, models with latent variables, which are also known as LISREL models, enable researchers to include theoretical constructs modeled by underlying (observed) variables. These models entail some additional concerns. Although dealing with identification and estimation issues are important to all non-recursive models, it should be noted that non-recursive models without latent variables are increasingly infrequent in journals of public administration and other social sciences. Examples of nonrecursive models without latent variables can be found in the older literature (Asher, 1983), as well as in economics for modeling macro- and micro-economic behavior.

A. The Identification Problem

The problem of identification is that nonrecursive models sometimes provide insufficient information to allow estimation. This situation is somewhat analogous to that in algebra in which models with more variables (unknowns) than equations do not have unique solutions. In the terminology of simultaneous equations, this is called under-identification. Likewise, when mod-

els have more equations than variables, several sets of algebraic solutions exist (assuming that no equations are linear combinations of others). Only when the number of variables and equations are equal, does a unique solution exist. These latter conditions are called, respectively, over-identification, and exact (or just) identification. The absence of feedback loops in recursive modeling is the constraint that ensures that such models are estimable.

Rules have been developed to determine the identification status of simultaneous equations. Briefly, the order condition states that in order for equations in a model of M equations to be identified, each equation must omit M-1 variables that appear elsewhere in the model. This is a necessary but not sufficient condition. The rank condition states that at least one nonzero determinant can be constructed of M-1 rows and columns, after omitting all columns of coefficients not having a zero entry in the equation in question, and omitting the row of coefficients of that equation (Asher, 1983). Each of these conditions is assessed separately for each equation in the model. Application of these rules is quite cumbersome, and does not provide much guidance when latent variables are involved.

Rather, researchers rely on statistical computer packages which provide messages warning researchers of under-identification. The condition of over-identification is resolved by using alternative estimation techniques (Two Stage Least Squares, see below) and does not present much of a practical problem. There are three basic strategies to obtain identification in under-identified models: (1) by imposing constraints on coefficients, (2) by adding new exogenous variables, and (3) by making assumptions about error terms. A common restriction is that certain parameters are zero, i.e., the elimination of paths from the model. In some instances, constraints in the form of linear combinations of variables are used. In sum, not all non-recursive models are estimable, and adaptations are often required. Researchers should discuss such strategies in their reporting.

B. The Estimation Problem

Nonrecursive models cannot be estimated with OLS because the endogenous variables are correlated with the error terms in the equations in which they appear as explanatory (i.e., independent) variables (Long, 1988). This violates the assumption that error terms are not correlated with independent variables. Using OLS to estimate the path coefficients in nonrecursive models produces biased and inconsistent estimates. The standard errors of other variables in equations will be smaller, causing researchers to wrongfully reject null hypotheses. Thus, alternative estimation techniques are called for. These techniques are instrumental variables (IV) approaches, and maximum likelihood (ML) methods.

Two Stage Least Squares (2SLS) is a widely used IV approach. Consider the following simultaneous equations:

$$X_1 = a_1 + b_1X_2 + b_2X_3 + e_1 \quad (3)$$

$$X_2 = a_2 + b_3X_1 + b_4X_4 + e_2 \quad (4)$$

In this model, X_1 and X_2 are mutually dependent endogenous variables, and X_3 and X_4 are exogenous variables. It can be shown that e_1 is correlated with X_2 in Equation 3, and that e_2 is correlated with X_1 in Equation 4. In 2SLS, the basic strategy is to estimate X_1 in Equation 4 as a function of all of the exogenous variables in the model. Hence:

$$X_1 = a_3 + b_5X_3 + b_6X_4 + e_3 \quad (5)$$

The estimates of b_5 and b_6 from Equation 5 are used to predict the values of a new variable, X_1'. X_1' is called an instrumental variable and is substituted for X_1 in [4]. Importantly, X_1' is

uncorrelated with e_2. Instrumental variables resemble their original variables when the number of exogenous variables in the model is large. In the second stage, OLS is used to estimate the parameters in the revised model Equation 4, in which X_1' is substituted for X_1. Hence:

$$X_2 = a_2 + b_3X_1' + b_4X_5 + e_4 \quad (6)$$

Subsequently, 2SLS is also used to obtain an instrumental variable of X_2, X_2', which is used to estimate X_1 in Equation 3. Although 2SLS estimates are biased, they are consistent which means that they will be close to the true parameter values in large samples. 2SLS estimates over-identified models by using all instruments simultaneously, thus providing a unique set of parameters. Berry provides a comparison of OLS and 2SLS estimates for nonrecursive models (Berry, 1984). He also discusses indirect least squares (ILS) IV approaches, but these are uncommon in nonrecursive models because they require exact identification.

Maximum likelihood (ML) is a family of estimation techniques that estimate predictors such that their value leads to estimates that best fit the observed sample observations. Referring to the above equations, limited information maximum likelihood (LIML) estimators try to find the best linear combinations of X_1 and X_2 (i.e., $c_1X_1 + c_2X_2$), called X* such that, when substituted as the dependent variable in Equation 3, (1) b_1 is significantly different from zero and (2) the excluded exogenous variable X_4 adds little or no explanatory power to the model. It can be shown that these restrictions imply that, in the above example, the linear combination of X_1 and X_2 should be chosen to minimize the ratio of the residual variance after regressing X* on X_3 over the residual variance after regressing X* on X_3 and X_4. The ratio is called lambda (Pindyck and Rubinfeld, 1981).

This least variance ratio, lambda, is applied to each regression in the model. When the equation is over-identified, lambda can be substantially greater than 1. Full information maximum likelihood (FIML) differs from LIML in that it applies the ML concept to the *entire* simultaneous equation system, rather than piecemeal equations. FIML provides more efficient estimators than LIML, but is mathematically quite complex. Likewise, 3SLS also uses information about the entire system, and is applied after 2SLS. 3SLS estimates usually have slightly smaller variances than those of 2SLS. Although 3SLS may be used, either 2SLS or a ML technique must be used. In practice, nonrecursive models without latent variables typically use 2SLS. LISREL applications (involving latent variables) usually use ML techniques for their final solutions. This is the LISREL default. Dwyer provides a general discussion of estimating simultaneous equation models (Dwyer, 1983).

C. LISREL Models

The popularity of LISREL modeling is based in the fact that it enables researchers to readily incorporate latent variables. Indeed, there are no applications in recent public administration journals of non-recursive models without latent variables. LISREL is a general model that involves (1) a set of structural equations and (2) confirmatory factor analysis (CFA). In CFA, researchers attempt to infer latent, unobserved factors from the underlying, measured (observed) variables, using the covariance structure of the latter. These latent variables are subsequently used to examine structural relationships among variables (Long, 1985). This latter activity distinguishes CFA from exploratory factor analysis. In LISREL terminology, a measurement model defines the relationships between observed variables and latent variables, and the structural model defines the relationships among the latent variables.

A challenge for many SAS and SPSS users is that LISREL models are specified through parameter matrices rather than structural equations or pull-down menus. Also, the LISREL language is unlike SAS and SPSS. SAS offers a LISREL alternative called PROC CALIS, but the

structural equations are specified in ways that are dissimilar from usual SAS syntax (e.g., in PROC GLM). Thus, LISREL requires learning new software language. The LISREL manual provides many detailed examples, with annotated examples of reporting. LISREL programs have control lines, such as ''DA'' for data, ''MO'' for model, etc., following which the user provides input. A useful feature in LISREL for Windows (version 8) are camera-ready outputs of path diagrams with coefficients, t-test statistics, etc. The following is a synopsis of essential LISREL concepts.

LISREL uses the exogenous/endogenous terminology to refer to latent variables, and the respective observed variables are called x- and y-variables. Latent exogenous variables are known as ksi-variables, and latent endogenous variables as eta-variables. Users must specify the following relationships among the x-, y-, ksi- and eta-variables, and error terms: (1) between x-variables and latent exogenous (ksi-) variables (called the lambda-x matrix); (2) between y-variables and latent endogenous (eta-) variables (the lambda-y matrix); (3) among latent endogenous variables (the beta matrix); and (4) from latent exogenous to latent endogenous variables (the gamma matrix). Other relationships involve associations (i.e., correlations that are not causal relationships) among (5) the latent exogenous variables (the phi matrix) and (6) the latent endogenous variables (the psi matrix), and vectors of (7) the error terms of x-variables (the theta-delta matrix), and (8) error terms of y-variables (the theta-epsilon matrix).

LISREL users specify their models by indicating (1) which matrices exist (hence, allowing for sub-models) and (2) which parameters are to be estimated. LISREL requires users to specify which parameters are free (to be estimated), fixed (assigned specific values, e.g., zero), and constrained (unknown, but equal to one or more parameters). The default values of the lambda-x, -y, and beta matrices are fixed, which means that users only have to specify those values that are to be estimated. The fixing and freeing of parameters is shown in the following two important examples. First, because latent variables are unobserved, it is necessary to assign measurement scales to these variables. This is usually done fixing one value (usually a one) in each of the columns of the lambda-x and lambda-y matrices. These columns refer to the latent variables. Failure to fix such values causes the model to be unestimable. Second, it is possible to enter x- and y-variables directly into the model by fixing their error term to zero and by specifying them as the only underlying variable of respective ksi- and eta-variables. Furthermore, unlike path analysis in recursive modeling, measurement errors can be assumed by setting the Cronbach coefficient of reliability to less than 1.00, for example, 0.85. This is accomplished by constraining the error terms (Joreskog and Sorbom, 1989).

The size of LISREL matrices is automatically defined by the number of x-, y-, ksi-, and eta-variables specified in the model. When x-, y-, ksi-, or eta-variables are not identified in the model, LISREL assumes that users are specifying a sub-model. When only x- and y-variables are defined, a recursive or nonrecursive model without latent variables is specified. When only x- and ksi-variables are present, or only y- and eta-variables, the user is specifying a factor analytic measurement model. When y- and eta-variables are specified, the gamma matrix allows for considering second and higher order (causal) effects among latent variables. Part of learning LISREL involves identifying which non-default parameter matrices are best used in these instances. For example, a path analysis in which most paths are hypothesized, is best accomplished by setting all beta parameters free, and specifying those that are fixed (i.e., not estimated). Another example is that CFA requires that the measurement scales are set through the phi- rather than lambda-matrices.

When the specified model is under-identified, LISREL provides a warning message of relationships that are not identified, or the statement that ''the information matrix is not positive definite.'' This matrix is generated by LISREL during the estimation process. To assist the identification effort, LISREL also generates ''modification indices'' as a user-specified output

options. A large, positive modification index indicates that a fixed parameter (i.e., relationship) will be identified if set free. LISREL will also estimate the value of the fixed parameter if set free.

LISREL also assesses the overall goodness of fit of the model. A chi-square test-statistic is calculated which compares the variance-covariance matrix of the data to that of the ML-estimated model. Insignificant chi-square values indicate close correspondence and, hence, good overall fit of the model. Insignificant chi-square values do not imply that the model is correct, only that it is consistent with the data. Nor does it imply that the model is unique or the best, as other models may also have insignificant chi-square values. This chi-square test-statistic is sensitive to departures from large sample sizes and the normality of observed variables. LISREL also produces a Root Mean Square Residual (RMSR) measure which can be used to compare the fit of different models, and an Adjusted Goodness of Fit measure which, for well-specified models, is usually between 0.94 and 1.00. Standardized residuals can also be used to help determine the source of lack of fit problems, as can modification matrices. T-tests are also provided for assessing the statistical significance of individual relations.

Nonrecursive modeling with latent variables requires that users proceed in the following order: (1) Define the measurement and structural models in theoretical terms, and draw the relationships. (2) Determine which LISREL (sub-)model represents the theoretical model. (3) Identify the hypothesized relationships as LISREL matrix parameters. (4) State the model in LISREL syntax (using examples from the LISREL manual). (5) Run (i.e., estimate) the model, and debug syntax errors. (6) When the model is under-identified, make appropriate changes in parameters. Re-run the model as necessary. (7) After the model is estimated, examine the Chi-square statistic, modification matrices, and other statistics to assess the goodness of fit. (8) Modify the model as necessary to ensure fit. (9) Report the results of the final model.

D. Applications

Perry and Miller discuss a LISREL model based on 1986 survey data collected by the Merit Systems Protection Board (Perry and Miller, 1991). The purpose of this analysis is to assess the effectiveness of the Senior Executive Service (SES), as perceived by SES employees. Data are analyzed from about 1700 surveys completed by SES employees. Survey items include employees' assessment of individual motivation, individual, agency and program performance, the effectiveness of rewards, accuracy of appraisals, public confidence, the role of career executives in policy-making, the enforcement of prohibited personnel practices, and so on. The model uses twenty-eight survey items which define twelve latent variables. Latent exogenous (i.e., ksi-) variables are identified with an asterisk (*) below.

Perry and Miller's initial model hypothesizes that program performance is affected by agency performance which, in turn, is affected by the role that SES employees play in policy-making,* as well as individual competence, performance, and motivation. These latter three individual measures are affected by the perceived effectiveness of performance rewards*, appraisal accuracy* and the quality of political executives* (i.e., SES' bosses). Finally, public confidence in their agencies is hypothesized to be related to agency performance, as well as the enforcement of prohibited personnel practices.* Three latent variables are based on single variables (political roles, program performance, and public confidence). The initial model also includes a variable, called ''rational deployment,'' that is not hypothesized to be associated with any other variable, but which is associated with individual performance in their final model. This variable is a single survey item regarding SES' assessment of efforts by agency head to rationally deploy senior executives based on their abilities.

The authors discuss that this initial model yields a very poor fit. The initial chi-square is

3079 (df = 323), which is statistically significant. The GFI is 0.83. They also report that nine parameters (i.e., relationships) have very large modification indices of over 100, and that four hypothesized paths are statistically insignificant. Based on these results, the initial model is modified by fixing and freeing some parameters. The final model has a chi-square of 955 (df = 307), a DFI of 0.95, and a maximum modification index of 10.83. The authors report the factor analysis of the relations between x- and ksi-variables, and y- and eta-variables. The standardized and unstandardized factor loadings are reported, as well as the R^2 (squared multiple correlation or SMC) of the x- and y-variables with respective latent variables. Each of these lambda parameters is significant at the 0.01 level.

Perry and Miller also report the significant coefficients of the gamma and beta matrices, which are also shown in a path diagram of the final model. Figure 5 only shows the relations among latent variables, but many other applications also show relations with x- and y-variables. Perry and Miller discuss: (1) the most important paths (based on the size standardized coefficients), (2) paths that are negative (but hypothesized as positive), (3) paths that were hypothesized but found to be insignificant, (4) and the most important effects on the variable of interest and agency performance. In this regard, their model shows that the strongest effect on agency performance is the enforcement of prohibited personnel practices. They also find that agency performance does not cause program performance, and that the latter is only, but strongly, influenced by performance rewards.

There are many examples of LISREL models in the literature, though not many in public administration. Some ''typical'' full-model applications include those that analyze complex relations between job satisfaction and life satisfaction (Judge and Watanabe, 1993); between health beliefs and preventive dental behavior (Chen and Land, 1986); between perceived group success and personal motivation (Riggs and Knight, 1994); between coordination, control and performance at the U.S. General Accounting Office (Gupta et al., 1994); between coping styles and deviant behavior among high school students (Kaplan and Peck, 1992); and between economic marginalization in the global economy and political conflict among developing countries (Moaddel, 1994). Some studies with different applications are those that use x- and y-variables that are index variables of survey items (Amato and Booth, 1995); that examine alternative model specifications using statistical reduction of Chi-square as a measure of improved model fit (Williams and Hazer, 1986); and those that use the CFA sub-model for scale validation (Houts and Kassab, 1990; Bagozzi et al., 1991).

Although no systematic assessment of LISREL modeling exists, some comments can be made about the above studies. In general, models are theory-driven, and conclusions are linked back to theory. Concerns about data quality of data are often more thoroughly addressed in LISREL studies than in recursive modeling, which may reflect the CFA aspect of LISREL modeling. However, assumptions about the normality of variables are seldom addressed, which are relevant to the validity of goodness of fit statistics. Few authors examine alternative model specifications. Also, some models use very few control variables, and theoretically obvious control variables are sometimes missing. None of the above studies used any external measures to ensure criterion-validity. A common practice in journals of psychology and many social sciences, but not in public administration, is the reporting of the correlation matrix, means and standard deviation, which enables readers to replicate results. By carefully assessing the reporting practices in the above articles, researchers can improve the quality of their work.

E. Comparison of LISREL and Path Analysis

LISREL models often suggest paths that are different or absent from those obtained through path analysis in nonrecursive modeling. This is the result of adjusting initial LISREL models

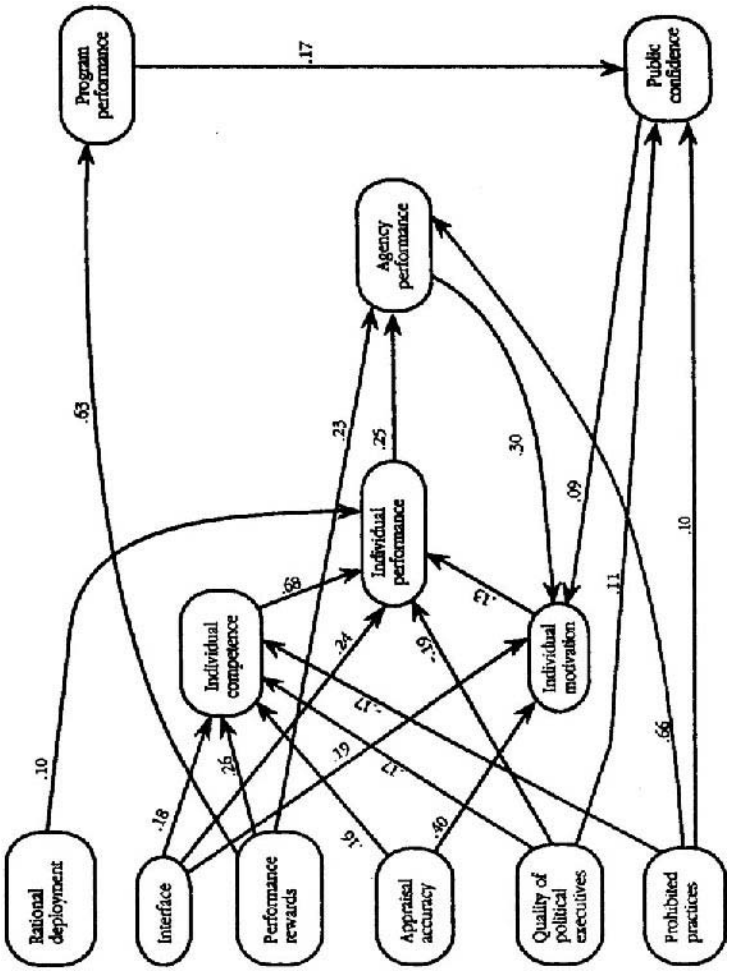


FIGURE 5 Revised structural model for SES program effects.

based on overall goodness of fit statistics, which are not available in path analysis. Although path analysis in recursive modeling, too, adjusts initial models, this is done on the basis of path coefficients, which is also done in final LISREL models. Gregson (1992) compares two path analysis and LISREL causal models. He finds that (1) the final path analysis models in both studies have unacceptable goodness of fit statistics; (2) acceptable LISREL models are only obtained when measurement errors (of x- and y-variables) are assumed; (3) such acceptable LISREL models do not include the final path analysis model; and (4) final LISREL models yield different conclusions because some relations are eliminated, added, or made stronger. The differences are nuances, which can be important.

Emerson and Van Buren (1992) also compare path analysis with LISREL, but in a single study. Their conclusions are not strictly comparable to Gregson, because they fail to report overall goodness of fit measures. They show two alternative LISREL models, one which uses measurement errors of endogenous variables, and another with latent variables, which yield similar (but not identical) results. The measurement errors (Cronbach alpha's) concern the study's endogenous variables, and these alpha's are empirically determined through respective predictor variables. The article is inconclusive about dissimilarities with the initial path analysis model, because of complications arising from improper operationalization of key variables. The LISREL analysis brought these to light. Emerson and Van Buren argue that LISREL provides a more accurate assessment of hypothesized relations, because it allows researchers to incorporate measurement errors. Gregson also argues that LISREL enables researchers to assess the overall goodness of fit.

VI. CONCLUSIONS

Causal modeling is an important technique that is used when relationships between variables are complex, indirect and multi-directional. This chapter provides criteria for assessing causal models, and it discusses initial, recursive, and nonrecursive applications. Initial applications are qualitative, and are used to represent relationships among theoretical concepts. Recursive and nonrecursive applications are quantitative data analysis strategies that estimate hypothesized paths (i.e., relations) among variables. When no feedback loops are present among variables in the model, nonrecursive methods can be used. When feedback loops are present, nonrecursive methods must be used. Causal models can also incorporate unobserved, latent variables, in which instance LISREL should be used. LISREL models can also be used to estimate recursive models, and offer some advantages in this regard.

This chapter has highlighted the need for sound theory and the use of appropriate methods. Causal modeling is theory-driven. Researchers specify and justify the paths that they are interested in. Although data analysis often suggests alternative paths, these too must be justified by theory. No computer program can generate models. Researchers must also use appropriate methods. Recursive models use estimation techniques that are straightforward extensions of multiple regression, but with some additional assumptions. Nonrecursive models require, at the very least, alternative estimation techniques and, in the case of LISREL, learning new statistical software.

Finally, regardless of the technique, researchers must meet the additional challenge of communicating their results with research colleagues, practitioners, and others who are often unfamiliar with causal modeling methods. Such audiences are often unable to interpret complex path diagrams and statistics. They may also harbor the incorrect idea that a given path diagram is the only correct model, rather than one among a family of plausible models. Thus, researchers who wish to use causal modeling must hone their communication (i.e., persuasion) as well as

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

research skills. To improve communication, researchers should ask the following questions about their reported results:

Is the research purpose straightforward and clearly stated? Are key hypotheses stated?
Are limitations of the data acknowledged? (for example, arising from the source or scope of data)
Is a straightforward and concise description of the causal method given? Are limitations acknowledged (e.g., multiple final models)?
Are rival hypotheses (i.e., control variables) clearly indicated? How are they incorporated in the model?
Are the results clearly displayed? Are the principal conclusions clearly stated? Is the reader assisted in interpreting the results?
Are these results consistent with previous studies, initial hypotheses, and alternative estimation methods?
What policy implications follow from the research findings? What further support exists for these implications?

REFERENCES

Amato, P.R. and A. Booth (1995). ''Changes in Gender Role Attitudes and Perceived Marital Quality,'' *American Sociological Review, 60*:58–66.

Asher, H. (1983). *Causal Modeling* (second edition), Sage, Newbury Park, CA.

Babbie, E. (1994). *The Practice of Social Research* (seventh edition), Wadsworth Publishing, Belmont, CA.

Bacot, H., T. Bowen, and M.R. Fitzgerald (1994). ''Managing the Solid Waste Crisis: Exploring the Link between Citizen Attitudes, Policy Incentives, and Siting Landfills,'' *Policy Studies Journal, 22*:229–244.

Bagozzi, R.P., Y. Yi, and L.W. Phillips (1991). ''Assessing Construct Validity in Organizational Research,'' *Administrative Science Quarterly, 36*:421–458.

Berman, E.M. and J.P. West (1995). ''Municipal Commitment to Total Quality Management: A Survey of Progress,'' *Public Administration Review, 55*:57–66.

Berry, W.D. (1984). *Non-recursive Causal Models*, Sage, Newbury Park, CA.

Birnbaum, I. (1981). *An Introduction to Causal Analysis in Sociology*, Macmillan Press, London, UK.

Blalock, H.M. (1957). *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill, NC.

Boswell, T. and W.J. Dixon (1990). ''Dependency and Rebellion: A Cross-National Analysis,'' *American Sociological Review, 55*:540–559.

Bozeman, B. and S. Loveless (1987). ''Sector Context and Performance: A Comparison of Industrial and Government Research Units,'' *Administration and Society, 19*:197–235.

Burby, R.J. and L.C. Dalton (1994). ''Plans Can Matter! The Role of Land Use Plans and State Planning Mandates in Limiting the Development of Hazardous Areas,'' *Public Administration Review, 54*: 229–238.

Chen, M.S. and K.C. Land (1986). ''Testing the Health Belief Model: LISREL Analysis of Alternative Models of Causal Relationships Between Health Beliefs and Preventive Dental Behavior,'' *Social Psychology Quarterly, 49*:45–60.

Clogg, C. and A. Haritou (1993). *The Regression Method of Causal Inference and a Dilemma with this Method*, Paper presented at the Conference on ''Causality in Crisis?,'' held at the University of Notre Dame, October 15–17.

Davis, J.A. (1985). *The Logic of Causal Order*, Sage, Beverly Hills, CA.

Dwyer, J.H. (1983). *Statistical Models in the Social and Behavioral Sciences*, Oxford University Press, New York, NY.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Emerson, M.O. and M.E. Van Buren (1992). ''Conceptualizing Public Attitudes toward the Welfare State: A Comment,'' *Social Forces, 71*:503–510.
Fong, E. (1994). ''Residential Proximity Among Racial Groups in U.S. and Canadian Neighborhoods,'' *Urban Affairs Quarterly, 30*:285–297.
Goodwin, D.D. (1988). Causal Modeling in Family Research, *Journal of Marriage and the Family, 50*: 137–147.
Gregson, T. (1992). ''The Advantages of LISREL for Accounting Researchers,'' *Accounting Horizons, 6*:42–48.
Gupta, P.P., M.W. Dirsmith, and T.J. Fogarty (1994). ''Coordination and Control in a Government Agency,'' *Administrative Science Quarterly, 39*:264–284.
Hasenfeld, Y. and T. Brock (1991). ''Implementation of Social Policy Revisited,'' *Administration and Society, 22*:451–479.
Houston, D.J. and S.M. Delevan (1994). ''A Comparative Assessment of Public Administration Journal Publications,'' *Administration and Society, 26*:252–271.
Houts, S.S. and C. Kassab (1990). ''Use of LISREL in Scale Validation,'' *Psychological Reports, 67*: 1059–1063.
Johnston, J. (1972). *Econometric Methods*, McGraw-Hill, New York, NY.
Johnston, J. (1984). *Econometric Methods*, third edition, McGraw-Hill, New York, NY, p. 501.
Joreskog, K.G. and D. Sorbom (1989). *LISREL 7: A Guide to Program and Applications*, SPSS Inc., Chicago, IL, p. 153.
Judge, T.A. and S. Watanabe (1993). ''Another Look at the Job Satisfaction-Life Satisfaction Relationship,'' *Journal of Applied Psychology, 78*:939–948.
Kaplan, H.B. and B.M. Peck (1992). ''Self-Rejection, Coping Style and Mode of Deviant Response,'' *Social Science Quarterly, 73*:903–919.
Kelley, H.H. and J. Lewin (1991). ''Situations and Interdependence,'' *Journal of Social Issues, 47*:211–233.
Kravchuk, R.S. (1993). ''The 'New Connecticut': Lowell Weicker and the Process of Administrative Reform,'' *Public Administration Review, 53*:329–339.
Lavee, Y. (1988). ''Linear Structural Relationships (LISREL) in Family Research,'' *Journal of Marriage and the Family, 50*:937–948.
Lind, E.A., C.T. Kulik, M. Ambrose, M.V. de Vera Park (1993). ''Individual and Corporate Dispute Resolution: Using Procedural Fairness as a Decision Heuristic,'' *Administrative Science Quarterly, 38*: 224–251.
Long, J.S. (1985). *Confirmatory Factor Analysis*, Sage, Beverly Hills, CA.
Long, J.S. (ed.) (1988). *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*, Sage, Newbury Park, CA.
McClendon, M.J. (1994). *Multiple Regression and Causal Analysis*, Peacock Press, New York.
Moaddel, M. (1994). ''Political Conflict in the World Economy: A Cross-National Analysis of Modernization and World-System Theories,'' *American Sociological Review, 59*:276–303.
Morgan, C.S., M. Afflick, and O. Solloway (1990). ''Gender Roles, Attitudes, Religiosity and Food Behavior: Dieting and Bulimia in College Women,'' *Social Science Quarterly, 71*:142–151.
Mulaik, S.A. (1987). ''Toward a Conception of Causality Applicable to Experimentation and Causal Modeling,'' *Child Development, 58*:18–32.
Perry, J.L. and K.L. Kraemer (1994). ''Research Methodology in the Public Administration Review 1975–1984,'' *Research in Public Administration* (J.D. White and G.B. Adams, eds.), Sage, Thousand Oaks, CA, pp. 93–109.
Perry, J.L. and T.K. Miller (1991). ''The Senior Executive Service: Is It Improving Managerial Performance?'' *Public Administration Review, 51*:554–563.
Pindyck, R. and D. Rubinfeld (1981). *Econometric Models and Economic Forecasts*, McGraw-Hill, New York, NY.
Riggs, M.L. and P.A. Knight (1994). ''The Impact of Perceived Group Success-Failure on Motivational Beliefs and Attitudes: A Causal Model,'' *Journal of Applied Psychology, 79*:755–766.
Roberts, G.E. (1995). ''Developmental Performance Appraisal in Municipal Government: An Antidote for a Deadly Disease?,'' *Review of Public Personnel Administration, 15*:17–43.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Robertson, P.J. and S.J. Seneviratne (1995). ''Outcomes of Planned Organizational Change in the Public Sector: A Meta-Analytic Comparison to the Private Sector,'' *Public Administration Review, 55*: 547–557.

Schneider, S. (1993). ''Examining the Relationship between Public Policies: AFDC and Medicaid,'' *Public Administration Review, 53*:368–380.

Simon, H. (1957). ''Spurious Correlations: A Causal Interpretation,'' *Models of Man*, John Wiley, New York, NY.

Smith, K.G., K.A. Smith, J.D. Olian, H.P. Sims, D.P. O'Bannon, and J.A. Scully (1994). ''Top Management Team Demography and Process: The Role of Social Integration and Communication,'' *Administrative Science Quarterly, 39*:412–438.

Stokey, E. and R. Zeckhauser (1977). *A Primer for Policy Analysis*, Norton and Co., New York, NY.

Streib, G. (1992). ''Professional Skill and Support for Democratic Principles,'' *Administration and Society, 24*:22–40.

Taggart, W.A. and R.S. Winn (1993). ''Imprisonment in the American States,'' *Social Science Quarterly, 74*:736–749.

Theil, H. (1971). *Principles of Econometrics*, McGraw-Hill, New York, NY.

Upchurch, D.M. and J. McCarthy (1990). ''The Timing of First Birth and High School Completion,'' *American Sociological Review, 55*:224–234.

Varady, D.P. (1990). ''Tax Abatements and Below Market Rate Mortgages To Attract Middle Income Families to the Central City: A Cincinnati Study,'' *Journal of Urban Affairs, 12*:59–74.

Weisert, C.S. (1991). ''Determinants and Outcomes of State Legislative Effectiveness,'' *Social Science Quarterly, 72*:797–806.

Welch, S. and J. Comer (1988). *Quantitative Methods for Public Administration* (second edition), Dorsey Press, Chicago, IL.

Williams, L.J. and J.T. Hazer (1986). ''Antecedents and Consequences of Satisfaction and Commitment in Turnover Models: A Re-analysis Using Latent Variable Structural Equation Methods,'' *Journal of Applied Psychology, 71*:219–231.

Wilson, L.A. and R.F. Durant (1994). ''Evaluating TQM: The Case for a Theory Driven Approach,'' *Public Administration Review, 54*:137–146.

Wright, S. (1934). ''The Method of Path Coefficients,'' *Annals of Mathematical Statistics, 5*:161–215.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

20
Economic Modeling

Ronald John Hy
University of Central Arkansas, Conway, Arkansas

I. INTRODUCTION

An important component of policymaking involves estimating what probably will happen if a policy change is implemented. Decisions, after all, are not made without an idea of what to expect. That is, decisions seldom are made without regard to estimated future consequences. Knowing the possible future effects of policy changes will affect an organization's decisions. Since many policy decisions involve fiscal matters, economic modeling is being used increasingly to estimate the future effects of policy changes.

One of the most striking developments in recent decades has been the increased emphasis on the use of statistical models to analyze the economic aspects of public policy issues. The value of these models is predicated not so much whether they hold true as it is in helping policy makers select satisfactory courses of action from among alternatives in order to increase the chances of attaining desired effects (Quade, 1982: p. 279).

A primary use of economic modeling, therefore, is directed toward improving fiscally based policy decisions. In addition, economic modeling helps identify limits beyond which further actions are not desirable and rates of progress are negligible.

Specifically, economic modeling:

Measures the impacts of changes in terms of volume of activity
Projects possible levels of activity as a bases for making decisions
Ascertains the effects of alternative policy changes

Some forms of economic modeling also measure the impact of policy changes on jobs, personal and corporate income, multipliers, as well as interindustry linkages. (Measuring the impact on jobs is exceedingly important politically, even though it often represents only a portion of the total impact of policy changes.)

The increasing complexities of policymaking coupled with its need to address a variety of economic and social problems have led to an expanded interest in economic modeling. As public funds begin to shrink and as the public becomes increasingly sensitive to the way their dollars are spent, policymakers are interested in learning more about the possible effects of changes before they occur.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

II. WHAT IS ECONOMIC MODELING?

Economic modeling uses quantitative techniques to untangle enormous amounts of available empirical economic data in order to describe and estimate the future impacts of policy changes before they occur, thus helping policymakers optimize their decisions. Economic modeling, then, is a mathematical system of sectoral allocation equations, each including several interdependent variables (Spyros and Wheelwright, 1973: p. 21). Relationships and impacts among variables are estimated by developing and computing a variety of rigorously defined equations.

Modeling, which combines data and theory, is a method of analysis that examines relatively stable patterns of the flow of goods and services among elements of an economy ''to bring a much more detailed statistical picture of the [economic] system into the range of manipulation by economic theory'' (Leontief, 1986: p. 4). The purpose of modeling is to reduce errors in estimates by bringing theory into closer association with data.

Economic theory constantly seeks to explain the operation of an economy in terms of interactions of key variables—namely, supply, demand, wages, and prices. This operation undoubtedly involves complex series of transactions where goods and services are exchanged by people. Consequently, there is a fundamental relationship between the volume of a sector's outputs and the amount of inputs needed to produce those outputs. The interdependence among sectors (industries) also is measured by a set of equations that express the balance between the total input and aggregate output of each good and service produced and used in a given time period (Leontief, 1986: p. 424).

The two of the most widely used economic models are predictive regression and input/output models. *Given the mathematical complexity of these models, it is virtually impossible to discuss their computations in a single chapter. Therefore, the principal objective of this chapter is to focus on comprehension rather than on computational knowledge*.

III. PREDICTIVE REGRESSION MODELS

Predictive regression models, which are reasonably useful when short-term forecasting is more important than is sectoral planning, are the simplest and most commonly used economic models. These types of models are designed to forecast outcomes based on specified activities.

Translated into its simplest terms, a predictive model is a set of definitions and assumptions that estimate the effect of specific activities (predictive variable(s)) on a public policy issue (dependent variable). These definitions and assumptions are expressed in mathematical terms which state that an impact is a function of activities. For instance, Keynes argued in part that consumption spending is a function of disposable income, or mathematically, Y (f) X where Y is consumption spending and X is disposable income. Simply put, if disposable income increases, consumption spending increases.

A predictive regression model, which is expressed as a set of equations, describes not only the way variables interact, but also the magnitude of that interaction. Such a model, for example would show not only how, but also how much a given tax cut would affect state spending. Consequently, a predictive model generates estimates of magnitude in addition to describing the relationships between and among variables.

The basic idea behind a predictive regression model is quite simple—if a patterned relationship exists between specified activities—predictor variable(s)—and a public policy issue—the dependent variable—and if that pattern continues with virtually no interruptions over a short

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 1 Types of Variations of Regression Models

	Simple regression	Multiple regression
Linear regression	One variable is used to estimate the dependent variable. Pattern of data fall along a straight plane.	More than one variable is used to estimate the dependent variable. Pattern of data fall along a straight plane.
Nonlinear regression	One variable is used to estimate the dependent variable. Pattern of data fall along a curved plane.	More than one variable is used to estimate the dependent variable. Pattern of data fall along a curved plane.

period of time, the predictor variables can be used to estimate the effect of activities on a policy issue.

A simple example illustrates this basic idea. Assume a predictive regression model is built stating that a state's sales tax (dependent variable) is affected by its personal income and employment (predictor variables). As the state's personal income and employment increases, the state's sales tax collections increase. A predictive model will show how much the state's income tax collections will increase in relation to changes in the state's personal income and employment.

A. Review of Regression Models

Essentially, there are two types of regression models—simple and multiple. A simple regression model uses a single predictor variable to estimate the direction and magnitude of a change on the dependent variable. Conversely, a multiple regression model uses two or more predictor variables to estimate the direction and magnitude of a change on the dependent variable.

Each of these two types of regression models, in turn, can use either a linear or nonlinear computational formula, depending on whether the pattern of the relationship is along a straight or curved line. When the pattern of the relationship between the predictor variables(s) and the dependent variables is a relatively straight, a linear regression model is appropriate. However, when the pattern of the relationship between the predictor variables(s) and the dependent variables is a curved, a nonlinear regression model is apropos. Table 1 illustrates the different types of predictive regression models. The appropriate regression model must be used in order to arrive at the most precise estimates—that is, the estimates with the narrowest range of estimate values.

A simple example describes this point. If a model estimates sales tax revenue for next year to be $1.7 billion, and the range of that estimate to be $0.7 billion, the range of the estimate is between $1 billion and $2.4 billion. On the other hand, if a model estimates sale tax revenue for next year at $1.2 billion and the range of the estimate to be $0.05 billion, the range of the estimate is between $1.15 billion and $1.25 billion, a much more precise estimate. The most appropriate regression model will be the one that yields the most precise estimate.

B. Determining the Most Appropriate Model

Unfortunately, there is no simple way to ascertain the pattern of the relationship between the predictor variables and the dependent variable, especially when more than one predictor variable

is used. One way, therefore, to determine the pattern of the relationship is to use the following modified trial and error method:

> *Calculate various linear and nonlinear regression models and select the one which yields the largest (strongest) coefficient of determination, the least amount of autocorrelation, and an acceptable probability of error*.

The *Coefficient of Determination* (R^2) is the fraction of variance explained by the model. (Since the coefficient of determination was discussed in Chapters 14 and 15, consult these chapters for further explanation.)

The three types of regression models that are generally used are linear, polynomial, and transformative. Since linear simple and multiple regression models are discussed in Chapters 14 and 15, the brief discussion presented below will focus on polynomial and transformative models. Listed below are the basic formulas for multiple linear, polynomial, and transformative predictive regression models. In practice, however, most predictive regression models are not linear.

Formulas for multiple, linear, polynomial, and transformative equations

Linear equation:

$$\ddot{Y} = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$$

Second degree polynomial equation:

$$\ddot{Y} = a + b_1X_1 + c_1X_1^2 + b_2X_2 + c_2X_2^2 + \ldots b_nX_n + c_nX_n^2$$

Third degree polynomial equation:

$$\ddot{Y} = a + b_1X_1 + c_1X_1^2 + d_1X_1^3 + b_2X_2 + c_2X_2^2\, d_2X_2^3 + \ldots b_nX_2n + c_nX_n^2\, d_nX_n^3$$

Transformative equations:

When data pattern turn slightly upward:

$$\sqrt{\ddot{Y}} = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$$

When data pattern turn upward more drastically:

$$\log \ddot{Y} = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$$

When data pattern turn slightly downward:

$$\ddot{Y} = a + b_1\sqrt{X_1} + b_2\sqrt{X_2} + \ldots b_n\sqrt{X_n}$$

When data pattern turn downward more drastically:

$$\ddot{Y} = a + b_1 \log X_1 + b_2 \log X_2 + \ldots b_n \log X_n$$

C. Polynomial Models

Polynomial models rely on each predictor variable's values and power(s) to estimate its impact on the dependent variable. Logically, this means that polynomials are used to bend the trend plane until it conforms to the nonlinear pattern of the data. A second degree polynomial bends the [trend] plane once; a third degree polynomial bend the plane twice; and so forth. There will always be one less bend than indicated by the degree of the polynomial equation. Solving a polynomial model is straightforward inasmuch as the procedure is identical to solving linear

models. The value of each predictor variable (X) and the power of the value for each predictor variable (e.g., X^2) are substituted into the model to estimate the value of the dependent variable.

D. Transformative Models

Often a transformative regression model is used with data whose patterns are nonlinear. Such a model is one in which either the predictor or dependent variables is transformed mathematically; and those products—rather than the original data, are used to compute a regression model. Such a model rearranges a nonlinear pattern of data in a way that a linear regression model can estimate the value of the dependent variable.

When a transformative regression model is used, data must be transformed before the model is computed, a relatively easy task with the aid of a computer. For instance, when using a transformative model which utilizes the square root of Y ($\sqrt{Y}$), the square root of the Y values is calculated and those transformed values instead of the actual values are used to compute the regression model. When a transformative model which alters the X value(s) is employed, identical mathematical functions are performed on the values of the predictor variable(s) ($\sqrt{Y}$ and log X).

IV. DIAGNOSTICS

The preceding section of this chapter focused on different types of predictive regression models, each depending on the number of predictor variables as well as the pattern of the relationship between the predictor variable(s) and the dependent variable. The utility of these models, however, depend on a number of regression-base assumptions that are part of regression models. The three primary assumptions that must be discerned and addressed when using predictive regression models are (1) autocorrelation, (2) multicollinearity, and (3) selection of appropriate variables.

A. Autocorrelation

Recall from Chapter 14 that a correlation between two variables describes statistically what happens to one variable (Y), if there is a change in the other variable (X). The degree of change is measured by a correlation coefficient, which varies between +1.00 and −1.00—with a coefficient of zero suggesting that no matter what happens to one variable (X), nothing much will happen to the other variable (Y). Autocorrelation is similar to a correlation coefficient except that it describes statistically a relationship (mutual dependence) among values of the same variable measured at different time intervals. Since predictive regression models rely on times series data, which are not made up of a set of randomly selected data points but rather of data points collected at periodic intervals, one of the most troublesome problems frequently encountered when using these models is autocorrelation.

Regression models assume that the residuals are independent of one another; that is, each data point does not have independent error terms. When the residuals are not independent, the model may treat the predictor variable as if it has been omitted. Rather than having the model explain the basic underlying patterns and let the residuals represent random errors, the model's residuals are included as part of the basic pattern (Wheelwright and Makridakis, 1973: p. 111). If this pattern is not eliminated, the regression model will not be so accurate as it would otherwise be.

With an autocorrelated series regression coefficients still will be unbiased, but many of

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

the associated statistics may be invalid. The standard errors of the regression coefficients as well as the estimated variance around the regression may be understated, and the t and F distributions may not be applicable.

Autocorrelation furnishes important information about the pattern of the data. When the autocorrelation is zero, the data points are completely random. When the autocorrelation is close to 1.00, the data points are seasonal or cyclical. Information gained from autocorrelation calculations can be utilized by many of the predictive regression models to arrive at optimal estimates.

B. Recognition

Standard diagnostics normally are displayed on a computer printout each time a model is fitted to the data. While many other statistics are available, the Durbin-Watson d statistic and the Ljing-Box Q statistic—along with the mean and standard deviation—are used most commonly to validate the estimates of a predictive regression model.

The *Durbin-Watson d statistic* checks for autocorrelation in the first lag of the residual errors. (Recall that autocorrelation describes the association—mutual dependence—among the values of the same variable at different time periods.) When autocorrelation exists, the residuals do not represent random error, as the model would suggest. Thus, the model, which is designed to describe the basic underlying pattern and estimate from that pattern, will not be so accurate as it would otherwise be. Testing for autocorrelation involves establishing a hypothesis stating that the first-lag autocorrelation is zero—that is, there is no autocorrelation (Stellwagen and Goodrich, 1997: p. 166).

The Durbin-Watson d statistic is based on the sum of the squared differences of the residuals. The following is a computational formula for the Durbin-Watson d statistic:

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e^2 t}$$

where

e_t = the residual of a given observation
e_{t-1} = the residual of the preceding observation

Because of certain statistical difficulties, the regions of certainty for rejection of the hypothesis is uncertain—unlike other significance tests presented in this book. Consequently, there is a region between two bounded limits where one can never be sure that autocorrelation exists. This is because of certain statistical difficulties.

After calculations are made, d is compared to the upper and lower bounds in a Durbin-Watson Test Bounds Table (found in most statistics books) for various significance levels, sample sizes, and independent variables. If d is below the lower bound, autocorrelation exists. If d exceeds the upper bound, autocorrelation does not exist. When d falls in the middle of the two bounds, autocorrelation may or may not exist. A significant limitation with the Durbin-Watson d statistic is that it is only applicable with regression models that include constant intercepts. Moreover, the statistic is not strictly reliable for models with lagged dependent variables. In such cases the Ljung-Box Q-statistic is a useful tool to determine the existence of an autocorrelation.

The *Ljung-Box Q-statistic* checks for autocorrelation in the first several lags of the residual errors. The statistic is used to test for overall autocorrelation of fitted errors of a model. The computational formula is a follows:

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$Q = T(T + 2) \sum_{i-1}^{L} (r_i^2/(T - i)$$

Where:

T = number of sample points
r = i'th autocorrelation coefficient
L = number of autocorrelation coefficients

As the formula indicates, Ljung-Box Q-statistic is the sum the squared autocorrelation. As such, it is zero only when every autocorrelation is zero. The larger the number of autocorrelations, the larger Q. When the Ljung-Box Q-statistic test is significant, the model needs to be improved. As a rule of thumb, the test is significant if its probability is greater than .99 (Stellwagen and Goodrich, 1997: p. 167).

At this point, it should be noted that several other diagnostic statistics which are beyond the scope of this book are, and can be, used, depending on the predictive regression model.

C. Solution

When the autocorrelation is either nonexistent or small, unstandardized regression models are appropriate. However, when the autocorrelation is too large, certain data transformation need to be applied to the data. The simplest way to correct for an autocorrelation is to use the method of first differences. Essentially, this method creates a new variable by computing the differences for each variable in the model and using this new variable to compute regression coefficients. In other words, $Y_i - Y_i + 1$. For instance, if a series has values of 6, 9, 7, 5, and 8, the new variable would consist of the values +3, −2, −2, and +3. For more sophisticated analyses, more complex regression-based models need to be used (e.g., exponential smoothing and Box-Jenkins for simples regression models and dynamic regression for multiple regression models).

D. Multicollinearity

A high level of multicollinearity exists when two or more of the predictor variables are correlated strongly with each other. Multicollinearity does not exist when the predictor variables correlate strongly with the dependent variable. Such a correlation can and should exist. Since multicollinearity occurs between or among predictor variables, it occurs only when multiple regression models are used.

Multicollinearity can be a problem because two or more highly correlated predictor variable makes it difficult to determine each of their effects on the dependent variable. (The stronger the correlation, the greater the problem.) When predictor variables are strongly correlated, the regression coefficients tend to vary widely, creating less precise estimates. Consequently, each coefficient may be found not to be statistically significant—even when a relationship between the predictor variables and the dependent variable actually exists. More specifically, multicollinearity results in large variances for the estimators, thus leading one to have skeptical confidence in the estimates because they may be very unreliable.

Stated succinctly, severe multicollinearity leads to:

1. Regression coefficients which can be so unreliable that they can be meaningless,
2. Impaired predictive accuracy, and
3. A standard error which is inordinately large (Gustafson, 1974: p. 138).

Predictive regression models are less sensitive to multicollinearity than are explanatory models because the former depend on the overall patterns of the relationships. Caution must be exerted,

however, since a predictive model assumes that the patterns among the predictive variables will continue to hold for the future. So long as this happens, multicollinearity is not a major problem. Nevertheless, multicollinearity is a frequent problem in short-term economic forecasting because of the high level of correlation between significant economic factors such as population, personal income, disposable income, taxes, and profits. One should be aware of its existence when collecting data so that the problem can be addressed as well as be aware of the fact that it is the less than perfect multicollinearity that causes most of the problems.

E. Recognition

A certain level of multicollinearity always exists because it is practically impossible to use meaningful predictor variables that are statistically independent of each other. A low level of multicollinearity does not affect the reliability of the regression coefficients, whereas a high level does. As a rule of thumb, *a high level of multicollinearity exists when a coefficient resulting from the correlation of two or more predictor variables is larger than .700* (Reinmuth, 1974: p. 44).

F. Solution

Fundamentally, there are two ways to alleviate multicollinearity—(1) eliminate all but one of the multicollinear predictive variables and (2) multiply the multicollinear variables and use that product as a variable to estimate the change in the dependent variable. While the first method is easier, the second is methodologically sounder since it enhances the predictive capability of the regression model (Hy et al., 1983: p. 315).

Eliminating the multicollinear variable(s). After the multicollinear variables have been identified, all but one of them can be removed from the regression model, since removal of such variable(s) may only slightly decrease the predictive power of the model. This is a useful approach, especially when the multicollinear variables measure the same concept. Selecting which variable to eliminate is a judgment call based on one's knowledge and expertise.

Employing multiplicative transformation. Another method used to eliminate severe multicollinearity is multiplicative transformation, a process which generates a new variable by multiplying the values of the multicollinear variables together. This new variable, in other words, replaces the original variables.

Although a multiplicative variable is artificial, it is quite useful in estimating the values of the dependent variable because such a variable represents the *combined effects* of two real variables, thus the predictive power of the model is not diminished. When using predictive multiple regression models, one is strongly encouraged to use multiplicative transformation to eliminate the affects of multicollinearity.

G. Selection of Variables

The variables in a predictive regression model must be selected carefully. Normally the predictor variable(s) are chosen by drawing on theoretical constructs that attempt to explain causality. These theoretical constructs, in other words, specify the predictor variable(s) which best describe and explain the direction and strength of change in the dependent variable.

When using a multiple regression model, two potential errors can occur as the predictor variables are selected. In the first place, an important predictor variable could be omitted from the model. Second, too many predictor variables could be included in the regression model.

Omitting a significant predictor variable implies that the dependent variable (Y) is not

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

affected by the omitted variable (X), even though it actually is. As a result, if Y is a function of X, but X is not included in the regression model, the estimates could be biased and inconsistent, since omitting a significant variable gives undo weight to the other predictor variable(s). Moreover, in most regression models with several predictor variables, the direction of the bias usually cannot be determined.

A tact sometimes used when developing a predictive regression model is to include all possible predictor variables and see which are the most important. Consequently, a regression model oftentimes is constructed using too many predictor variables, implying that there is nothing to lose by including many variables. This is not the case—costs do occur when using too many predictor variables.

Probably, the most important cost encountered when using too many variables is that the variance of the parameter estimators tends to increase with the number of predictor variables. The more predictor variables in the model, the larger the variance, the wider the confidence intervals, and the less precise the estimates.

These two potential errors create a dilemma. If significant predictor variables are left out of the model, the results can be biased and inconsistent. However, if too many predictor variables are included in the model, the precision of the estimates is affected negatively. The judgment of those constructing and interpreting a predictive regression model, thus, is crucial. Only predictor variables that are based on theoretical constructs and sound expert judgment need to be selected (Gustafson, 1974: p. 138).

V. A PREDICTIVE REGRESSION ILLUSTRATION

In keeping with the principal objective of this chapter, which is to focus on user comprehension rather than computational knowledge, the following illustration exemplifies how a predictive regression equation model can be used to estimate the impact future sales tax collections.

A. Problem to be Addressed

The State Department of Education is interested in locating a state boarding-type high school somewhere in the state. The state is willing to fund the operating costs of the school, but wants the selected locality to furnish adequate campus-like facilities. The school will operate on the standard school year and will eventually have a total enrollment of 300 junior and senior level high school students. Upon completion of the construction (at the beginning of the 1997–1998 school year), the school will enroll 150 students in their junior year. The following year, another 150 students will begin their junior year at the school.

Bay county proposed a 25-acre site which includes new construction of a 75,135 square feet facility for instructional and administrative support functions and renovation of five existing buildings for dormitories, dining, and multipurpose activities. The county will spend $8 million for new construction, renovations, and equipment purchases. The money will be generated by an additional one-half cent sales tax. *The basic question is, how much money will the county generate from a one year one-half cent increase in the sales tax*?

B. Assumptions

The analysis is limited to the two-year period, 1997 and 1998. The tax base will remain unchanged and no substantial economic downturn or upturn will occur before 1998.

Economic theory suggests that personal income is a good predictor of sales tax revenues. In other words, as personal income increases, disposable income increases. As disposable income

increases, people spend more on taxable goods and services. Consequently, sales tax revenues increase.

1. Data Analysis

Both county personal income and sales tax revenues were collected by quarter for the past six years, 1990–1995. (The county's fiscal year is a calendar year.) In addition, quarterly personal income estimates for 1996 and 1997 were obtained from state and national forecasting organizations.

A dynamic predictive regression model was used to estimate 1997 county revenues. The forecasted personal income figures, then, were inserted into the equation to estimate the amount of sales tax revenues that will be generated. They are as follows:

Forecast of County Sales Tax for 1997 and 1998, Based on Personal Income

1997		1998	
Quarter	Revenues	Quarter	Revenues
199701	$4,012,526	199801	$4,042,770
199702	$3,795,772	199802	$3,821,894
199703	$4,107,361	199803	$4,133,048
199704	$3,954,431	199804	$4,159,635
Annual total	$15,870,089		$16,157,347
R square	0.983		
Adjusted R square	0.978		
Durbin-Watson	1.79		

C. Findings

Taking into account the needed adjustments, as determined by revenue experts, the predictive models shows that in 1997 the county will generate approximately $15.8 million with a one cent sales tax and $7.9 with a one-half cent sales tax. In addition, it is estimated that the county would generate about $8.1 million ($16.2/2) if it kept the sales tax in effect for an additional year. The 1998 estimate indicates that the 1997 estimate is quite consistent with past data.

VI. INPUT/OUTPUT MODELS

The economic impact of an event on a particular geographic area is dependent upon the interrelationships among various industry sectors. (Because any expenditure by one industry sector (the purchaser of goods and services) involves at least one other sector (the seller of the goods and services), there will be some effect upon the economy of an area each time a purchase is made. Input/output modeling is a method of determining the magnitude of the impact based on the identified relationships among industrial sectors in an area.

In input/output modeling, the economy is represented as a variety of industry sectors which purchase goods and services from each other (and from themselves, as well) to produce outputs (e.g., raw materials, semifinished and finished goods, capital equipment, labor, and

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

taxes). Industry sectors can, and are, grouped into economic sectors on the basis of the product made, services rendered, or functions performed (e.g., agriculture, transportation, real estate, and services; including government).

The output produced by each sector or industry is sold and consumed, invested, or exported to either final users (e.g., consumers and government and industrial sectors) or as inputs to other sectors. Leakages of supply and demand from an area's economy are represented by imports and exports in an input/output model. Together these sectoral and consumer purchases represent the final demand for a product. Within this closed economy, therefore, total input must always equal total outputs.

An input/output model describes the transactions among these economic sectors in dollar values. Transactions among economic sectors include:

> sales of finished goods and services to meet final user demand, sales of raw materials and partially finished goods to intermediate users, sales to customers outside the economy being modeled, payments of wages and salaries to the labor and management forces (human resources), taxes to government to pay for publicly produced goods and payments for the use of capital, and depreciation allowances to recover the costs of capital goods used in production. The ordinary business sales and purchases transactions of an economy are systematically classified and tabulated so as to readily show the dollar value of trading among all the sectors of the economy (Grubb, 1974: p. 4).

VII. HISTORY OF INPUT/OUTPUT MODELING

The initial development of input/output models dates back to the mid-1700s with the publishing of Francois *Quesnay's Tableau Economique of 1758*, a volume that examined the French economy by incorporating the concepts of circular product flows and general equilibrium into an economic analysis. Then, in the 1870s, Leon Walras, building on Quesnay's analyses, developed a general equilibrium model for a national economy. Walras's model, however, remained a theory which was not empirically testable until the mid-1930s.

In 1936, Wassily Leontief simplified Walras's general equilibrium model so that its results could be estimated empirically. Leontief modified Walras's model by simplifying two of its basic assumptions. First, Leontief aggregated the large number of commodities in Walras's model so that the actions occurring in each industry or economic sector was measured by a group of commodities. Second, Leontief parsimoniously reduced the number of equations in the model by omitting both the supply equations for labor and the demand equations for final consumption. Finally, the remaining production equations in Walras's model were incorporated into Leontief's input/output model as linear, rather than nonlinear, equations (Richardson, 1972: p. 7).

By simplifying Walras's general equilibrium model, Leontief artificially reduced the number of equations and unknowns to a manageable number without distorting the major thrust of the model. For instance, a poultry plant undoubtedly produces a variety of products. Nevertheless, most of its products are sufficiently similar to be aggregated without distorting either the ''things'' needed to produce the products (inputs) or the number and types of products produced (outputs). In addition, by assuming that production functions are linear—although some certainly are not—Leontief was able to solve many of the implementation problems associated with Walras's theoretical model so that it could be tested empirically.

Following Leontief's lead, the United States Bureau of Labor Statistics published *The Structure of the United States Economy, 1919–1939*—a 96 sector table for the U.S. economy—in 1941 and continued this effort until 1944. (The Bureau of Labor Statistics resumed publishing

the table in 1964.) Then, in 1944 the first practical input/output model was developed to estimate the effects of the war's end on unemployment. In 1949, the Bureau of Labor Statistics developed a 200 sector input/output table (Richardson, 1972: p. 9).

Another important development was the incorporation of linear programming as an essential component to input/output modeling (Koopman, 1951: p. 125). Linear programming and input/output models are closely related since linear programming can convert input/output models into equations.

These two evolvements in the early 1950s led to the development of regional input/output models as it became increasingly possible to derive intersector transaction flow tables for regional gross outputs. A limitation, however, was that the format of the regional tables were almost identical to those of the national tables, even though regional economies are not identical to national economies. The regional input/output models are needed to represent regional production and interregional trade. As a result, this limitation coupled with the lack of available data meant that in the early 1950s the most important work on regional input/output models was conceptual instead of empirical (Isard, 1951: pp. 318, 328; Isard and Kuenne, 1953: pp. 289–301; Leontief, 1953: pp. 93–115; and Moses, 1955: pp. 803–832).

These conceptual studies, however, led to the rise of economic impact analyses which treated local direct inputs of an expanded output in a sector as an addition to the final demand. This development addressed the previously stated problem that regional production functions differ from nation production functions by abandoning regionally unadjusted national input coefficients in favor of regionally adjusted national coefficients to account for differences in regional production (Moore and Petersen, 1955: pp. 363–383).

Because of this development, regional input/output models demanded that regional coefficients of sales and purchase flows be used instead of assuming that national coefficients could be used as regional coefficients. That is, the sales and purchase flows of sectors at the national level could not be assumed to be similar to the regional level.

Because of the high cost of surveying sales and purchase flows for each major sector in each region of the country, recent efforts have focused on ways to adjust the national coefficients for a region. Some of these methods have become quite precise and are being used with increasing frequency.

It also should be noted that the United States is not the only country involved in developing input/output models. Considerable and substantial work has been and is being done in Great Britain, the Netherlands, Japan, and some former communist countries.

VIII. BASIC INPUT/OUTPUT MODEL

The model in Figure 1 incorporates five basic relationships, each consisting of a number of equations designed to measure certain functions. In this model, local consumption and demand, along with wages, prices, and profits determine employment. Capital demand depends on the relative cost of capital and labor and on local consumption and demand. Labor supply depends on population and wages, prices, and profits. Demand and supply interact to determine wages, prices, and profits. Local consumption and demand, together with wages, prices, and profits, determine market shares. Directly, and indirectly, these relationships are interrelated. Thus, estimates derived from the model are the result of satisfying all the equations simultaneously (Treyz, 1995a: pp. 1–9).

These transactions (sales and purchases) are systematically classified and tabulated in order to show the dollar value of trading among all sectors of the economy. A transaction (input/

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

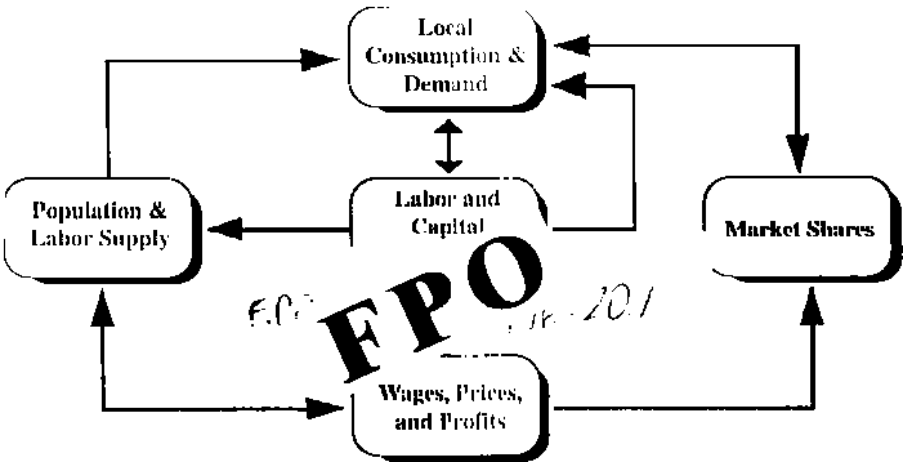


FIGURE 1 Basic model relationships. Adapted from Treyz, 1995a: pp. 2–3.

output) table summarizes the origins of the inputs and the destination of the outputs for all sectors of the economy.

Table 2 illustrates a simple transaction table consisting of three sectors—agriculture, manufacturing, and services. For purposes of this illustration each sector produces only one type of output. The three sectors are interdependent since they purchase their inputs from each other and in turn sell their outputs to each other. To produce outputs also requires labor. In this illustration, furthermore, all final goods are consumed and do not re-enter the production process as inputs (Yan, 1969: p. 6).

Row data shows output distributions to other sectors, while column data indicates the sources of inputs needed to produce goods and services. Reading across the row called agriculture, one finds that:

The total output of agricultural goods produced is \$400,
\$80 of the agricultural goods produced are ''sold'' to companies comprising the agricultural sector itself,
\$160 of the agricultural goods produced are sold to companies comprising the manufacturing sector,
None of the agricultural goods produced are sold to service sector companies, and
\$160 of the agricultural goods produced are sold to final consumers.

TABLE 2 Simple Input/Output Table

	Output to				
Input from	Agriculture	Manufacturing	Services	Final demand	Gross output
Agriculture	\$80	\$160	\$0	\$160	\$400
Manufacturing	\$40	\$40	\$20	\$300	\$400
Services	\$0	\$40	\$10	\$50	\$100
Labor	60	100	80	10	250

Source: Yan, 1969: p. 6.

Reading down the column called agricultural, one can see that the agricultural sector companies:

Consume $80 of its own products to produce $400 of total output,
Consume $40 of manufactured goods,
Consume nothing from companies in the service sector, and
Hire 60 persons to produce their goods.

Input/output tables also assume fixed proportions. This concept suggests that when the level of output is changed, the amount of input required to produce this new level changes proportionately. Accepting the concept of fixed proportions allows for the calculation of the amount of input needed to achieve a given level of input and for the development an input/output (transaction) coefficient table.

Table 3, a slightly more complex input/output table, illustrates the direct, indirect, and induced transaction coefficients for purchases for each sector. Each row shows the sector to which payments are made for purchasing inputs. Each cell within each column is the purchasing coefficient, that is, the percentage of $1.00 expenditure for inputs used by each sector listed in the column. For example, of each dollar spent by the manufacturing sector, almost 8.5 cents (8.46) is spent for agricultural related goods and 6 cents is spent for transportation.

These input coefficients are used, in a system of simultaneous linear equations to measure the magnitude of the transactions among of the economy (Grubb, 1974: p. 8). The system of equations is as follows:

$$X_i = \sum_{i\&j=1}^{\pi} a_{ij} X_j + FD_i$$

where

X_i = total annual dollar value of output of sector I
a_{ij} = dollar value of sales by sector I to sector j per dollar of output of sector j
X_j = annual dollar value of outputs of sector j
FD_i = annual dollar value of sales by sector I to final demand

The total annual outputs of sector i, X_i are accounted for in the i^{th} equation by summing the sales to other processing sectors ($a_{ij}X_j$) and the sales to final demands (FD_i). This set of simultaneous equations allows for the calculation of direct, indirect, and induced effects of the transactions.

The effects of any change in the economy, then, is determined by using matrix computations to ascertain the direct, indirect, and induced impacts of the change in the economy. From these mathematical processes the total economic effects (direct, indirect, and induced) can be computed to measure both the positive and negative impact of a proposed change in the economy.

A. Types of Economic Effects

Basically, input/output models compute three types of economic effects generated by an expenditure change in the economy. The models measure direct, indirect, and induced effects but are static rather than dynamic. *Direct effects* are changes associated with immediate changes in demands generated by employment, personal and household income, governmental expenditures, and private and public capital investment and formation.

Indirect effects are changes caused by the needs directly effected by businesses and governments. Essentially, they are interindustry impacts. These changes measure the effects on

TABLE 3 Direct, Indirect, and Induced Requirements

Purchases	Agriculture, forestry, fisheries	Mining	Construction	Manufacturing	Transportation	Communications	Utilities	Wholesale trade	Retail trade	Finance, insurance real estate	Education	Services
Agriculture, forestry, fisheries	1.22537641	0.01381689	0.03913482	0.08463068	0.02569174	0.01647700	0.01478750	0.02213544	0.02648238	0.02071452	0.03308053	0.02548514
Mining	0.07016532	1.13710134	0.07903034	0.16526314	0.05992353	0.03545022	0.42699929	0.04687679	0.05107547	0.04641886	0.07087945	0.05369148
Construction	0.02214800	0.00809733	1.01510726	0.01614288	0.02112891	0.00816496	0.01316231	0.01314784	0.01179763	0.02356364	0.01905686	0.01668937
Manufacturing	0.40302254	0.13944897	0.41883304	1.29278563	0.28294854	0.16573591	0.15496658	0.21266627	0.24872962	0.19809917	0.35017059	0.26202499
Transportation	0.06271180	0.02976966	0.06180456	0.06025327	1.06820127	0.03062685	0.02945890	0.04428404	0.04106694	0.03785431	0.05743343	0.05069589
Communications	0.02582966	0.01406657	0.02289225	0.01743794	0.04207247	1.01819875	0.01674030	0.03550498	0.03292383	0.03294260	0.03246196	0.03521565
Utilities	0.06266197	0.03149213	0.04508049	0.05131108	0.06241122	0.04266059	1.11117052	0.05748948	0.05843207	0.06066301	0.07989930	0.06113856
Wholesale trade	0.13785936	0.05149459	0.07286388	0.06362144	0.09421524	0.05671079	0.05424959	1.08098207	0.08311451	0.07902238	0.10891616	0.09699340
Retail trade	0.27153692	0.12155969	0.18963027	0.12420137	0.19374012	0.14317030	0.12561426	0.20602259	1.20451992	0.19730627	0.26225334	0.21684641
Finance, insurance, real estate	0.14070990	0.06615776	0.12361917	0.07224301	0.15865321	0.08212784	0.07433749	0.13362397	0.13407484	1.17213070	0.15674982	0.12893004
Education	0.07091660	0.05075866	0.04728889	0.03870723	0.06164289	0.06758011	0.07868178	0.05505502	0.05391564	0.07064601	1.06095950	0.05748923
Services	0.15385003	0.09486036	0.10557406	0.10555053	0.16989998	0.13374446	0.10669647	0.16810162	0.17372667	0.19532198	0.19541226	1.20436117

Source: Grubb, 1974: p. 6.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 4 Types of Multipliers

Employment multipliers	Total multipliers Direct, Indirect, and Induced
Construction/renovation	1.935
County spending	1.619
State spending	1.985
Personal income multipliers	
Construction/renovation	2.031
County spending	1.754
State spending	1.845

employment, household income, governmental expenditures, and private and public capital investment and formation added from industry purchases of all items needed to furnish a product. For example, construction contractors buy goods and services from other sectors who in turn purchase goods and services from suppliers, each of whom makes additional purchases from still other suppliers. Indirect effects measure the impacts of these purchases.

Induced effects are changes in spending patterns of households caused by changes in household income—generated by direct and indirect effects. These new expenditures are reintroduced into the economy as a new demand. Thus, the indirect effects are related to sector interaction, whereas the induced effects are related to consumption.

A demand to build a new bridge in a county illustrates these concepts. Building a bridge in a county causes the contractor to purchase various types of building materials from suppliers (*direct effect*). In turn, suppliers must buy materials from various manufacturers (*indirect effect*). Finally, this increase in demand for building products leads to income and employment increases that stimulate spending in the economy in general (*induced effect*). This process, to be sure, also works in reverse, permitting policy analysts to estimate the impact of reductions as well as expansions (Hy, 1995: pp. 3–4).

The total impact is the combined effect of the direct, indirect, and induced effects on the economy. The magnitude of each of the types of effects is determined by the size of the multiplier associated with each sector in a defined area. Based on the production functions generated through an analysis of input/output data from an area, the multiplier(s) for each sector can be calculated to provide a measure of the amount of direct, indirect, induced, and total impact of a given increase in demands of a particular sector. Table 4 illustrates typical employment and personal income multipliers provided by an input/output model for a given area.

B. Foundations of Economic Modeling

Input/output models are founded on system dynamics. Modeling, in other words, is based on interdependence and circular flows. That is, the models rely on circular, instead of linear, causality and on interdependent, rather than independent, relationships. In effect, such models abandon static, stimulus-response relationships so often used with regression based models (Barry, 1993: p. 118).

Systems dynamics assumes that all causal factors are connected in a circular process and effect each other. Analysis, thus, changes focus from linear to circular causality and from independent to interdependent relationships. (Figure 1 illustrates this point.) Mathematically, the

shift is away from correlation and regression approaches to operational modeling approaches which involve dynamic weighing—meaning that some circular loops dominate at certain periods in the analysis, followed by others, and so forth.

Dynamic models are theoretically conservative. Generally speaking, the estimates generated by these models are based on equilibrium theory, which means that an economy—be it national, state, or local—ultimately exists in a state of balance. When that balance is upset by some change, the economy will eventually correct itself over time, and return to a state of balance. As a result, over time estimates produced by these models tend to be somewhat conservative.

Equilibrium is achieved when the demand for and supply of a good or service is equal to each other and no more adjustments in the price and quantity traded are needed. Since sectors are not isolated from each other, a change in the equilibrium of one sector will affect the equilibrium of other sectors. Thus, to examine the effects of a change, the impact of that change on all affected sectors must be examined—a form of system dynamics.

C. Data Sources

Input/output models are built with data gleaned from various sources. No single data source can be used, however, since a variety of agencies gather, organize, and publish statistics. The Department of Labor is mainly, but not entirely, in charge of employment, wage, and cost of living statistics. Information on railroad and trucking freight is collected by the Interstate Commerce Commission, and information on air shipments is collected by the Federal Aviation Administration. The Federal Power Commission is the principal collector of data for electric and power companies, whereas the Department of Interior is the primary gatherer of coal and oil output data. While the Standard Industrial and Commodity Classifications are commonly adhered to, each agency feels free to use its own classification and definition and to determine on its own the frequency and timing of its statistical operations (Leontief, 1986: p. 425).

As a result of these decentralized data gathering sources and processes, input/output models incorporate data collected from a wide variety of sources. The three primary sources are: (1) the Bureau of Economic Analysis (BEA); (2) the Bureau of Labor Statistics (BLS); and (3) County Business Patterns (CBP).

The Bureau of Economic Analysis (BEA) has employment, wages, and personal income series data. These series contain data such as employment, personal income, wage and salary disbursements, other forms of labor income, proprietors' income, rental income, dividends, interest, transfer payments, and personal contributions for social insurance.

Another vital source of data used with these models is the Bureau of Labor Statistics (BLS) which furnishes data such as state and county employment, unemployment, and wage and salary figures. Yet another major data source is County Business Patterns (CBP) which, because it has ranges for suppressed data, is customarily used to estimate suppressed information. CBP data also are used to generate Regional Purchasing Coefficients (RPSs) for economic models. (RPCs are measures that show how much one sector purchases from another sector and, as such, is a major component of any economic model.)

These three primary data sources are supplemented frequently with data from other sources. Though not all-inclusive, Table 5 lists some of the major types of supplementary data and their sources.

D. Limitations

Since the 1930s input/output models have become a widely accepted technique for economic planning and decision making. Despite their popularity, these models must be used with care

TABLE 5 Supplementary Data and Their Sources

Data	Sources
Fuel and energy	State and price expenditure report
	Census of manufacturers
	Census of construction industries
	Census of service industries
	Census of retail trade
	Census of wholesale trade
Tax	Government finance
	Survey of current business
	BEA's wage and salary date
Cost of capital	Quarterly financial report for manufacturing
	Survey of current business
Gross state product	National income and product accounts
	BEA
	BLS
	Survey of current business
Housing prices	Census of housing
	National association of realtors regional and metropolitan growth rates

Source: Treyz, 1995a: pp. 4–16, 4–19.

since they possess certain weaknesses that can be debilitating to some types of economic impact analyses. These weaknesses can be grouped into the two categories—(1) reliance on ''historical'' data and (2) dependence on linear programming equations.

1. Historical Data

Input/output tables are based on periodically provided chronicled data. The construction and computation of input/output tables are complex and laborious. Immense amounts of data must be gathered, most of which is historical due to the time it takes to accumulate and collect such volumes of information. Consequently, input/output tables must be revised constantly as new data become available.

In addition to gathering and updating data an enormous amount of time is needed to solve a sizable number of simultaneous equations, inasmuch as outputs of each sector are dependent upon inputs of other sectors. The number of equations to be solved generally is two to three times the number of sectors into which the economy is divided, normally between 500–600 sectors.

More accurate estimates can be generated when input/output tables are kept up-to-date. Unfortunately, census data are gathered only every 10 years, and other survey data are costly to obtain. A promising alternative is to collect representative sample data so that estimates can be generalizable. Instead of gathering data from all sectors, only data from industries representative of a particular sector are collected. Sector input/output coefficients, then, are estimated from the sample data, reducing the time, effort, and cost of data collection and allowing input/output tables to be constructed more frequently. The reliability of the estimates, of course, must be verified.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

E. Linear Programming Equations

Linear programming equations provide a solution that can be applied under conditions of certainty. Certainty exists when a course of action definitely leads to a given result, even though the range of that result is unknown before application of the linear programming equations. These equations suggest the most efficient or least inefficient way to solve a problem. Linear programming is a powerful tool that allows one to either maximize an efficient course of action or minimize an inefficient course of action, given constraints (restrictions imposed on the solution of a problem). A typical set of linear programming equations include hundreds and sometimes even thousands of variables and constraints computed in an iterative process.

Linear programming equations are composed of several types of equations. The first, called the objective equation, states the relationship between the objective of the problem (minimizing inefficiency or maximizing efficiency) and the key variables. The other equations, called constraint equations, specify the limitations imposed on the solution. (There is one constraint equation for each constraint.)

Linear programming equations incorporate the following assumptions:

The problem can be formulated in quantitative terms

The relationships among key variables are linear; that is the variables have a constant and reciprocal ratio

The relationships among key variables are additive, meaning that the total effect of the variables is equal to the sum of the effects of each variable. Thus, all key variables are included in the equations

Linearity assumes that all statistical relationships among variables are proportional; that is, the coordinates fall on a straight plane. Thus, the use of large sets of linear equations signals a potential problem since an input/output model is the based on fixed proportional relationships between purchases and sales among sectors. While linear functions solve various types of empirical problems, the use of them also creates others, the primary one being that production functions may not be linear—especially in the agricultural, service and trade sectors.

This limitations, however, is not totally insoluble. Various nonlinear programming equations, which are similar to linear programming and which can be used when relationships among variables are assumed to be nonlinear, can be supplemented for linear equations. As Rubinstein (1975: p. 386) stated:

> Nonlinear programming problems can be solved by a number of techniques. One technique approximates the nonlinear functions as a combination of linear function segments (piecewise linear functions) and proceeds by and algorithm similar to that of linear programming. Some problems require, however, much more sophisticated techniques. While there is no single efficient solution method for the general nonlinear programming model, many efficient algorithms have been developed for certain classes of models.

Normally, however, the usual somewhat nonlinear economic assumptions of profit maximization, optimal resource allocation, and consume utility maximization are built into an input/output model as if they were linear (Richardson, 1972: p. 9).

2. Solution

When most recent transactions are not incorporated into the model, care must be exerted when interpreting the estimates. If major changes have or will occur and if they are not taken into account, the estimates furnished by the model can be misleading or incorrect. Customarily,

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

major transactional changes do not occur, and the estimates are relatively accurate. If it can be shown that the effect of the changes are negligible, an input/output model can be quite reliable.

The reliability of an input/output model also depends on the length of the period for which estimates are made. If estimates are projected for too long a period, the results will be unreliable. If, however, estimates are projected for only a couple of years into the future, the estimates may be quite reliable.

IX. AN INPUT/OUTPUT ILLUSTRATION

In keeping with the principal objective of this chapter, which is to focus on user comprehension rather than computational knowledge, the following illustration exemplifies how an input/output model can be used to estimate the impact of an economic change on a community.

A. Problem to be Addressed

The State Department of Education is interested in locating a state boarding-type high school somewhere in the state. The state is willing to fund the operating costs of the school, but wants the selected locality to furnish adequate campus-like facilities. The school will operate on the standard school year and will eventually have a total enrollment of 300 junior and senior level high school students. Upon completion of the construction (at the beginning of the 1997–1998 school year), the school will enroll 150 students in their junior year. The following year, another 150 students will begin their junior year at the school.

Bay county proposed a 25-acre site which includes new construction of a 75,135 square feet facility for instructional and administrative support functions and renovation of five existing buildings for dormitories, dining, and multipurpose activities. The county will spend $8 million for new construction, renovations, and equipment purchases. The money will be generated by an additional one-half cent sales tax.

Because of the construction and operation expenditures associated with the school, the presence of the school at this site will have a positive impact upon the economy of Bay county. For example, local construction contractors will be hired, local supplies purchased, and many school faculty and staff will likely live within the county. The purpose of this analysis is to determine the potential level of the economic impact felt throughout Bay county. *The basic question is, what specifically will the county receive for its $8 million investment*?

B. Assumptions

The analysis is limited to the three-year period of 1997 through 1999. The economic transactions associated with the school during this period have three major components: (1) construction activities for both the new and renovated buildings on the campus; (2) purchase of equipment for the school; and (3) actual operating budget of the school, including salaries and purchases.

Several assumptions—detailed below—underlie the analysis of the economic impact upon Bay county. The are based on information collected from various sources. The expenditure levels assumed to occur are depicted in Figure 2 (Hy, 1992).

1. 1997 Expenditure Assumptions

The transactions which will occur in 1997 consist of construction costs, operating expenditures, and equipment purchases. Construction costs of $6.5 million for the project were allocated en-

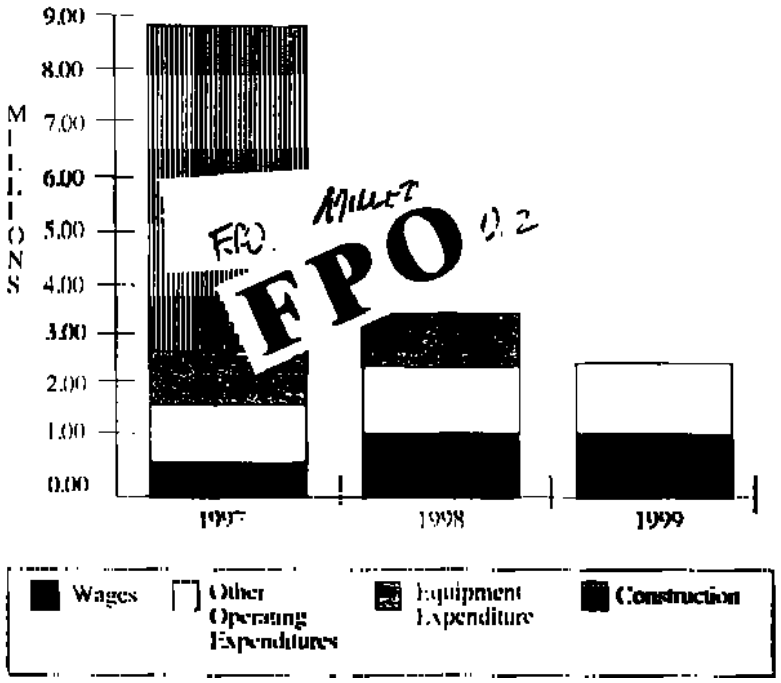


FIGURE 2 Expenditures assumptions 1997–1999 (in $ millions).

tirely to 1997, and funding is provided by Bay county. These expenditures will not be subject to state purchasing rules and regulations. Consequently, these dollars will be spent in Bay county whenever possible (decided by the model's input/output tables). Of the $6.5 million, $3.2 million will be spent on new construction. The remaining $3.3 million for renovation work. Since renovation of the existing facilities will be extensive, renovations have been treated in the analysis as new construction.

During the latter half of 1997, 150 students will be enrolled in the school. Operating expenditures and equipment purchases during 1997 amount to $1.6 million in state funds and $750,000 in county funds for equipment purchases. These equipment purchases will not subject to state purchasing rules and regulations. Therefore, it will be expected that most of these dollars will be spent in Bay county.

The state funds will be subject to state purchasing regulations. However, out of the $ 1.6 million, $508,980 consist of wages and salaries. It was assumed that most of the employees will live in Bay county and therefore, most of these dollars will be spent in Bay county. The remaining $1,091,020 is subject to state rules and regulations and programmed into the model accordingly.

2. 1998 Expenditure Assumptions

During 1998, 300 students will be enrolled in the school, construction will be completed, and some continuing purchases of equipment will be made. Equipment purchases totaling $750,000 will be made with county funds, under the same 1997 assumption—most of these dollars will be spent in Bay county. The $2.5 million operating budget for the school in 1998 consists of $1,032,000 in wages and salaries, which will be spent primarily in Bay county; and $1,468,000 in operating expenses which will be subject to state purchasing regulations.

3. 1999 Expenditures Assumptions

The 1999 scenario is similar to 1998, with the exception that the county will no longer be purchasing equipment for the school. The remaining state expenditures will be broken down in the same manner and amounts as is the case in 1998 (and, it is assumed, will continue for years to come with the possibility of at least increases for inflation).

C. Additional Assumptions

The impact analysis does not include any funds generated during summer months. Although it is very likely that some revenue generation will take place during the summer, there are no data currently available upon which to appropriately estimate such an amount. This analysis also is limited to the impact of the expenditures made by the state and county on the school. It does not include the impact of tax dollars generated by the county from additional wages and salaries.

D. Findings

Using an input/output model, the negative impact of taking $8 million out of the county's economy as well as the positive impact of the investments made to build and operate the high school have been calculated for each of the first three years of the project.

1. 1997 Impact

During 1997, there are two distinct sets of events: the construction of the school, and the initial operation of the school with 150 students enrolled.

a. Effects of Construction Expenditures The direct effects of construction activities amount to $6.5 million dollars in final demand, $1.5 million in employee compensation income, and an additional 95 jobs produced. Table 6 describes the effects of the construction expenditures on Bay county. (See the glossary at the end of this illustration for a definition of terms.) These amounts are the direct result of expenditures on construction, including all materials purchased and labor supplied to the job site.

The indirect effects trace the impact of purchases of construction materials on other local suppliers. Final demand for these firms amounts to $1.6 million, producing nearly $600,000 in employee compensation from an additional 32 jobs.

The induced effects produced by the construction activities are the result of consumer expenditures generated by the new dollars flowing through the retail economy. This amounts $5.6 million in final demand, producing $1.7 million in employee compensation income from 102 new jobs.

The total impact of construction activities in 1997 are a summation of the direct, indirect, and induced effects and amount to $12.1 million in final demand, $14.8 million in total industry output, $3.9 million in employee compensation income, and 229 new jobs.

b. Effects of Operating Expenditures The school will also generate an impact in 1997 as a result of expenditures including wages, other operating expenses (purchases of equipment and supplies), and county equipment purchases. Table 7 traces the impacts of these expenditures on Bay county.

The direct effects of the 1997 operating expenditures amount to $2.5 million dollars in final demand and $890,000 in employee compensation income from an additional 53 jobs produced. These amounts are the direct result of expenditures on staffing, equipping, and supplying the school.

The indirect effects trace the effect of the purchases of equipment and supplies on local suppliers the vendors. Total industry output for these firms amounts to $1.0 million, producing nearly $200,000 in employee compensation and an additional 10 jobs.

The induced effects produced by the operating expenditures are the result of consumer expenditures generated by the new wage income dollars flowing through the economy. This amounts to $3.2 million in final demand, $3.8 million in total industry output, producing just under $1 million in employee compensation income from 51 new jobs.

The total impact of the wages, other operating expenditures, and county equipment purchases in 1997, amount to $5.7 million in total demand, $7.4 million in total industry output, $2.0 million in employee compensation income, and 114 new jobs.

Because both of these events will be occurring within Bay county in 1997, it is appropriate to sum the impacts in order to judge the full positive impact upon the county. The first section in Table 8 is a summation of Tables 6 and 7 and details the results of the total investment in the community.

The total economic impacts on Bay county in 1997 amount to $17.8 million in total final demand, of which $12.1 million is due to construction/renovation expenditures. The additional impact of $5.7 million in total final demand is the result of $1.6 million budget for wages and other operating expenses and $750,000 in additional county purchases of scientific equipment. The total number of jobs created in Bay county is estimated to be 343. Of these, 229 jobs are construction related, while 114 are related to operating expenditures and equipment purchases. Thus, the total first year impact ratio is approximately 2:1 ($2 dollars for every dollar spent). The percentage contribution by the county to the total first year investment is approximately 82 percent.

C. Total Impact The above mentioned figures, however, do not account for the negative economic impact of taking $8 million out of the county's economy in 1997 via a one-half cent sales tax increase. The second section in Table 8 shows these impacts. Total industry output is reduced by $4.6 million and employees compensation income by $1.3 million from the loss of 104 jobs.

Net Impact. Despite these losses, the net economic impact is positive, as shown in the last section of Table 8. Total industry output will be increased by $17.5 million and employee compensation income by $4.5 million from the creation of 238 jobs. The losses occur in primarily in wholesale and retail trade, while the gains are in the service, construction, and government (including education) sectors of the county's economy.

2. 1998 Impact

The impact in 1998 will be substantially less than the 1997 impact, because construction activities will be completed during 1997. Some county equipment purchases will continue, however, and the school's operating budget in 1998 will be increased. In addition, the one-half sales tax levy will expire at the end of 1997, meaning that there will be no negative economic effects generated by a forced reduction in personal consumption spending.

The impact of the total investment in the school in 1998 is detailed in Table 9. (Since the direct, indirect, and induced effects were discussed during the examination of the 1997 impact, the examination of the 1998 and 1999 impacts will focus only on the total effects.)

a. Total Impact The total economic effect on Bay county in 1998 amount to $7.9 million in total final demand resulting from $2.5 million budget for wages and other operating expenses, and $750,000 in additional county purchases of scientific equipment. The total number of jobs created in the county is estimated to be 150, all of which are related to operating expenditures

TABLE 6 1997—Estimated Impact of Construction/Renovation Expenditures

Industry	Final demand	Total industry output	Employee comp income	Property income	Total POW income	Total value added	Number of jobs
Direct effects							
Agriculture, forestry, and fisheries	$0	$0	$0	$0	$0	$0	0
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$6,499,956	$6,499,956	$1,515,780	$1,335,184	$2,850,963	$2,892,794	94.95
Manufacturing	$0	$0	$0	$0	$0	$0	0
Trans, communications, and utilities	$0	$0	$0	$0	$0	$0	0
Wholesale and retail trade	$0	$0	$0	$0	$0	$0	0
Finance, Insurance, and real estate	$0	$0	$0	$0	$0	$0	0
Services	$0	$0	$0	$0	$0	$0	0
Govt. enterprise and special	$0	$0	$0	$0	$0	$0	0
Total—direct effects	$6,499,956	$6,499,956	$1,515,780	$1,335,184	$2,850,963	$2,892,794	95
Indirect effects							
Agriculture, forestry, and fisheries	$0	$19,663	$7,640	$3,381	$11,021	$11,773	0.73
Mining	$0	$1,127	$376	$250	$626	$626	0.01
Construction	$0	$19,537	$4,008	$4,509	$8,642	$8,642	0.26
Manufacturing	$0	$80,028	$20,539	$5,511	$26,551	$26,927	0.75
Trans, communications, and utilities	$0	$143,149	$47,215	$37,196	$84,787	$89,171	1.84
Wholesale and retail trade	$0	$119,479	$45,462	$19,663	$64,874	$81,281	1.91
Finance, Insurance, and real estate	$0	$62,996	$14,403	$27,052	$41,454	$46,214	0.83
Services	$0	$1,182,266	$449,862	$227,060	$677,047	$690,949	25.49
Govt. enterprise and special	$0	$20,539	$8,516	$501	$8,892	$8,892	0.28
Total—indirect effects	$0	$1,648,785	$598,021	$325,123	$923,895	$964,473	.32

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Induced effects							
Agriculture, forestry, and fisheries	$65,125	$167,195	$10,144	$46,464	$56,108	$57,986	2.09
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$126,618	$22,543	$25,298	$47,967	$48,969	1.44
Manufacturing	$177,841	$234,575	$46,214	$11,773	$58,237	$59,113	2.12
Trans, communications, and utilities	$399,265	$573,474	$124,113	$146,406	$270,518	$306,838	4.05
Wholesale and retail trade	$1,551,849	$1,630,124	$609,418	$278,409	$888,077	$1,066,669	29.64
Finance, insurance and real estate	$1,308,508	$1,582,032	$182,349	$873,799	$1,056,274	$1,205,936	12.75
Services	$1,900,642	$2,135,092	$664,899	$490,064	$1,155,589	$1,290,974	47.25
Govt. enterprise and special	$176,213	$234,324	$76,146	$13,025	$89,421	$89,421	2.82
Total—induced effects	$5,579,442	$6,683,433	$1,735,826	$1,885,238	$3,622,191	$4,125,907	102
Total effects*							
Agriculture, forestry, and fisheries	$65,125	$186,858	$17,784	$49,846	$87,129	$69,633	2.81
Mining	$0	$1,127	$376	$250	$626	$626	0.01
Construction	$6,499,956	$6,646,111	$1,542,581	$1,364,991	$2,907,572	$2,950,404	96.67
Manufacturing	$177,841	$314,352	$67,379	$17,784	$84,787	$86,416	2.92
Trans, communications, and utilities	$399,265	$716,373	$171,579	$183,351	$355,181	$396,134	5.9
Wholesale and retail trade	$1,551,849	$1,749,478	$654,755	$298,196	$952,826	$1,147,825	31.55
Finance, insurance and real estate	$1,308,508	$1,644,902	$196,752	$900,977	$1,097,603	$1,252,275	13.56
Services	$1,900,642	$3,317,232	$1,115,387	$717,124	$1,832,762	$1,982,173	72.74
Govt enterprise and special	$176,213	$254,863	$84,662	$13,776	$98,313	$98,313	3.1
Total effects	$12,079,398	$14,831,297	$3,851,255	$3,546,296	$7,396,800	$7,983,800	229

* Data may not sum to total effects due to rounding.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 7 1997—Estimated Impact of Wages, Other Operating Expenditures, and County Equipment Purchases

Industry	Final demand	Total industry output	Employee comp income	Property income	Total POW income	Total value added	Number of jobs
Direct effects							
Agriculture, forestry, and fisheries	$0	$0	$0	$0	$0	$0	0
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$0	$0	$0	$0	$0	0
Manufacturing	$0	$0	$0	$0	$0	$0	0
Trans, communications, and utilities	$28,598	$28,598	$4.961	$7.004	$11,965	$14,299	0.24
Wholesale and retail trade	$0	$0	$0	$0	$0	$0	0
Finance, insurance, and real estate	$0	$0	$0	$0	$0	$0	0
Services	$1,435,025	$1,435,025	$576,199	$97,614	$673,812	$673,812	36.37
Govt. enterprise and special	$1,067,902	$1,067,902	$308,896	$82,601	$391,615	$391,615	16.18
Total—direct effects	$2,531,526	$2,531,526	$890,056	$187,218	$1,077,392	$1,079,727	.53
Indirect effects							
Agriculture, forestry and fisheries	$0	$9,776	$2,918	$1,897	$4,961	$5,253	0.24
Mining	$0	$0	$0	$0		$0	0
Construction	$0	$391,477	$67,848	$76,311	$144,159	$147,223	3.73
Manufacturing	$0	$20,281	$4,961	$1,751	$7,004	$7,296	0.19
Trans, communications, and utilities	$0	$176,405	$34,289	$51,069	$85,357	$97,614	0.86
Wholesale and retail trade	$0	$57,051	$21,449	$9,484	$30,933	$38,520	0.81
Finance, insurance, and real estate	$0	$166,337	$20,136	$96,447	$116,436	$133,508	1.63
Services	$0	$129,568	$30,933	$43,335	$74,122	$79,229	2.39
Govt. enterprise and special	$0	66,175	$15,835	$5436	$21,271	$21,271	0.6
Total—indirect effects	$0	1,017,071	$198,368	$285,729	$484,243	$529,913	.10

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Induced effects							
Agriculture, forestry, and fisheries	$37,353	$96,301	5,399	$27,139	$32,684	$33,413	1.02
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	73,101	$13,132	$14,591	$27,723	$28,307	0.72
Manufacturing	$102,429	$135,550	$26,702	$6,712	$33,559	$34,435	1.05
Trans, communications, and utilities	$230,392	$331,216	$71,642	$84,628	$156,270	$177,281	1.99
Wholesale and retail trade	896,033	941,411	$351,935	$161,085	$512,874	$616,032	14.69
Finance, insurance, and real estate	$755,522	$913,543	$105,347	$504,703	$609,904	$696,283	632
Services	$1,097,681	$1,232,794	$384,473	$282,774	$667,246	$745,600	23.44
Govt enterprise and special	$82,364	$109,662	$35,687	$6,145	$41,832	$41,832	1.41
Total—induced effects	$3,201,774	$3,833,577	$994,316	$1,087,776	$2,082,092	$2,373,182	.51
Total effects*							
Agriculture, forestry, and fisheries	$37,353	$106,514	$6,317	$28,890	$37,499	$38,520	1.28
Mining	$0	$146	$0	$0	$0	$0	0
Construction	$0	$464,577	$80,980	$90,902	$171,882	$175,530	4.44
Manufacturing	$102,429	$155,394	$32,392	$8,463	$41,293	$41,584	1.28
Trans, communications, and utilities	$258,844	$535,927	$110,892	$142,554	$253,300	$288,902	3.11
Wholesale and retail trade	$896,033	$998,462	$373,530	$170,569	$543,661	$654,406	15.5
Finance, insurance, and real estate	$755,522	$1,080,026	$125,337	$601,003	$726,486	$829,936	7.95
Services	$2,532,706	$2,797,095	$991,313	$424,014	$1,415,035	$1,498,496	62.21
Govt. enterprise and special	$1,150,267	$1,243,739	$360,300	$94,300	$454,600	$454,600	18.18
Total effects	$5,733,154	$7,381,881	$2,083,060	$1,560,695	$3,643,755	$3,981,974	114

* Data may not sum to total effects due to rounding.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 8 1997—Estimated Tax Effects and Impact of Wages, Other Operating Expenditures, and County Equipment Purchases

Industry	Final demand	Total industry output	Employee comp income	Property income	Total POW income	Total value added	Number of jobs
Total effects of construction wages, other operating expenditures, and county equipment purchases							
Agriculture, forestry, and fisheries	$102,478	$293,372	$26,101	$78,736	$104,628	$108,153	4.09
Mining	$0	$1,273	$376	$250	$626	$626	0.01
Construction	$6,499,956	$7,110,688	$1,623,561	$1,455,893	$3,079,454	$3,125,934	101.11
Manufacturing	$280,270	$469,746	$99,771	$26,247	$126,080	$128,000	4.20
Trans., communications, and utilities	$658,109	$1,252,300	$282,471	$325,905	$608,481	$685,036	9.01
Wholesale and retail trade	$2,447,882	$2,747,940	$1,028,285	$468,765	$1,496,487	$1,802,231	47.05
Finance, insurance, and real estate	$2,064,030	$2,724,928	$322,089	$1,501,980	$1,824,089	$2,082,211	21.51
Services	$4,433,348	$6,114,327	$2,106,700	$1,141,138	$3,247,797	$3,480,669	134.95
Govt. enterprise and special	$1,326,480	$1,498,602	$444,962	$108,076	$552,913	$552,913	21.28
Total effects	$17,812,552	$22,213,178	$5,934,315	$5,106,990	$11,040,555	$11,965,774	343.00
Tax effects							
Agriculture, forestry, and fisheries	($29,500)	($78,400)	($8,400)	($18,900)	($27,400)	($28,900)	(1.37)
Mining	($2,100)	($3,400)	($300)	($1,600)	($2,000)	($2,600)	(0.03)
Construction	$0	($127,900)	($43,500)	($20,800)	($64,200)	($65,000)	(3.21)
Manufacturing	($117,000)	($161,000)	($29,200)	($14,100)	($43,700)	($44,700)	(1.60)
Trans., communications, and utilities	($245,700)	($324,400)	($59,100)	($89,300)	($148,300)	($163,100)	(2.69)
Wholesale and retail trade	($1,194,200)	($1,223,600)	($623,800)	($146,600)	($770,400)	($949,500)	(51.26)
Finance, insurance, and real estate	($1,262,400)	($1,331,500)	($98,500)	($683,200)	($781,500)	($997,300)	(7.32)
Services	($1,033,500)	($1,138,600)	($435,500)	($170,200)	($604,900)	($627,900)	(34.58)
Govt. enterprise and special	($224,000)	($276,700)	($76,600)	($49,000)	($125,600)	($125,700)	(2.39)
Total effects	($4,108,400)	($4,665,500)	($1,374,900)	($1,193,700)	($2,568,000)	($3,004,700)	(104.45)
Net Effects							
Agriculture, forestry, and fisheries	$72,978	$214,972	$17,701	$59,836	$77,228	$79,253	2.72
Mining	($2,100)	($2,127)	$76	($1,350)	($1,374)	($1,974)	(0.02)
Construction	$6,499,956	$6,982,788	$1,580,061	$1,435,093	$3,015,254	$3,060,934	97.90
Manufacturing	$163,270	$308,746	$70,571	$12,147	$82,380	$83,300	2.60
Trans., communications, and utilities	$412,409	$927,900	$223,371	$236,605	$460,181	$521,936	6.32
Wholesale and retail trade	$1,253,682	$1,524,340	$404,485	$322,165	$726,087	$852,731	(4.21)
Finance, insurance, and real estate	$801,630	$1,393,428	$223,589	$818,780	$1,042,589	$1,084,911	14.19
Services	$3,399,848	$4,975,727	$1,671,200	$970,938	$2,642,897	$2,852,769	100.37
Govt. enterprise and special	$1,102,480	$1,221,902	$368,362	$59,076	$427,313	$427,213	18.89
Total effects	$13,704,152	$17,547,678	$4,559,415	$3,913,290	$8,472,555	$8,961,074	238.55

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

and purchases. Thus, the total second year impacts are a ratio of about 2.4:1. The percentage contribution by the county to the total second year investment is reduced from 82% to approximately 23%, due to the elimination of construction renovation obligations. Although the second year impacts are lower, the contributions from Bay county to the project are also much less.

3. 1999 Impact

In 1999 the total impact of the school upon the county will again decline, since the county will no longer be investing money into the construction or equipping of the school. During that year, the school will be operating at full capacity of 300 students, and it is during 1995 that the operating expenditures can be looked upon as typical of what the school can be anticipated to spend in the following years. The impacts are presented in Table 10.

a. Total Impact The total economic impacts on Bay county in 1995 amount to $5.6 million in total final demand resulting from continuing state investment of $2.5 million toward the school's operating budget. The total number of jobs created in the county is estimated to be 100, all of which are related to the continuing operating expenditures. Thus, the total third year impact is a ratio of about 2.2:1. By the third year Bay county's financial obligations are negligible. The percentage contribution by the county to the third and succeeding years investment is reduced to zero. In other words, after the third year, all the money will come from outside the county (state money) to inside the county and can be assumed to be ''new'' money for the county's economy.

E. Summary

Despite simplistic theoretical assumptions, input/output models have exhibited staying power. Although the degree of interest in using input/output models at the national level is slackening, interest in using input/output models at the regional and substate levels is increasing rapidly, primarily because they can be used when data shortages prevent hardcore empirical analysis. Input/output models can be applied to answer a wide range of regional and substate economic questions. It is possible to reduce the errors of estimates by applying input/output models to a specific and narrow set of questions when alternative methods are either inappropriate or data are unavailable.

The types of information that usually can be gleaned (depending on the model) are:

Direct and indirect production changes
Total payroll costs
Direct and indirect consumer demand
Household spending patterns
Industry Multipliers
Wages and Salaries
Total value of production
Amount added to cost of production
Disposable income
Government spending
Investment spending
Labor and occupational supply and cost
Business and tax credits

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 9 1998—Estimated Impact of Wages, Other Operating Expenditures, and County Equipment Purchases

Industry	Final demand	Total industry output	Employee comp income	Property income	Total POW income	Total value added	Number of jobs
Direct effects							
Agriculture, forestry, and fisheries							
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$0	$0	$0	$0	$0	0
Manufacturing	$0	$0	$0	$0	$0	$0	0
Trans., communications, and utilities	$39,078	$39,078	$6717	$9617	$16,486	$19,539	0.31
Wholesale and retail trade	$0	$0	$0	$0	$0	$0	0
Finance, insurance, and real estate	$0	$0	$0	$0	$0	$0	0
Services	$1,974,986	$1,974,986	$793,017	$134,332	$927,349	$927,349	47.85
Govt. enterprise and special	$1,436,890	$1,436,890	$415,674	$111,202	$526,876	$526,876	21.19
Total—direct effects	$3,450,955	$3,450,955	$1,215,407	$255,151	$1,470,711	$1,473,764	69
Indirect Effects							
Agriculture, forestry, and fisheries	$0	$13,281	$3969	$2595	$6564	$7022	0.31
Mining	$0	$153	$0	$0	$0	$0	0
Construction	$0	$536,717	$93,117	$104,565	$197,529	$201,956	4.88
Manufacturing	$0	$27,172	$7022	$2442	$9770	$10,075	0.23
Trans., communication, and utilities	$0	$241,645	$46,864	$70,219	$116,777	$134,027	1.12
Wholesale and retail trade	$0	$78,309	$29,614	$12,975	$42,589	$52,664	1.07
Finance, insurance, and real estate	$0	$228,822	$27,630	$132,653	$160,283	$183,638	2.14
Services	$0	$177,837	$42,437	$59,686	$102,123	$108,534	3.17
Govt. enterprise and special	$0	$89,229	$21,245	$7405	$28,650	$28,650	0.78
Total—indirect effects	$0	$1,393,165	$271,896	$392,541	$664,285	$726,566	14

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Induced effects							
Agriculture, forestry, and fisheries	$51,296	$132,653	$7,938	$36,789	$44,726	$45,795	1.37
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$100,444	$18,013	$20,150	$36,010	$38,773	0.94
Manufacturing	$141,201	$186,386	$36,789	$9159	$46,253	$47,016	1.39
Trans., communications, and utilities	$316,749	$455,050	$98,459	$115,861	$214,473	$243,477	2.63
Wholesale and Retail Trade	$1,231,428	$1,293,556	$483,595	$221,037	$704,632	$846,597	19.3
Finance, insurance, and real	$1,038,478	$1,255,699	$144,712	$693,336	$538,201	$956,963	8.28
Services	$1,508,335	$1,693,957	$527,711	$388,952	$916,663	$1,024,434	30.76
Govt. Enterprise and Special	$111,202	$147,865	$48,196	$8255	$56,330	$56,330	1.84
Total—induced effects	$4,398,683	$5,265,609	$1,365,413	$1,493,540	$2,859,289	$3,259,385	67
Total effects*							
Agriculture, forestry, and fisheries	$51,290	$146,086	$11,907	$39,536	$51,290	$53,122	1.69
Mining	$0	$153	$0	$0	$0	$0	0
Construction	$0	$637,161	$110,977	$124,562	$235,692	$240,729	5.83
Manufacturing	$141,201	$214,015	$44,269	$12,212	$56,481	$57,854	1.66
Trans., communications, and utilities	$355,980	$735,926	$152,039	$195,697	$347,889	$396,890	4.06
Wholesale and retail trade	$1,231,428	$1,371,713	$513,057	$234,318	$747,374	$898,956	20.37
Finance, insurance and real estate	$1,038,478	$1,484,369	$172,495	$826,142	$998,331	$1,140,601	10.43
Services	$3,483,320	$3,846,933	$1,363,317	$583,123	$1,946,288	$2,060,775	81.8
Govt. enterprise and special	$1,548,093	$1,673,985	$485,114	$126,863	$611,856	$611,856	23.81
Total effects	$7,849,790	$10,110,340	$2,853,174	$2,142,454	$4,995,201	$5,460,783	150

* Data may not sum to total effects due to rounding.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 10 1999—Estimated Impact of Wages, Other Operating Expenditures, and County Equipment Purchases

Industry	Final demand	Total industry output	Employee comp income	Property income	Total POW income	Total value added	Number of jobs
Direct effects							
Agriculture, forestry, and fisheries	$0	$0	$0	$0	$0	$0	0
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$0	$0	$0	$0	$0	0
Manufacturing	$0	$0	$0	$0	$0	$0	0
Trans., communications, and utilities	$39,880	$39,880	$6859	$9890	$16,750	$19,940	0.3
Wholesale and retail trade	$0	$0	$0	$0	$0	$0	0
Finance, insurance, and real estate	$0	$0	$0	$0	$0	$0	0
Services	$1,031,935	$1,031,935	$414,433	$70,189	$484,622	$484,622	23.93
Govt. enterprise and special	$1,436,937	$1,436,937	$415,683	$111,215	$526,898	$526,898	20.63
Total—direct effects	$2,508,752	$2,508,752	$836,975	$191,294	$1,028,269	$1,031,459	45
Indirect effects							
Agriculture, forestry, and fisheries	$0	$10,528	$3350	$1914	$5264	$5583	0–5
Mining	$0	$160	$0	$0	$0	$0	0
Construction	$0	$534,711	$92,522	$103,848	$196,210	$200,517	4.65
Manufacturing	$0	$20,897	$5105	$1755	$7497	$7497	0.16
Trans., communications, and utilities	$0	$212,481	$39,242	$62,213	$101,295	$116,928	0.89
Wholesale and retail trade	$0	$66,360	$25,045	$11,007	$36,211	$44,347	0.89
Finance, insurance, and real estate	$0	$131,285	$16,590	$74,974	$91,564	$104,645	1.18
Services	$0	$118,204	$27,756	$39,720	$67,477	$72,741	2.02
Govt. enterprise and special	$0	$79,296	$17,455	$6982	$24,437	$24,437	0.63

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Total—indirect effects	$0	$1,173,923	$227,064	$302,413	$529,956	$576,696	11
Induced effects							
Agriculture, forestry, and fisheries	$35,892	$92,362	$5583	$26,002	$31,266	$32,223	0.9
Mining	$0	$0	$0	$0	$0	$0	0
Construction	$0	$70,029	$12,443	$14,038	$26,640	$27,118	0.63
Manufacturing	$98,105	$129,530	$25,842	$6,381	$32,542	$33,021	0.91
Trans., communications, and utilities	$221,095	$318,083	$68,753	$80,877	$149,789	$169,889	1.75
Wholesale and retail trade	$860,132	$903,521	$337,704	$154,415	$492,119	$591,181	12.9
Finance, insurance, and real estate	$725,337	$876,881	$101,136	$484,462	$585,119	$668,389	5.54
Services	$1,053,470	$1,183,319	$368,491	$271,822	$640,313	$715,128	20.55
Govt. enterprise and special	$76,304	$101,490	$33,040	$5611	$38,900	38.90	1.24
Total—induced effects	$3,070,335	$3,675,216	$952,992	$1,043,607	$1,996,689	$2,275,849	44
Total effects*							
Agriculture, forestry, and fisheries	$35,892	$103,369	$8933	$28,076	$37,009	$37,966	1.18
Mining	$0	$160	$0	$0	$0	$0	0
Construction	$0	$604,900	$104,964	$117,885	$222,849	$227,635	5.28
Manufacturing	$98,105	$151,384	$31,585	$8,614	$39,880	$40,997	1.12
Trans., communications, and utilities	$260,815	$569,965	$114,535	$153,139	$267,994	$306,438	2.93
Wholesale and retail trade	$860,132	$969,722	$362,589	$165,422	$528,171	$635,368	13.78
Finance, insurance, and real estate	$725,337	$1,008,166	$117,566	$559,277	$676,843	$772,874	6.72
Services	$2,085,564	$2,333,778	$811,478	$381,891	$1,192,412	$1,272,651	46.51
Govt. enterprise and special	$1,513,116	$1,617,598	$466,179	$123,932	$589,986	$589,986	22.5
Total effects	$5,578,962	$7,359,042	$2,017,830	$1,538,236	$3,555,144	$3,883,914	100

* Data may not sum to total effects due to rounding.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

F. Uses

Over the years, a variety of local issues have been analyzed with economic models. Though not all-inclusive, some are:

New plant development
Facility expansion
Utility cost increases
New port development
Rate changes
Spending increases
Tourism increases
Welfare changes
Gambling establishment development
Shopping and entertainment complexes
Transportation systems
Changing gasoline prices
Solid waste management regulations (Treyz, 1995b: p. 14).

X. THE NEXT STEP: DYNAMIC MODELS

Although dynamic economic models are beyond the scope of this chapter, a brief mention of them is helpful to understand more fully economic modeling.

Given that input/output models are static and linear, the next logical step in the modeling process is to use dynamic models, which better approximate actual economic behavior. In other words, dynamic models account for more factors and relationships than do input/output models. Dynamic models also incorporate economic responsiveness (feedback) based on economic theory. The inclusion of economic responsiveness allows the models to respond to changes in policy and to embody into the analysis labor and product substitution. Economic responsiveness also incorporates the element of time into dynamic models by recognizing that over the near term supply and demand are fixed, but that large changes to the economic system will affect both supply and demand curves over a longer period of time.

The structure of dynamic models are built by coupling theoretical descriptions of an economy with empirically estimated relationships of that economy. To handle such elaborate processes, dynamic models require more advanced mathematical methods such as linear differential and nonlinear programming equations.

Like input/output models, dynamic models have certain inherent limitations. They are as follows:

Appropriate data are not always available,
Data are not always in the form needed to address policy questions
Effects of unknowns such as legislative and judicial decision are not incorporated into the models, and
Coefficients derived from historic responses to changes in supply and demand mean that results are only valid over the ranges of past changes, thus making the impact of larger changes much more speculative.

Despite these limitations, dynamic models are widely used, primarily because of the different types of economic models they correspond most closely to actual economic behavior. This characteristic allows for numerous scenarios to be analyzed with various levels of detail and

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

frequency. In addition, the evolution in the power and capacity of computers have made dynamic economic models more useful and understandable.

A note of caution is in order. When using dynamic models, the subsequent suggestions need to be considered seriously and followed rigorously:

Make sure all spending and cost accounts are balanced,
Keep track of all the large numbers—be sure they reconcile,
Do experiments—use each key variable in a simple setting to confirm that it is acting as expected,
Always look at both the population and economic results,
When comparing more than one scenario, use the same methodology, and
Use graphs and an assortment of exhibits (Treyz, 1995b).

XI. CONCLUSION

Economic modeling has become an important tool for policy analysis since it identifies the economic impact of public policy changes. (It does not, however, incorporate nonfiscal costs and benefits.) As a result, economic modeling provides an excellent way to analyze and simulate policy changes, but it should not be used as the sole criterion for accepting or rejecting changes in public polices. Modeling is not a substitute for decision making, but it does provide useful information for the decision making process.

Despite some limitations, economic modeling provides an excellent way to analyze and estimate policy impacts before they occur. Modeling indicates what is likely to happen given a set of assumptions or actions. Economic modeling improves public policy analysis by making it more systematic and explicit. Even in its inchoate form, modeling is indispensable for the systematic understanding of the functioning and malfunctioning of an economy as well as for deciding which adjustments should be made to produce corrective actions (Leontief, 1986: p. 425). The implications for policy analysis are exceptional.

GLOSSARY

Direct Effects—The production changes associated with the immediate effects of formal demand changes.
Employment—The number of jobs (annual average) in an industry, including self employment.
Employee Compensation Income—Total payroll costs (wages, salaries, and benefits) paid by local industries.
Final Demand—The demand for goods and services from ultimate consumers, rather than from other producing industries.
Indirect Effects—the production changes in back-ward-linked industries caused by the changing input needs of directly effected industries.
Induced Effects—The changes in regional household spending patterns caused by changes in household income (generated from direct and indirect effects).
Multipliers—Type I multipliers are the direct effect (produced by a change in final demand) plus the indirect effect divided by the direct effect. Type III multipliers are the direct effect plus the indirect and induced effects divided by the direct effect.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

REFERENCES

Gustafson, R. (1974). ''The Role of Stratification in the Use of Multiregression as Applied to Single-Family Residences,'' in International Association of Assessing Officers, *The Application of Multiple Regression Analysis in Assessment Administration*, Chicago: in International Association of Assessing Officers.

Hy, R. et al. (1983). *Research Methods and Statistics*, Cincinnati, Ohio: Anderson Publishing Company.

Hy, R. et al. (1992). *An Assessment of the Economic Impact of Locating the Arkansas Math and Science School*, Little Rock, Ark.: UALR.

Isard, W. (1951). ''Interregional and Regional Input-Output Analysis: A Model of a Space Economy,'' *Research Methods and Statistics*: 33.

Isard, W. and R. E. Kuenne (1953). ''The Impact of Steel upon the Greater New York-Philadelphia Urban Industrial Region,'' *Research Methods and Statistics*: 35.

Koopmans T.C. (1951). *Activity Analysis in Production and Allocation*, New York: John Wiley and Sons.

Quade, E.S. (1982). *Analysis for Public Decisions*, 2nd. ed., North Holland, New York.

Leontief, Wassily W. (1986). *Input-Output Economics*, 2nd. ed. Oxford University Press, New York.

Leontief, W. W. (1953). ''Interregional Theory,'' in W. Leontief, H. Cheney, P. Clark, J. Duesenberry, A. Ferguson, A. Grosse, M. Holzman, W. Isard, and H. Kistin (eds.), *Studies in the Structure of the American Economy*, New York: Oxford University Press.

Leontief, W.W. (1953). ''Interregional Theory,'' in W. Leontief, et al. (eds.), *Studies in the Structure of the American Economy*, New York: OUP.

Makridakis, S. and S.C. Wheelwright (1973). *Forecasting Methods for Management*, 5th ed. John Wiley and Sons, New York.

Moore, F.T. and J.W. Petersen (1955). ''Regional Economic Reaction Paths,'' *American Economic Review*: 45.

Moses, L. (1955). ''A General Equilibrium Model of Production, Interregional Trade, and Location of Industry,'' *Research Methods and Statistics*: 42.

Reinmuth, J.E. (1974). ''The Use of Multivariate Statistical Methods in Assessment Analysis, with Special Emphasis on the Problem of Multicollinearity,'' in International Association of Assessing Officers, *The Application of Multiple Regression Analysis in Assessment Administration*, Chicago: in International Association of Assessing Officers.

Richardson, H.W. (1972). *Input—Output and Regional Economics*, New York: John Wiley and Sons.

Richmond, B. (1993). ''Systems Thinking: Critical Thinking Skills for the 1990s and Beyond,'' *Systems Dynamics Review*: 9, Summer.

Rubinstein, M. (1975). *Patterns of Problem Solving*, Englewood Cliffs, N.J.: Prentice-Hall.

Samuelson, P. (1951). ''Abstract of a Theorem Concerning Substitutability in Open Leontief Models'' in T.C. Koopmans (ed.), *Activity Analysis in Production and Allocation*, New York: John Wiley and Sons.

Stellwagen, E. and R. Goodrich (1997). *Forecast Pro for Windows*, Belmont Cal.: Business Forecast Systems.

Treyz, George I. (1995a). *Model Documentation for the REMI EDSF-53 Forecasting and Simulation Model*, Regional Economic Models, Inc., Amherst, Mass.

Treyz, G.I. (1995b). ''Policy Analysis Applications of REMI Economic Forecasting and Simulation Models.'' *International Journal of Public Administration*: 18.

Yan, C.-S. (1969). *Introduction to Input-Output Economics*, New York: Holt, Rinehart, and Winston.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

21
Computer Simulation

David Kane
Numeric Investors, Cambridge, Massachusetts

I. INTRODUCTION

The use of computers to model social phenomena is both common and controversial. On one hand, it is obvious that virtually everyone *uses* computers, at a minimum for word processing, literature searches and data analysis. Indeed, the simple act of using a hand-held calculator to determine the natural logarithm of a number is an example of computer simulation. On the other hand, the use of full-blown computer simulations has not met with widespread acceptance in the academic or the professional community involved in public administration. This chapter will provide an overview of computer simulation as a method of studying, understanding, predicting, and controlling social phenomena.

In theory, there is not a clear distinction between computational work and noncomputational work. Consider a model of an auction (Friedman and Rust, 1991; Andreoni and Miller, 1995) for FCC licenses. That model will specify the agents (Federal government, large companies, small companies, FCC bureaucrats), their preferences (for money, power, prestige), their resources (time, information, wealth), and their strategies for taking part in the auction. Interest will then focus on how different formulations of the rules of the auction lead to different outcomes in terms of who wins control of which frequencies at what price.

Standard formal models will assume that preferences, resources, and strategies take mathematical forms which are tractable. In this context, ''formal'' simply means mathematical. Instead of claiming that a company prefers higher profits to lower, we specify a mathematical relationship which implies the same thing. For example, an agent's preferences for money will often be described by a concave (positive first derivative and negative second derivative) function. This is a perfectly reasonable assumption since it makes intuitive sense that the first hundred (or million) dollars is more valuable to an individual (or firm) than the second. However, it is not always plausible or convenient to restrict the analysis to this class of mathematical specifications. In such circumstances, a computational approach provides a different method for attacking the same problem.

A computer model will also need to determine these factors. Perhaps the main difference between computer and non-computer modeling is that computer modeling requires *specified* functional forms. That is, an analytic model will often assume that the preference function is concave without being precise about the exact functional form. Any results, however, hold true for *any* function which matches the assumptions.

A computer only executes a program which must be written by some human. That program

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

simply specifies a series of steps to be carried out. A researcher could, given enough time and patience, execute the same steps herself and thereby do computational modeling without using a computer. In an analogous manner, it is possible to perform standard modeling—modeling in which the analyst can easily derive the solution without the aid of a computer—on a computer. Simply program in the model and press a button.

Consider the distinction between computer-aided addition and manual (human-only) addition. It is clear that the underlying process—adding numbers—is the same in both cases. It is also clear that certain circumstances, such as adding up the number of dollars in my wallet, call for manual addition. Other circumstances, such as adding up the number of dollars in a bank, call for computer addition. It would be possible to use a computer to add up the number of dollars in my wallet, but it might not be worth the trouble to do so. Similarly, it would be possible for a human to add up all the dollars in the bank, but it would take an extremely long time to do so. Which type of addition, manual or computational, we choose to use depends on the particular application at hand. The reason that computer-aided analysis has made such inroads into the public administration community is because, for a variety of applications, it is turning out to be the more convenient choice. Consider the following pedestrian example.

A. Traffic

One of the major motivations for the use of computational models in public administration is the recognition that problems involving the management of human society are *intrinsically* complex. People are difficult to manage because they have their own agendas and resources, their own plans and capabilities. Moreover, people change over time and learn from experience.

Consider the contrast between a river flood and a traffic jam. The standard tools of science, of hydrology and meteorology and mechanical engineering, are adequate to the task of designing and building dams and levees for the purpose of preventing floods. One must estimate how much rain is likely, how that amount of rain will raise the level of the river, how such increases in river flow will affect the course and strength of river currents, and so on. With this information, an engineer can go about designing a dam which will, one hopes, have the intended effect.

The problem of managing traffic flow, on the other hand, is infinitely more complex than the problem of managing water flow. Water molecules, unlike drivers, do not think, adapt, anticipate, or plan ahead. Water molecules obey the laws of physics and that is all they do. People, on the other hand, while also restricted by the laws of physics, have the ability to change their actions in ways not anticipated or desired by the public administrator seeking to improve traffic flow. For example, if an additional lane added to a congested highway will not have the desired effect (reduced congestion) if the very addition of the lane causes more people to use the highway. People who had traveled by other means or at other times are induced to start using the road at a time when they had previously not done so.

We can divide up the approaches to the study of traffic management (or of any other topic in public administration) into four broad categories.

1. *Experience*. A trained administrator, after having worked for many years in a highway or transportation department, will often have a very good idea about the likely effects of proposed road improvements. Experience will have taught her that certain changes are likely to have certain effects, that some rules of thumb apply to roads into a city while others apply to those out from a city, that lane widenings in one part of the system are likely to affect traffic patterns in distant areas, et cetera. Experience is a wonderful teacher. Polanyi (1958) referred to this sort of information as ''personal

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

knowledge'' and insisted that it was at least as important as more formalized sets of information and procedures.

2. *Formal models*. In conjunction with experience, the use of formal (mathematical) models of traffic flow may be helpful. Formal models in the social sciences are inspired by the success of abstract mathematical frameworks in the natural sciences, especially physics. The idea is to reduce the complications of actual traffic flow to a few important variables (number of vehicles per minute, average speed, average spacing between vehicles, probability of an accident) and then look for a mathematical relationship among these variables. Discovering such a relationship (even an approximate one) would help the administrator to better understand the behavior of traffic patterns. See Stokey and Zeckhauser (1978) for a good introduction to this framework in the context of public administration.
3. *Computer models*. A standard computer model of traffic flow picks up where the formal modeling process leaves off. The difficulty with formal models is that they are occasionally *intractable*. That is, even though there is a well-specified mathematical relationship between the variables, it is difficult if not impossible to ''solve'' the model by hand. There is not an analytic solution. The only method for determining what the future evolution of the variables will be, or to estimate the effect of changes in the value of some variables, is to use a computer. In economics, Klein (1947) and Orcutt (1952) were among the earliest formulations of this approach.
4. *Complex models*. There is a subtle but important distinction between standard computer models—which represent the state of the art in public administration—and ''complex'' models. The distinction turns on the absence of evolution, adaptation and/or learning in standard models, computational or not. The behavior of drivers evolves over time; they adapt to new highway rules and configurations; they learn new strategies for accomplishing their goals within the constraints of the environment created by public administrators. Understanding these phenomena is crucial to an accurate description of traffic patterns and, most importantly, to a reliable forecast of how proposed changes will affect the functioning of the system. To date, little if any work in public administration has employed these techniques. Anderson, Arrow and Pines (1988) provide an overview of the sorts of techniques which may someday prove useful.

On one level, the problem of traffic jams seems simple. Too many cars are trying to use too little road at the same time. One can either increase the amount of road or decrease the number of cars. On another level, it is clear that the problem of traffic jams is significantly more complex than the problems involved in controlling river flooding. The fundamental difference, from a modeling perspective, is that people learn and change their behavior while water molecules do not. Both areas will benefit from the use of computer simulation, but, because the citizens being modeled are as intelligent and resourceful as the public administrator doing the modeling, it is possible that computer simulation will prove even more important to an official in the transportation department than it has already proved to his colleague in the department of public works.

B. Goal

In either case, the goal of computer simulation is the same: to provide the administrator with an accurate model of the phenomena under consideration. The administrator would like to know both the likely course of events—the severity of traffic jams—in the absence of any changes in public policy and conditional on such changes. Perhaps the administrator has four options

for highway improvement, three changes in the roadway configuration and the ability to make no changes. It would be extremely useful for him to know what the effect of the four options would be on traffic flow. Such information would both allow him to select among the four options and to perhaps conclude that public funds would be better spent elsewhere. If instead of a highway the issue were dam building, there is no doubt that simulation would be one of the tools available for the administrator—or, more likely, the engineer contracted to do the work. The question, then, is whether or not simulation is useful in areas far afield from its origin in physics and engineering.

The purpose of this chapter is to provide an overview of the use of computer simulation in public administration, broadly understood. Currently, computer simulations, and the conceptual framework of which they are part and parcel, are being used in a wide variety of policy settings. Everything from the traffic patterns in a city; to the construction of electoral districts (Gelman and King, 1994); to the forecasting of tax revenues (Citro and Hanushek, 1991a) and Medicare spending (Feldstein, 1971); to the modeling of decentralized policy making (Kollman et al., 1995); to the making of foreign policy (Taber, 1992); to the distribution of incomes (Huberman, 1990; Durlauf, 1995); to efforts to halt the spread of tuberculosis (Brewer et al., 1996); to the importance of cooperation (Clearwater et al., 1991); to the evaluation of macroeconomic policy (Hansen and Heckman, 1996; Kydland and Prescott, 1996; Sims, 1996); to behavior in a marketplace for coffee (Misgley et al., 1995), movies (DeVany and Walls, 1995), phone systems (Lane and Maxfield, 1995), fresh fish (Weisbuch et al., 1995), international trade (Vriend, 1995) and financial services (Arifovic, 1996; Arthur, 1995); to the effects of emission controls on the earth's climate (Lempert et al., 1995; Bankes and Lempert, 1996) is being modeled with the aid of computers. Whether and to what extent these efforts are successful is one of the questions that we will explore. There can be no question, however, that the use of computer simulation has grown exponentially over the previous 20 years and that this use—given the continuing decline in the cost of computational resources and their increasing ease of use—will continue to grow. In the movie *The Graduate*, the advice proffered to Dustin Hoffman's character emphasized the importance of ''plastics.'' The use of computer simulation in public administration is receiving similar praise, not the least in this chapter. Whether that praise is justified, only time will tell.

The next section provides an example of previous work on computer simulation in each of two broad categories: applied micro-policy and applied macro-policy. Only by considering the nuts and bolts of an actual simulation is it possible to appreciate the power and problems associated with this technique. However, I will refrain from any discussion of *specific* computational software. The field is changing too rapidly to make any such discussion of summary of anything more than passing interest. Instead I will focus on the conceptual issues which lie at the heart of any modeling exercise. Section III provides a brief review of current academic work on computer simulation. The fourth section summarizes the advantages and disadvantages of computer simulation in public administration. Section V concludes.

II. EXAMPLES

In broad overview, computational work can be divided into two categories: micro-policy and macro-policy. Micro-policy refers to modeling at a ''low'' level. In general, it is concerned with the actions of particular individuals, usually persons. Macro-policy, on the other hand, is concerned with the behavior of broad aggregates. Instead of considering the behavior of unique individuals, the macro approach focuses on the behavior of variables which are derived from

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

the aggregate behavior of many individuals. Both approaches arose in economics. See Orcutt (1957) for motivation and Citro and Hanushek (1991b) for a survey of historical techniques.

For example, a micro approach to forecasting the effects of a job-training program on the unemployment rate in an individual city would focus on the characteristics of the individuals which the program is designed to help. It would consider their education, employment history, family status, and level of community involvement. The heart of the simulation would lie in these microeconomic details. In an extremely advanced application, the analyst would even attempt to ascertain the importance of *interactions* among the participants in the program. The power of computer simulation is precisely its ability to consider such interactions. For example, perhaps the likelihood of an individual finding employment after completing the program is heavily influenced by whether or not her sibling (also in the program) has found employment. That is, one is much more likely to find a job—or to search hard for one, or to attach great importance to finding one—if a close friend or relative has succeeded in doing so.

Applied macro-policy, on the other hand, is concerned with the behavior of aggregate statistics. The focus is not on whether individual A got a job after completing the program, but on total employment in an entire community with such a program in place. This summary statistic would then be compared to what total employment would have been in the absence of the program or to what total employment actually is in a (similar) community which did not institute the program. The great advantage of macro-policy is that one does not need detailed micro-information. The focus is on the relationship among employment, income, new business start-ups, and the like. These relationships may be every bit as complex as those captured by micro-simulation, but they occur at a higher level. In many situations, macro-simulation is the only option because the micro data is unavailable or too expensive to collect. In others, the analyst may feel that available theory concerning the relationships at the macro-level is more developed and accurate than that concerning the micro-level. These two exceptions aside, however, it is generally considered better to model at the lowest possible level.

A. Applied Micro Policy

The most common use of computer simulation in public administration today is in applied micro policy. Consider a proposed change in the tax laws. A state government is considering increasing the income tax rate on individuals earning more than $50,000 from 4% to 5%. Before implementing this change, however, the state legislators and governor would like to know what the likely effects of this new policy will be. These effects can range from the obvious and direct—How much will tax revenue change?—to the subtle and indirect—How many fewer high income individuals will move into the state? The attractiveness of the change depends on the size of these effects. In many ways, of course, this is both a micro and a macro level question. It is micro in that it deals with the actions of identifiable, if not actually identified, individuals. But it is also macro in that it deals with effects at a ''high'' level, that of a state. For now, consider the purely micro-aspects of the problem.

On the simplest level, one barely needs the use of a computer in order to provide a back-of-the-envelope calculation as to the effect of the proposed change. If we assume that high income individuals will make (and report) the same amount of income as last year—or an amount of income consistent with the trend in income growth—then we would expect tax receipts from these individuals to increase to the exact same extent as tax rates. Instead of paying 4% of their income, these people would pay 5% of that same income. Tax receipts would be 25% higher than they otherwise would have been.

Of course it is unlikely that high income tax payers will be as cooperative as this scenario assumes. In fact, there is little reason to believe that tax payers are unaffected by tax rates. The

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

difficulty lies in determining the size of the effect. Computer simulation is a tool for estimating how much tax revenue will actually enter the coffers of the state. For at least 35 years, the goal has been the same:

> Thus, if an adequate representation of the economic system were available, alternative tax policies could be tried upon it. By using such a model to generate the behavior that would follow from each tax policy, a base would be provided for inferences concerning the yields and relative effects of applying each of these different policies to the real economy. If the model of the economic system used were sufficiently accurate, this would be of enormous value (Orcutt, 1960, p. 896).

But the devil, as always, is in the details—or, in this case, in the code. A simulation of tax policy, or of anything else, is only valuable to the extent that it accurately represents the connections between policy options and actual effects. To explore the subtleties involved in creating such a model, the next section will examine a simulation of the effects of various public health interventions on the spread of tuberculosis.

1. Tuberculosis

Imagine that the public health administrator of a large city has been granted an extra million dollars in her budget for the express purpose of controlling tuberculosis (TB). She has a variety of worthy projects which she could spend the money on. How should she allocate theses funds in order to have the largest impact on the spread and effects of TB?

Notice that we have already abstracted away from many of the actual concerns of an administrator. We are not concerned with which strategy will increase the odds of getting more funding next year or which will have more beneficial effects in relation to other projects or which plan will be easiest to implement. Instead, the goal is simple prediction. If strategy A is followed, then X people will contract tuberculosis next year and Y of those will die from it. If strategy B is followed, then X and Y will be different. If no plan is implemented, then X and Y will be different—and presumably higher—yet again. It would be extremely valuable if the administrator could determine, even approximately, what X and Y will be, given the implementation of various strategies.

Brewer et al. (1996) implement such a model. They divide the entire population of the United States into three age groups: ≤ 15, 15–44, and ≥ 45. Within each age group, the population is divided into 18 clinical states based on two factors: TB status and human immunodeficiency virus (HIV) status. The TB states are:

1. Very low risk—general population.
2. Low risk—known TB test conversion more than two years ago.
3. High risk—recent TB test conversion or partially treated active TB.
4. Active drug sensitive TB.
5. Active drug resistant TB.

The HIV states are:

1. Uninfected—no known infection or a negative HIV test.
2. HIV—infected with the virus which causes AIDS.
3. AIDS—based on 1993 Centers for Disease Control and Prevention (CDC) definition.

Each individual in the U.S. population is classified according to both TB and HIV status. So, one category is having drug sensitive TB and being HIV negative. The model then specifies the probabilities which govern the transition from one state to another in a given year. For

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

example, there is a certain chance that a random individual will contract TB in the next year. There is a different probability that another individual with TB will die from it. In some cases, it is impossible for an individual to go from one specified state to another. An individual can not progress from HIV infected to HIV negative. The probability for this transition is therefore set to zero.

Notice that HIV status is an important part of the model even though the main focus is on TB. The reason for this is that the rate of TB is much higher among HIV-infected people than it is in the general population. Moreover, people with HIV are much more susceptible to TB, and more likely to die from it, because of their weakened immune system. It would have been possible to ignore these factors by folding the HIV population into the general U.S. population, but that would have degraded the model's accuracy. It is one of the primary benefits of computer simulation that, in cases in which data is available about an important subpopulation, the model can be expanded to use that data.

Once the entire population has been placed into one of the 18 categories and yearly transition probabilities between categories have been calculated, it is possible to ''run'' the model, to calculate how many individuals will be in each of the 18 states in each year in the future. In this case, the model is focused on two outcomes: the number of new cases of tuberculosis and the number of deaths resulting from tuberculosis each year. As a by product, the model also produces many other numbers of interest. For example, it estimates the number of new cases of HIV infection each year as well. But, as in any model, there is only so much that can be studied at once. So the central concern is with TB, both new cases and deaths.

In order to understand the construction and purpose of the model, it is necessary to review a few of the basic facts of tuberculosis as a disease (see Brewer et al., 1996). TB is spread from person to person. It is common in the third world and occurs at a rate of 9.4 per 100,000 in the United States. This rate has been essentially unchanged for 7 years, in contrast to the significant progress which had been made in previous years. Rates of TB infection are especially high among HIV sufferers and immigrants. There is a vaccine for TB, bacille Calmette-Guérin (BCG). Because BCG is expensive and because TB is not viewed as a common in the U.S. or as dangerous as other diseases like polio, the vaccine is not widely disseminated. Once infected with TB, a patient can receive treatment in the form of INH chemoprophylaxis (INH) which will reduce both his individual risk of mortality as well as the likelihood that he will infect other individuals. Once a patient infected with TB develops an active case, he may be hospitalized, depending on the severity of the case. Whether hospitalized or not, a person with active TB should undergo an extensive series of treatments for the disease. These will greatly reduce his risk of death and increase his rate of recovery, *provided that he completes the treatment course*.

It is exactly this proviso which makes projecting the effect of public health interventions so difficult. In general, one would assume that increasing the amount of resources devoted to treatment of a diseases would decrease the incidence and severity if that disease. But that is not necessarily true in the case of INH as a treatment for tuberculosis. When a patient fails to complete the INH course of treatment, he can develop and spread a strain of TB which is insensitive to standard treatments and, therefore, more dangerous. There is a positive relationship between the number of people treated with INH and the number of people who develop drug resistant TB. These are the sorts of dynamics which Brewer et al. felt were important and sought to capture.

In conducting their simulations, Brewer et al. tested five different interventions, both singly and in various combinations. They concluded that:

> In our simulations, a combination of treatment strategies, increasing entry rates and improving effectiveness, had a mutually reinforcing effect in preventing future TB cases. The synergy

> of this combination suggests that over-reliance on one aspect of treatment, such as improving effectiveness, is unlikely to be as efficacious as a program directed at both entry rates and effectiveness. Finally, these results indicate that the prevalence of HIV infection or multiple drug resistant TB in the population is an important consideration when selecting TB control measures (Brewer et al., 1996, p. 1902).

In particular, they concluded that treatment (hospitalization and other interventions after the onset of active TB) was much more effective than prevention (BCG vaccinations and INH) prior to the onset of active TB in the general population. More public health dollars should be directed to the former and, if the public health budget is fixed in size, less to the latter. A complex computer simulation, by explicitly capturing the dynamics of disease within the general population as well as within specified high-risk groups, has thereby helped to answer a concrete question of public administration.

B. Applied Macro-Policy

The primary difference between micro- and macro-policy is the issue of *aggregation*. In the TB example, each individual in the population is explicitly modeled. If we wanted to, we could examine the paths that individuals take from one state to another. The level of detail is at that of individual people. Now, in general, an administrator is not concerned with whether or not person X contracts tuberculosis. Instead, she is focused on how many people in total become infected. She cares about the behavior of population statistics. She models the individual only because it allows her to predict the behavior of the aggregate with greater accuracy.

Modeling and measuring the behavior of individuals and then aggregating the results for the entire population is one method for determining the overall statistics that are the primary focus of public administration. But, instead of this micro-approach, it is also possible to model at the macro-level; to abstract away from the behavior of individuals and look only at the behavior of the aggregate statistics which are made up of many individuals. In the TB example, this would consist of modeling *directly* the relationship, for example, between the prevalence of TB infection, the distribution of the population among various age groups, the number of new cases and other aggregate variables. Because data at the aggregate is both cheaper and more accessible than data at the individual level, the aggregate approach has a longer history. However, as the next section demonstrates, that history has not always been free from controversy.

1. The Limits to Growth

Perhaps the main reason why computer simulation is viewed with such suspicion in much of the academic and professional community is because of the controversy surrounding the work of The Club of Rome and the publication of *The Limits to Growth*. The Club of Rome was an informal organization of industrialists, academics and policy makers convened by Dr. Aurelio Peccei at the Academia dei Lincei in Rome in 1968. The members of The Club were interested in the quintessential ''big picture.'' Given current trends in population growth, agricultural production, and economic development, what would a plausible vision of the future look like? What did the succeeding decades hold in store for mankind and, to the extent that this future was problematic, what could be done about it?

Conceptually, this question is straightforward, even obvious. From the perspective of 1968, it was clear that variables of interest—world population, living standards, pollution levels, raw material costs—would attain specific values by the year 2000. it was also clear that certain values for these variables were more likely than other values. It was almost inconceivable that

world population would stabilize at the then current value of 3.6 billion. It was equally unlikely that world population would triple to above 10 billion. The question then became: Which values between these two extremes were most likely?

Up to this point, The Club of Rome had done nothing particularly different from would-be seers from time immemorial. What made their prognications original and influential was their decision to use computer simulation as the centerpiece of their efforts. They decided to work in conjunction with the System Dynamics group at MIT, headed by Jay Forrester. In 1971, Forrester published *World Dynamics*, a culmination of sorts to the more than 10 years that he and his colleagues at MIT had devoted to the ''systems'' approach to the study of human society. The model which he developed formed the basis for the one used by Meadows and her co-authors for *The Limits to Growth*.

The systems approach—termed general system theory by von Bertalanffy (1968) and system dynamics by Forrester (1968)—involves focusing on the relationships among the component parts of a complex system. The traditional approach of science since the Renaissance can be characterized as *reductionist*. Take an interesting phenomenon, say the human body, and break it into parts, say individual organs. Once you understand the behavior of the parts, you understand the behavior of the whole. The reductionist approach has proven to be an immensely powerful method for increasing our understanding of the world. Moreover, it provides a standard direction for the forward progress of science. Simply continue breaking the objects of study into smaller and smaller parts. Study the organs which comprise the body, the cells which comprise the organs, the proteins which comprise the cells, the molecules which comprise the proteins, and so on.

The systems approach seeks to turn in a different direction. The basic idea is that to fully understand the functioning of a large system, say New York City, one needs more than a knowledge of its constituent parts. One also needs an understanding of how those parts fit together, how each affects and is, in turn, effected by the others. A city is characterized, not simply by the people, firms, and infrastructure that constitute it, but by the relationships among these components (Forrester, 1969). One can not understand the whole simply by understanding the parts. Reductionism is a necessary part of the science, but it is not sufficient. A study of the system—whether it be a highway, a city, a country, or the entire world—necessitates a focus on the connections and relationships among its parts.

The World3 model is the heart of *The Limits to Growth*. It is a direct descendant of the World2 model used by Forrester in *World Dynamics*. The goal of both models is to provide an explicit, yet concise, description of the most important parts of the world ''system.'' Obviously, no model can ever hope to capture more than an almost trivial amount of the complexity of the real world. But, it should still be possible to describe the most important elements of the world system.

The World3 model sought to do so by focusing on five elements of the world system which were thought to be most important: population, food production, industrialization, pollution, and the consumption of nonrenewable natural resources. Each of these parts of the world is extremely complex and worthy of intensive study in its own right. The point of World3, however, is to consider the relationships among these variables. More people mean more pollution. More people require greater food production. An increase in population also increases industrialization and the use of resources. At the same time, however, greater industrialization makes food production more difficult. The more land that is used for industry and cities, the less there is available for farming. And the less food produced, the slower population could grow.

It is clear that the actual world is filled by ''feed-back'' loops of this type. Not only do the major subsystems of the world interact, but this interaction can feed back into itself. This

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

much is not controversial. What is, however, is the manner in which the Limits to Growth model assumes that this feedback occurs. In World3, the key functional form is the exponential. World population, resource use, pollution and other variables were all modeled as increasing exponentially. Indeed, one of the five chapters in the book is devoted to a description of the mechanics and dangers of exponential growth. Moreover, it was clear that many of the parameters that they were most interested in had grown exponentially in the recent past. Other functional forms (linear, logarithmic) would not have fit the available data nearly as well.

And yet, once the decision had been made to model the world as being dominated by exponential growth and finite resources, the conclusions were inescapable. Meadows et al. (1972, p. 24) write:

Our conclusions are:

1. If the present growth trends in world population, industrialization, pollution, food production, and resource depletion continue unchanged, the limits to growth on this planet will be reached sometime within the next one hundred years. The most probable result will be a rather sudden and uncontrollable decline in both population and industrial capacity.
2. It is possible to alter these growth trends and to establish a condition of ecological and economic stability that is sustainable far into the future. The state of the global equilibrium could be designed so that the basic material needs of each person on earth are satisfied and each person has an equal opportunity to realize his individual human potential.
3. If the world's people decide to strive for this second outcome rather than the first, the sooner they begin working to attain it, the greater will be their chances of success.

2. *Criticisms of* The Limits to Growth

Their assured tone notwithstanding, Meadows and her co-authors met with significant criticism both at the time and many years later. This criticism can be broken into two parts: complaints about the substance of their economic reasoning and warnings concerning the use of computer simulation as a tool. Beckerman (1987) summarizes the standard economic complaint well when he writes about *The Limits to Growth* that:

Its main defects included:

1. Failure to allow for the fact that changes in the balance between demand and supply for any material had, over the past, eventually led to changes in the price which provided the stimulus, where necessary, to the discovery of new resources, to the development of substitutes, to the technological improvements in methods of exploration, extraction and refinement, to substitution in the products in which they are embodied and so on. History is filled with dire predictions that if the demand for a certain product continued to grow as before, the known resources would be used up in *x* years time, and all of them have been shown by events to be absurd. The concept of ''known resources'' is a misleading one; society only ''knows'' of the resources that it is worth discovering given present and prospective demands, costs, and prices.
2. Thus the technique of inserting fixed supplies—even with some assumptions concerning eventual finite increases in these supplies—into a computer and then confronting them with indefinitely expanding demands, which must eventually overtake the supplies, bears no resemblance to the way demands and supplies have developed over

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

the past and has no foundation in economic analysis or the particular analysis of technological innovation.

3. Furthermore, even if the concept of ''finite resources'' made sense, slower growth would not enable society to continue indefinitely: it would merely postpone the fateful day of reckoning. If resources were really ''finite'' the only way the indefinite existence of society could be ensured would be to cut standards of living to infinitesimally low levels, and this did not seem to be politically feasible in democratic countries.
4. Pollution per unit of output was being reduced and could be reduced very much more if the correct pricing policies were introduced to internalize the externalities that pollution represented. This was a problem of resource allocation at any point of time and has nothing to do with resource misallocation over time, which is what the claim that growth was excessive amounted to. Indeed, pollution tended to be worse in the poorest countries and less resources were made available to reduce pollution to optimal levels in conditions of low and slowly rising incomes.
5. World food supplies had been rising faster than population for several decades and faster economic growth seemed to lead to slower population increases, rather than the reverse. The acute food shortages in many parts of the world reflected gross maldistribution of world food supplies. Slowing down the growth rate of the USA was not likely to increase availability in those parts of Africa constantly threatened by famine. If anything, insofar as it meant less aid to such countries, it would only aggravate their condition.

Beckerman's point is that, because the assumptions made by the *Limits to Growth* team are faulty, the conclusions derived from their simulation are useless. His is an attack on the substance, rather than the methods, of Meadows, et al. Other commentators were unwilling even to grant that the methodology of simulation itself had any value as a tool for understanding the likely course of future events. For example, Cole et al. (1973, p. 8), in writing about the Limits to Growth team, insist that

> They argue that in understanding the behavior of complex systems, computer models have great advantages. This view is unexceptional if we are considering the number of variables, complex interactions and speed of calculation. But it can easily and dangerously be exaggerated into what is best described as computer fetishism. The computer fetishist endows the computer model with a validity and independent power which altogether transcends the mental models which are its essential basis. Because of the prevalence of this computer fetishism it cannot be repeated to often that the validity of any computer calculation depends entirely on the quality of the data and the assumptions (mental models) which are fed into it. Computer models cannot replace theory.

And, if fetishism is not enough of a sin, Berlinski (1976, pp. 52–54) adds hubris to the list of accusations. He writes that

> *The Limits to Growth* and *World Dynamics* are ambitious and sustained efforts to see in human and social systems the elements of a dynamical system amenable to description and analysis by means of differential equations. *The Limits to Growth* stands forward and just slightly to the left of Jay Forrester's *World Dynamics*: behind them both are the lessons of *Urban Dynamics* (A Study in Slums), *Industrial Dynamics*, and *Principles of Systems*, the eminencies of applied dynamics.
>
> Things are bad and getting worse; by the end of the century a point of crisis will have arrived with all the inevitability of Death. This is the mostly Malthusian message of *The Limits to Growth* and *World Dynamics*. Professors Forrester and Meadows are voices made dolorous by the awesome powers of the exponential function.

> Nor do they hope for *technological* succor: the computer is a harsh taskmaster. Those splendid initiatives to which optimists habitually appeal—dynamic rice, colorful condoms, mulched plankton—prove languidly incapable on simulation of affecting the coming catastrophes to anything more than a marginal extent. The Ecological Evil One has so arranged the affairs of this world that a society that has successfully evaded disaster on account of failing natural resources is sure to encounter it as a result of intolerable overcrowding, gross pollution, or blighted crops.
>
> This is not a position that Professor Forrester has reached as a mere servant of fashion (Hacke ordinaire). His is no hastily and ill thought out document put together to satisfy a morbid popular taste for gloomy prognostications; it is a *theoretical* proposition, an apotheosis of Method.

C. Suggested Improvements

Other critics of computer simulation have not been so unremittingly negative. In a popular, and sympathetic, account, Kelly (1994, p. 447) writes:

> Twenty years later, the Limits to Growth simulation needs not a mere update, but a total redo. The best use for it is to stand as a challenge and a departure point to make a better model. A real predictive model of planetary society would:
>
> 1) spin significantly varied scenarios,
> 2) start with more flexible and informed assumptions,
> 3) incorporate distributed learning,
> 4) contain local and regional variation, and
> 5) if possible, demonstrate increasing complexification.

Points 1, 2, and 4 are mostly quibbles. It is obvious that a model which is able to generate a wider variety of scenarios is better than one which is not. For example, it would be good if the *Limits to Growth* model could forecast the results of dramatic drops in birth rates among educated and affluent women around the world. It would be even better, of course, if the *Limits to Growth* model had *forecast* that drop, but, since essentially no one else did, this would seem to be an unreasonable complaint. It is also obvious that more flexible assumptions are better than less flexible ones. For example, a model which allows for the inclusion of a wide range of forecasts concerning future supplies and price of various raw materials would be better than a model ''hardwires'' specific assumptions in from the beginning. It is similarly obvious that a model which allows for greater variation among different parts of the world would be an improvement. The relationship between agricultural production and population growth in the United States is different from that in India. By assuming away these distinctions, the *Limits to Growth* model ignores important parts of the reality which it seeks to capture.

But the key point is that some important parts of reality must always be ignored. That is the whole reason for building a model, for simplifying, in the first place. Once we determine that we have the time and the resources (computational and otherwise) to build a model of size X, we must determine what are the most important aspects of reality to include, given the constraints under which we are operating. It would be preferable not to have any constraints, to be able to include everything in the model that we can think to include; but that's not possible. The question then becomes, what aspects of the *Limits to Growth* model should be *removed* to make room for these proposed additions? Or, why are these the most important additions, out of the space of all model improvements, to make?

Conceptually, one can view the model building process as an attempt to create a *mapping* from the space of possible assumptions about the world to the space of possible conclusions or forecasts (Leamer, 1978). Consider the specific problem of forecasting the population of the

world 50 years from now. Obviously, if energy supplies run out and food production drops, the number of people will be less than it would be in the absence of these events. A good model will allow for different assumptions. For example, if pollution continues to increase exponentially, then world population will be 6 billion. If pollution does not increase at this rate, then the best forecast is for 10 billion. Perhaps there is some ''best'' forecast which utilizes the ''best'' available assumptions. But, on most topics, administrators will differ in their assumptions and, therefore, in their forecasts. Yet the mapping from assumptions to forecasts, the links which connect one particular set of assumptions to a forecast and a different set of assumptions to a different forecast, should be explicit. Then the administrator can judge for herself the validity of the model and the plausibility of the forecasts which it provides.

III. ACADEMIC

Why should public administrators, concerned as they are with the knitty-gritty of solving actual problems, be concerned with the purely ''academic'' uses of computer simulation? After all, there are now dozens of examples of computer simulation used successfully in public administration. Why not study these examples instead of the more esoteric explorations of academia?

The main reason is that what is esoteric today may be commonplace tomorrow. In particular, the main focus and techniques of computer simulation are undergoing dramatic change. In rough outline, we can divide the use of computer simulation into two broad categories: nonadaptive and adaptive. The nonadaptive category includes well over 95% of the current computer simulations used in public administration. This is the traditional use of computer-as-calculator. A nonadaptive simulation does not allow the entities which are being simulated to ''evolve'' or change over time. Agents, be they individuals or families or firms, are programmed, with the best available data, to behave as they have in the past.

But the past is not always prologue. People learn and adapt and change. Capturing this critical aspect of social life is the next big challenge for computer simulation in public administration. This section provides a brief overview of some of the main currents in contemporary academic work.

A. Complex Adaptive Systems

There is a sense in which the standard economic analysis of problems in public administration is fundamentally flawed (Holland and Miller, 1991; Arthur, 1991). The flaw derives from the assumption that problems in public administration can be specified using mathematical forms which are *tractable*. Consider the problem of deciding the optimal number of police officers to employ. A standard economic approach would call for the description of the marginal benefit of each additional police officer. In general, the benefits from police would be *assumed* to have some mathematical form, such as logarithmic, with pleasing properties such as being monotonically increasing with a negative second derivative. And, while it is possible to justify these properties using commonsense criteria—more police are always better, but the first few police officers are the most important—the analysis often proceeds as if the functional form has captured most of the important aspects of reality.

The problem is that tractable mathematical forms fail to capture a, perhaps the, critical characteristic of the actual practice of public administration. They fail to capture *disagreement*. In every issue of public administration, there is disagreement. Often this disagreement has obvious causes. People disagree about how many police should be hired because they have different priorities or different biases. But, there is more to disagreement than just values. Even people

with same values, people willing to make the same sorts of trade-offs, often disagree about the appropriateness of a given decision. They disagree because their predictions concerning the likely effects of a given policy differ.

Consider just one (possible) benefit of hiring a single police officer: the number of crimes that will be committed next year *relative* to the number that would have been committed in the absence of hiring an officer. Presumably, the first number is smaller than the second. Yet reasonable, well-informed public administrators disagree about the size of that decrease. This disagreement can lead them to disagree about the necessity of hiring an additional officer.

One of the benefits of computer simulation as an aid to public administration is that it provides a framework in which such disagreement is sensible, even expected. Under this view, administrators are seen as operating in extremely *complex* environments. The problem of estimating the effect on crime rates of additional police officers is difficult in a manner that the problem of estimating the effect of a new dam on river flow is not. That is, there is much more disagreement even, or perhaps especially, among the experts in the former area. Criminology is a harder, or less advanced, science than hydrology because humans are more difficult to understand than water molecules.

The current catch phrase for modeling this sort of disagreement is ''complex adaptive systems.'' Choosing the best number of police officers to hire is a *hard* problem because no one really knows the correct answer. Real world problems are more difficult than the problems which standard models construct and solve. Simon (1976, 1945) was among the first to emphasize the intrinsic complexity of public administration. His work (Simon, 1955, 1959, 1978, 1979) set the stage for an examination of the central assumption of neoclassical economic theory: that agents are best modeled as *maximizing*. Instead, Simon argued that agents ''satisfice,'' that they are unable to find the best answer and are, instead, satisfied with a merely good one. Given that agents do not (necessarily) find the optimal solution, a method for progressing from their current state, for ''adapting'' within a complex environment must be defined. Complex adaptive systems are agents which satisfice, which seek to improve their current status, not by instantly moving to the optimum, but instead by searching for and discovering improved positions in a manner analogous to biological evolution (Nelson, 1995).

Other economists (Day, 1967; Winter, 1971) have applied the satisficing approach within the neoclassical tradition. Cyert and March (1992 [1963]) provide an early example of a computer simulation along these lines. ''Bounded rationality'' (Conlisk, 1996) is another term which captures the same concept. An agent whose rationality was not bounded, i.e., one who was omniscient, would be able to maximize in even the most complex environment. Only boundedly rational agents need to satisfice. Nelson and Winter (1982) and Krugman (1996) provide extensive arguments for this outlook while Sargent (1993) surveys recent computational advances within the context of macroeconomics.

B. Landscapes

The metaphor of a landscape is both seductive in its simplicity and connected directly to the evolutionary framework derived from notions of satisficing and bounded rationality. Simply put, complex systems ''adapt'' by searching for peaks on a landscape. Due originally to Wright (1932), the idea originated in a biological context. Consider the genotype of a specific animal, for example a minnow. That particular genotype will yield a specific phenotype of a minnow. Small alterations in the genotype will yield (mostly) small variations in the phenotype. A minnow with different genes might be slower or faster, bigger or smaller, smarter or dumber, and so on. Consider a measure of the overall ''fitness'' of a specific minnow. Fitness might be

measured in length of life, number of offspring, number of offspring who survive to adulthood or in any other reasonable way.

The landscape metaphor is then a relationship between the space of possible genotypes and their relative fitness. Imagine that we can describe each possible genotype as a point in the *XY* plane. Then the fitness of that genotype can be plotted above the plane in the *Z* coordinate. Connecting different fitness points if their associated genotypes are next to one another yields a three dimensional landscape. Imagine a mountain range in which the peaks correspond to minnows which are very fit; they are smart and large and swim fast. The valleys in the landscape correspond to genotypes which produce minnows which are less fit. Evolution can then be viewed as a search over the space of possible genotypes for peaks in the fitness landscape.

The analogy to fields like public administration is fairly straightforward. Instead of the space of possible genotypes, consider the space of possible transportation networks for a city. Two different networks are ''close'' in the space of all possible networks if they are not too dissimilar. Perhaps one differs from the other only in that it has fewer freeway exits. Two networks would be ''far'' apart in the space of possible networks if they were significantly different. Perhaps one has commuter trains and large freeways while the other relies on buses. For each possible transportation network, there is an associated fitness: the cost effectiveness, the decrease in pollution, convenience, etc. Different networks will have different fitness. Connecting the fitness points of neighboring networks will create a landscape. The job of the public administrator then becomes to search for high points in the transportation landscape. Kauffman (1993) provides an encyclopedic discussion about and motivation for the landscape metaphor. See Kollman et al. (1992) for a political application and Kauffman and Macready (1995) for an economic one. The landscape metaphor is appealing because it captures, in an intuitive way, many of the complexities of the real world.

C. Cellular Automata

Cellular automata (CA) are popular tools for exploring complex phenomena. Consider a checkerboard in which each of the 64 squares may be either white or black. Each square has a maximum of 8 neighbors: three above, three below, and two to the sides. Squares along the edges have only 5 neighbors; the four corner squares have three each. Assume that each square ''wants'' to be the same color as a majority of its neighbors and has the ability to ''change'' its color in order to achieve this goal. Consider a random initial state of white and black among the 64 squares and a particular ordering of decisions—say from right to left and then top to bottom—whereby individual squares ''decide'' whether or not to change color. Question: What is the evolution of this system over time?

This classic example of a simple model, due to Schelling (1978), for studying the dynamics of segregation is an example of a cellular automata. In a cellular automata there are a number of ''cells'' each of which may take on one of a discrete number of ''states.'' In the checkerboard/segregation model, each square is a cell and there are two possible states: black and white. Time moves forward in discrete steps. At each step, one or more cells is ''updated'' according to specific ''rules'' which govern the evolution of the states over time. These rules are based on the states of a subset of cells, sometimes including the specific cell being updated. In the checkerboard model, a single cell is updated at each time step and the order of updating is row then column. The rule for updating is: If your color matches that of a majority of your neighbors, then do not change colors. Otherwise, change colors.

One of the main advantages of cellular automata as a modeling tool is its flexibility. Instead of updating the cells in order around the square, it is possible to update them at random, or to

update them all simultaneously. Instead of having the updating rule depend of the values of all the neighboring cells, it is possible to have it depend on the values of just the two (or one) neighboring cell in the same row, or the values of all the neighboring cells *and* the neighbor's neighbors. Instead of having just two states, it is possible to have more than 2. Instead of having the ''edges'' of the model behave differently, it is possible to have ''periodic boundary conditions,'' which is simply physics lingo for having the bottom of the square connect to the top and the right edge to the left, thereby forming a torus. At the same time, of course, all this flexibility is a danger, as Page (1996) points out, because any results generated by a cellular automata model should be checked tested using other reasonable updating rules, timing, et cetera.

Weisbuch et al. (1996) employ a cellular automata model to explore the spatial dynamics of pollution spread. They point out that standard tools in economics are useful for studying the spread of pollution over time, for considering the discounted costs and benefits of the adoption of pollution control technology. But studying the spread of pollution over a spatial region such as a metropolitan region is much more difficult. To tackle this problem, Weisbuch et al. construct a cellular automata with dimension 32×32 cells or grid squares. They imagine that a single agent occupies each cell and faces a choice between purchasing a car with or without pollution control technology. The dilemma is that pollution control devices cost extra to the individual who purchases them while benefiting both that individual (less pollution in her cell) and the individuals in neighboring cells. Each cell suffers from an amount of pollution which is a function of that generated by the occupant of that cell and of neighboring cells. Pollution spreads across the lattice (or grid) of cells in accordance with physically reasonable equations. In essence, the spread of pollution depends on the amount generated and on the gradient—the difference in pollution between two locations—across the lattice. The higher the gradient, the faster the spread.

At each time period, individual agents decide between the purchase of two types of cars: polluting and nonpolluting. They make their decisions on the basis of two pieces of information. First, is their prior expectations concerning the utility of the two choices. Nonpolluting cars are expensive while polluting cars are cheap. However, pollution in an agent's cell, whether its source is the agent's own car or not, is unwanted. The second piece of information is the experience of her neighbors. Each agent polls her neighbors to determine (1) the type of car which they currently use and (2) their current utility.

Within this framework, agents have no *direct* knowledge of the mechanism via which pollution spreads or even, strictly speaking, of pollution itself. All they know is that polluting cars are cheap, but neighborhoods with lots of such cars tend to be places in which people have lower utility. The parameters are adjusted so that an agent would be best off in a world in which everyone else bought nonpolluting cars except for her and worst off in a world in which the opposite were true. In other words, each agent's choices have significant consequences on the utilities of other agents.

Weisbuch et al. (1996, p. 406) conclude by noting that:

> Another interesting result is the fact that invasion of the polluted region by non-polluting devices does not always proceed from the non-polluted region. When mixed metastable attractors are reached, the polluted mixed region can be invaded from islets of non-polluting devices that started as fluctuations inside this region. Revolutions don't start in the most advanced countries, but rather in the most retrograde.

As can be seen from this quote, however, the primary purpose of academic computer simulation is *not* the accurate representation of social reality. Instead, the purpose is to develop

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

tools, both theoretical and applied, which may someday be used—in general by someone else—for this purpose. Whatever else may be said about cellular automata, their relationship to any Marxian theory of social change is speculative at best.

IV. DISADVANTAGES AND ADVANTAGES OF COMPUTER SIMULATION

An excellent summary of the disadvantages of computer simulation as a tool in public administration is provided in Cole et al. (1973, p. 12):

- By giving the spurious appearance of precise knowledge of quantities and relationships which are unknown and in many cases unknowable.
- By encouraging the neglect of factors which are difficult to quantify such as policy changes or value changes.
- By stimulating gross over-simplification, because of the problem of aggregation and the comparative simplicity of our computers and mathematical techniques.
- By encouraging the tendency to treat some features of the model as rigid and immutable.
- By making it extremely difficult for the non-numerate or those who do not have access to computers to rebut what are essential tendentious and rather naive political assumptions.

Even though they were made in the context of the debate surrounding *The Limits to Growth* more than 20 years ago, these criticisms are just as applicable to the computer models of today, both macro and micro. Computer models can, and often do, lead to spurious precision, neglect of certain factors and excessive simplification. They can become rigid in their assumptions and opaque in their analysis. And yet we must predict. We must have some method for forecasting, however roughly, the likely effects of different policy options. As Forrester (1971, p. 123) points out in the context of World2:

> On the other hand, the model presented here is probably more complete and explicit than the mental models now being used as a basis for world and national planning. The human mind is not adapted to interpreting the behavior of social systems. Over the long history of evolution it has not been necessary for man to understand these systems until very recent historical times. Evolutionary processes have not given us the mental skill needed to properly interpret the dynamic behavior of the systems of which we have now become a part.

Once a model grows beyond a certain size, we have no choice but to simulate it on a computer. There is no other option. To forbid the use of the computer as a tool is to restrict ourselves to using models of almost childish size.

V. CONCLUSION

In many ways, the case in favor of computer simulations as a tool for public administration seems overwhelming. Consider Forrester's (1971, p. 126) argument, as relevant today as it was 25 years ago, and increasingly plausible as a practical matter.

> Our social systems are far more complex and difficult to understand than our technological systems. Why, then, do we not use the same approach of making models of social systems and conducting laboratory experiments on those models before we try new laws and government

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

> programs in real life? The answer is often stated that our knowledge of social systems is insufficient for constructing useful models. But what justification can there be for the apparent assumption that we do not know enough to construct models but believe we do know enough to directly design new social systems by passing laws and starting social programs? I am suggesting that we do indeed know enough to make useful models of social systems. Conversely, we do not know enough to design the most effective social systems directly without first going through a model-building experimental phase. But I am confident, and substantial supporting evidence is beginning to accumulate, that the proper use of models of social systems can lead to far better systems, and to laws and programs that are far more effective than those created in the past.

The previous 25 years have not been kind to Forrester's prediction. It is difficult, even impossible, to *prove* that the use of a particular computer model lead to ''more effective'' laws or programs. It is obvious that simulation is more common, used more frequently. But, how could one demonstrate that this fact has had beneficial consequences? If the U.S. Treasury had been forbidden from using simulation tools for the previous few years, would tax policy have been worse? Would tax laws have been less effective?

It is by no means clear that the answer is Yes. Consider the stinging rebuke made by Berlinski 30 years ago against Forrester's *World Dynamics*:

> Here is a fat book covering 250 pages and crammed with computer-theoretical arcana. Half the work is delivered to the reader in the form of a computer printout. Recondite charts dance across the pages; there are learned references to the DYNAMO compiler, pages and pages of densely printed input-output charts, and, finally, flow charts featuring intricately drawn arrows in numbers approaching the transcendental.
>
> *Urban Dynamics* carries the ordinary systems-analytic hunger for the general to the point of baroque splendor, for in it Professor Forrester has assayed to explain the growth and decline not of any particular city, not even of a group of particular cities, but of urban area *überhaupt*. Progress on this order has been formerly unobtainable, primarily because
>
> > the influences operating within a city are so subtly and intricately interconnected that the human brain—whose response is conditioned by exposure to simple systems—finds it all but impossible to trace cause and effect.
>
> Professor Forrester, whose own brain has presumably smashed through the barrier of simple systems, has been sustained in his analysis by communion with the powers of systems theory. . . .
>
> Not only are cities systems, they are amenable to study by *general principles* of systems good everywhere and for all systems. These principles are hinted at in *Urban Dynamics* and expounded more fully in a separate text entitled *Principles of Systems*. The theory gets plotted in Chapter 4, devoted exclusively to the *structure* of systems. However, when one attends closely to the details, one finds little in the way of explication. The notion of *feedback* is never fully explained. Evidently, positive feedback is simply a barbarism denoting growth, while negative feedback has something to do with servomechanisms. But one cannot be sure. Terms like ''decision'' and ''decision mechanism'' get dragged in without much explanation:
>
> > As used here the decision process is one that controls any systems action. It can be a clear explicit human decision. It can be a subconscious decision. It can be a governing process in biological development. It may be the valve and actuator in the chemical plant. It can be the natural consequences of the physical structure of the system. Whatever the nature of the decision process, it is always embedded in

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

> a feedback loop. The decision is based on the available information; the decision controls an action that influences the system level; the new information arises to modify the decision stream.
>
> Connoisseurs will want to read this paragraph backward as well as forward (Berlinski, 1976, pp. 39–45).

If we are going to have polemicists, it is appropriate that they be as engaging and amusing as Berlinski. But, make no mistake, he is a polemicist. One can imagine him writing in the 17th century as a critic of the latest scientific advances of another era. About this upstart Newton, Berlinski might say:

> Not only is this new mathematics of Mr. Newton, a system which is supposed to apply to many things, to be all things to all men, but that same system deals with the invisible. This is a calculus of the *infinitesimal*. How can Mr. Newton see the infinitesimal? He does not say. How are we supposed to make use of this magical stuff, the mathematics of the invisible? No clues are offered. In point of fact, Mr. Newton has gone the wizards and charlatans who have proceeded him one better: not only must we accept on the faith the claims that he makes, as we must with the less abstruse tricksters among us, but we must also accept *without proof* the very tools which he proposes to prove that his theories concerning the magical, invisible force of ''gravity'' are correct. Mr. Newton seeks, not merely to pull himself up with his own theoretical bootstraps but to use those same bootstraps to hoodwink an excessively credulous audience.

Of course, in actuality, Berlinski would say nothing of the sort (Berlinski, 1995). Moreover, he would relish the opportunity to point out the absurdity of a proponent of computer simulation comparing himself and his colleagues to men such as Newton. Yet the point remains that a refusal to use computers as a tool in public administration restricts one to techniques and approaches which are not significantly advanced beyond those available to Newton's own contemporaries. Perhaps the art and science of public administration has not advanced much in the previous three centuries. One hopes that readers of this handbook would disagree.

A. Simulation as Statistics

The point of this chapter has been to review the use of computer simulation as a tool in public administration. This is not, nor could it possibly be, a guide to constructing actual simulations. That is a topic worthy of one book or many. Moreover, the field of simulation is changing so fast that any *specific* advice would be quickly rendered obsolete. Twenty-five years ago, a thorough training in computer science along with years of computer programming experience was required in order to construct even rudimentary simulations. Now, anyone with access to a desktop PC and a spreadsheet program such as Microsoft Excel can construct simulations of surprising power and flexibility. With each passing year, power and ease of use increases. A model like World3, instead of requiring a team of highly trained scientists and programmers and thousands of dollars worth of equipment and months of time, could now be implemented by a pair of first year undergraduates as a final project. Computer simulation is not going away anytime soon.

This fact suggests that public administrators need to know, if not how to implement computer simulations, then how to understand and analyze them. The proper method for doing so, as I have argued throughout, is to think of simulation as a subset of statistical inference. We seek to create a *mapping* from the space of possible assumptions about the world to the space of possible conclusions (Leamer, 1978; Lempert et al., 1995). If we assume X, then Y

follows. If, on the other hand, we assume *W*, then *Z* is the appropriate conclusion. Sometimes, most assumptions lead to the same conclusion. That is, whether we assume that *X or W* is true, the conclusion is *Y* in both cases. This makes public administration easy. Sometimes, each change in assumptions leads to a change in conclusion. This makes public administration hard.

Consider the example of a straightforward regression analysis in which we are concerned with the relationship between the dependent variable *Y* and an independent variable X_1. In particular, we seek to determine whether the relationship between X_1 and *Y* is statistically significant. The answer to this question will sometimes depend on which other variables are included in the regression equation. For example, what if, when X_2 is included in the regression, the coefficient of X_1 is statistically significant, but, if, instead of X_2, X_3 is included, the coefficient is insignificant? In that case, the key is the ''truth'' concerning whether X_2 or X_3 should be included. That choice determines the conclusion on the topic of statistical significance. But we can never know the ''true'' model. We can only make assumptions of various degrees of plausibility. Reasonable people will disagree over which of these assumptions is correct.

But, if the statistical analysis—or computer simulation—has been done well, reasonable people should not disagree about the mapping. The purpose of a computer simulation is to provide that mapping. It should make clear which assumptions lead to which conclusions. The foregoing analysis would then suggest certain rules of thumb for judging a simulation. A good simulation is:

- *Calibrated*: Accurate data is included in the construction of the simulation. The values for the parameters match, as closely as possible, empirical observation. One of the strong points of the *Limits to Growth* model and, indeed, of the entire school of systems dynamics, is a focus on accurate calibration. Brewer et al. have a similar focus on the epistemological data underlying their model of tuberculosis spread. Any (accurate) data which can be captured in the model should be captured.
- *Checked*: The functioning of the model is compared to the actual functioning of the real world. For example, the outputs of the model should be compared to actual outputs by ''running'' the model on data from an earlier time period. If all parameters are set to the appropriate values for 1988, does the model accurately predict the (known) outputs for 1989? If it does not succeed in this task, then there is little reason to believe that its predictions for the future will be any more realistic. One of the most damning criticisms of the *Limits to Growth* model concerned its failure to provide exactly this sort of benchmark. Meadows et al. failed to check its performance over other historical periods. In fact, it seems fairly clear from the structure of the model that such an exercise would have resulted in the sorts of 20-feet-of-horse-manure-in-the-streets-of-New-York-City predictions which have bedeviled trend-forecasters since time immemorial.
- *Flexible*: A good model should be flexible enough to answer a variety of questions, not about other topics, but about changes in the assumptions concerning this particular topic. If a user of the model believes that the value of a particular parameter is *X*, then, even if the builder of the computer simulation believes that the value of that variable is *Y*, there should be a method for running the simulation with Y in place of *X*.

These rules of thumb will not guarantee that a computer simulation will produce accurate results. But, a model which is inconsistent with this advice is likely to be of little if any practical

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

use. Ultimately, the accuracy and power of computer simulation as a tool in public administration will only advance in conjunction with its increasing use.

REFERENCES

Anderson, P. W., K. J. Arrow, and D. Pines, eds. (1988). *The Economy as an Evolving Complex System*, Reading, Massachusetts: Addison-Wesley.

Andreoni, J. and J. H. Miller (1995). ''Auctions with Artificial Adaptive Agents,'' *Games and Economic Behavior*, *10*: 39–64.

Arifovic, J. (1996). ''The Behavior of the Exchange Rate in the Genetic Algorithm and Experimental Economics,'' *Journal of Political Economy*, *104*(3): 510–541.

Arthur, W. B. (1991). ''Designing Economic Agents that Act Like Human Agents: A Behavioral Approach to Bounded Rationality,'' *American Economic Review, Papers and Proceedings*, *81*(2): 353–359.

Arthur, W. B. (1995). ''Complexity in Economics and Financial Markets,'' *Complexity*, *1*(1): 20–25.

Bankes, S. and R. J. Lempert (1996). ''Adaptive Strategies for Abating Climate Change: An Example of Policy Analysis for Complex Adaptive Systems,'' in *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, pp. 17–25.

Beckerman, W. (1987). *Limits to Growth*, in Eatwell, Milgate, and Newman, pp. 192–194.

Berlinski, D. (1976). *On Systems Analysis: An Essay Concerning the Limitations of Some Mathematical Methods in the Social, Political, and Biological Sciences*, Cambridge, Massachusetts: MIT Press.

Berlinski, D. (1995). *A Tour of the Calculus*, New York: Pantheon Press.

Brewer, T. F., S. J. Heymann, G. A. Colditz, M. E. Wilson, K. Auerbach, D. Kane, and H. V. Fineberg (1996). ''Evaluation of Tuberculosis Control Policies using Computer Simulation,'' *Journal of the American Medical Association*, *276*(23): 1898–1903.

Citro, C. F. and E. A. Hanushek, eds. (1991a). *Improving Information for Social Policy Decisions*, Vol. I, Washington, D.C.: National Academy Press.

Citro, C. F. and E. A. Hanushek, eds. (1991b). *Improving Information for Social Policy Decisions*, Vol. II, Washington, D.C: National Academy Press.

Clearwater, S. H., B. H. Huberman, and T. Hogg (1991). ''Cooperative Solution of Constraint Satisfaction Problem,'' *Science*, *254*: 1181–1183.

Cole, H. S. D., C. Freeman, M. Jahoda, and K. L. R. Pavitt, eds. (1973). *Models of Doom: A Critique of the Limits to Growth*, New York: Universe Books.

Conlisk, J. (1996). ''Why Bounded Rationality?'' *Journal of Economic Literature*, *34*(2): 669–700.

Cyert, R. M. and J. G. March (1992 [1963]). *A Behavioral Theory of the Firm*, Second ed., Cambridge, Massachusetts: Blackwell.

Day, R. H. (1967). ''Profits, Learning and the Convergence of Satisficing to Marginalism,'' *Quarterly Journal of Economics*, *81*(2): 302–311.

DeVany, A. and W. D. Walls (1995). ''Information, Adaptive Contracting, and Distributional Dynamics: Bayesian Choice, Bose-Einstein Statistics, and the Movies,'' Working paper, University of California, Irvine.

Durlauf, S. N. (1995). ''Neighborhood Feedbacks, Endogenous Stratification, and Income Inequality,'' Working paper, University of Wisconsin at Madison.

Eatwell, J., M. Milgate, and P. Newman, eds. (1987). *The New Palgrave: A Dictionary of Economics*, London: Macmillan Press.

Feldstein, M. S. (1971). ''An Econometric Model of the Medicare System,'' *Quarterly Journal of Economics*, *85*(1): 1–20.

Forrester, J. W. (1961). *Industrial Dynamics*, Cambridge, Massachusetts: MIT Press.

Forrester, J. W. (1968). *Principles of Systems*, Cambridge, Massachusetts: Wright-Allen Press.

Forrester, J. W. (1969). *Urban Dynamics*, Cambridge, Massachusetts: MIT Press.

Forrester, J. W. (1971). *World Dynamics*, Cambridge, Massachusetts: Wright-Allen Press.

Friedman, D. and J. Rust, eds. (1991). *The Double Auction Market: Institutions, Theories, and Evidence*, Number XIV, in ''Santa Fe Institute Studies in the Sciences of Complexity,'' Reading, Massachusetts: Addison-Wesley.
Gelman, A. and G. King (1994). ''Enhancing Democracy through Legislative Redistricting,'' *American Political Science Review*, *88*(3): 541–559.
Hansen, L. P. and J. J. Heckman (1996). ''The Empirical Foundations of Calibration,'' *Journal of Economic Perspectives*, *10*(1): 87–104.
Holland, J. H. and J. H. Miller (1991). ''Artificial Adaptive Agents and Economic Theory,'' *American Economic Review, Papers and Proceedings*, *81*(2): 365–370.
Huberman, B. A. (1990). ''The Performance of Cooperative Processes,'' *Physica D*, *42*: 38–47.
Kauffman, S. (1993). *Origins of Order*, New York: Oxford University Press.
Kauffman, S. and W. Macready (1995). ''Technological Organizations and Adaptive Evolution,'' *Complexity*, *1*(2): 26–43.
Kelly, K. (1994). *Out of Control*, Reading, Massachusetts: Addison-Wesley.
Klein, L. R. (1947). ''The Use of Econometric Models as a Guide to Economic Policy,'' *Econometrica*, *15*(2): 111–151.
Kollman, K., J. Miller, and S. Page (1992). ''Adaptive Parties in Spatial Elections,'' *American Political Science Review*, *86*: 929–937.
Kollman, K., J. Miller, and S. Page (1995). ''On the Possibility of States as Policy Laboratories,'' Working paper, Department of Political Science, University of Michigan.
Krugman, P. (1996). *The Self-Organizing Economy*, Cambridge, Massachusetts: Blackwell.
Kydland, E. E. and E. C. Prescott (1996). ''The Computational Experiment: An Econometric Tool,'' *Journal of Economic Perspectives*, *10*(1): 69–85.
Lane, D. and R. Maxfield (1995). ''Foresight, Complexity and Strategy,'' Working paper, 95-12-106, Santa Fe Institute.
Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: John Wiley & Sons.
Lempert, R. J., M. E. Schlesinger, and S. C. Bankes (1995). ''When We Don't Know the Costs or the Benefits: Adaptive Strategies for Abating Climate Change,'' Working paper, RAND.
Meadows, D. H., D. L. Meadows, J. Randers, and W. W. Behrens, III (1972). *The Limits to Growth*, New York: Universe Books.
Misgley, D. F., R. E. Marks, and L. G. Cooper (1995). ''Breeding Competitive Strategies,'' Working paper, 95-06-052, Santa Fe Institute.
Nelson, R. (1995). ''Recent Evolutionary Theorizing About Economic Change,'' *Journal of Economic Literature*, *33*(1): 48–90.
Nelson, R. R. and S. G. Winter (1982). *An Evolutionary Theory of Economic Change*, Cambridge, Massachusetts: Harvard University Press.
Orcutt, G. H. (1952). ''Toward Partial Redirection of Econometrics,'' *Review of Economics and Statistics*, *34*(3): 195–200.
Orcutt, G. H. (1957). ''A New Type of Socio-Economic System,'' *Review of Economics and Statistics*, *39*(2): 116–123.
Orcutt, G. H. (1960). ''Simulation of Economic Systems,'' *American Economic Review*, *50*(5): 893–907.
Page, S. E. (1996). ''On Incentives and Updating in Agent Based Models,'' Working paper, California Institute of Technology.
Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*, Chicago: The University of Chicago Press.
Sargent, T. (1993). *Bounded Rationality in Macroeconomics*, Oxford: Clarendon.
Schelling, T. C. (1978). *Micromotives and Macrobehavior*, New York: W. W. Norton.
Simon, H. A. (1955). ''A Behavioral Model of Rational Choice,'' *Quarterly Journal of Economics*, 69: 99–118.
Simon, H. A. (1959). ''Theories of Decision-Making in Economics and Behavioral Science,'' *American Economic Review*, *49*(3): 253–283.
Simon, H. A. (1976 [1945]). *Administrative Behavior*, Third ed., New York: The Free Press.

Simon, H. A. (1978). ''Rationality as Process and Product of Thought,'' *American Economic Review*, *68*(2): 1–16.
Simon, H. A. (1979). ''Rational Decision Making in Business Organizations,'' *American Economic Review*, *69*(4): 493–513.
Sims, C. A. (1996). ''Macroeconomics and Methodology,'' *Journal of Economic Perspectives*, *10*(1): 105–120.
Stokey, E. and R. Zeckhauser (1978). *A Primer for Policy Analysis*, New York: W. W. Norton.
Taber, C. S. (1992). ''POLI: An Expert System Model of U.S. Foreign Policy Belief Systems,'' *American Political Science Review*, *86*(4): 888–904.
von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*, Revised ed., New York: George Braziller.
Vriend, N. J. (1995). ''Self-Organization of Markets: An Example of a Computational Approach,'' *Computational Economics*, *8*: 205–231.
Weisbuch, G., A. Kirman, and D. Herreiner (1995). ''Market Organization,'' Working paper, 95-11-102, Santa Fe Institute.
Weisbuch, G., H. Gutowitz, and G. Duchateau-Nguyen (1996). ''Information Contagion and the Economics of Pollution,'' *Journal of Economic Behavior and Organization*, *29*: 389–407.
Winter, S. G. (1971). ''Satisficing, Selection, and the Innovating Remnant,'' *Quarterly Journal of Economics*, *85*(2): 237–261.
Wright, S. (1932). ''The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution,'' in *Proceedings of the Sixth International Conference of Genetics*, pp. 356–366.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

22
Data Envelopment Analysis: An Introduction

Patria D. de Lancer
University of Illinois at Springfield, Springfield, Illinois

Data development analysis (DEA) is a powerful linear programming technique for accessing the efficiency of organizations providing similar services. Charnes et al. (1978), introduced DEA as a technique that could be used specifically by the public sector to measure efficiency because managers of such decision making units (DMUs) are not free to divert resources to other programs for their profitability or attractiveness. Thus, the DEA approach provides a method of ascertaining the amount of resource conservation and/or output augmentation involved from improvements in program or managerial efficiency. In addition, the data to which DEA is applied, are not weighted by reference to market prices or other economic indicators.

Notwithstanding, to date, there have been many novel applications of DEA in both private and nonprofit sectors as well. DEA has been used in settings including education, hospital and physician evaluation, courts systems, nursing services, banking, and highway maintenance.[1] Also, it has been used to determine the efficiency of farms and coal mining, the beverage and brewing industries, the efficiency of baseball players, and the airline industry.[2]

Economists refer to DEA as the nonparametric approach to production theory or the measure of the efficiency of production by economists (Diewert and Mendoza, 1995). In general, nonparametric methods do not depend on specific population distributions; they do not require samples from normally distributed populations. One of the advantages of nonparametric methods is that of generality because they are inherently resistant to outliers and skewness and can use categorical variables, ranks, and frequency (Watson et al., 1993). As long as the assumptions underlying nonparametric methods hold the methods can be more powerful than other parametric methods.

I. HOW DOES DATA ENVELOPMENT ANALYSIS WORK?

DEA is an application of linear programming that measures the efficiency of any DMU as the maximum of a ratio of weighted outputs to weighted inputs subject to the condition that the similar ratios for every DMU be less than or equal to unity. DEA measures the ''relative efficiency'' of DMU producing similar outputs and using similar inputs. It is called ''relative efficiency'' because a hypothetical composite DMU is constructed based on all DMU's in the reference group. The rest of the DMU's are then evaluated relative to this efficient DMU.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

In this context, efficiency rating is relative to some maximum possibility so that always efficiency (E) is > 0 but $< = 1$. DEA is based on pareto-optimality condition. Pareto Efficiency or as it is also called, Pareto-Koopmans Efficiency, was explained by Bessent and Bessent (1980) as:

> A DMU is not efficient in producing its output (from given amounts of input) if it can be shown that some other DMU or combination of DMUs can produce more of some output, without producing less of any other output and without utilizing more of any resource. Conversely, a DMU is efficient if this is not possible.

Outputs and inputs are combined objectively based on this criteria (Nunamaker, 1983).

The mathematical models of DEA as introduced by Charnes, Cooper and Rhode in 1978 (the CCR models) are presented in Appendix 1 (Charnes et al., 1994). This ratio formulation of DEA models yields an objective evaluation of overall efficiency and identifies the sources and estimates the amounts of the identified inefficiencies.

DEA analysis is used to establish a best practice group of units to determine which units are inefficient compared to the best practice groups and estimate the magnitude of inefficiencies present. It tells which units should be able to improve productivity and the amount of resource savings and/or output augmentation these inefficient units must achieve to meet the level of efficiency of best practice units. A DMU is identified as inefficient only after all possible weights have been considered to give that DMU the highest rating possible consistent with the constraint that no DMU in the data set can be more than 100% efficient.

A DMU is considered to be technically inefficient, in terms of resource conservation, if some other units, or some combination of other units can: produce at least the same amounts of all outputs; use less of at least one resource; and accomplish the above with at least the same difficulties in terms of the environment. In terms of output augmentation, a unit is inefficient if some other unit or combination of units can use no more of any inputs; produce at least the same amounts of all outputs and more of at least one output; and accomplish the above with at least the same difficulty in terms of environmental constraints.

Thus, accordingly, the two dimensions of efficiency in DEA, based on Pareto efficiency can be summarized as (Bessent and Bessent, 1980 and Nunamaker, 1983)[1]:

1. Output Orientation. A DMU is not efficient in output production if the level of one or more outputs can be increased without decreasing other outputs and without utilizing more inputs.
2. Input Orientation. A DMU is not efficient in converting its inputs to outputs if other DMUs can produce the same level of output by using fewer inputs.

The DEA approach is based on building a composite DMU which is a convex combination of other DMU's inputs and outputs. This assumption of convexity is equivalent to assuming that if two production possibilities are observed in practice, then any production plan which is a convex weighted combination of the two production possibilities are observed in practice, then any production plan which is a convex weighted combination of the two production possibilities is also achievable.

The convexity assumption, together with minimum extrapolation manifests itself in that DEA estimates the efficient production frontier in a piecewise linear fashion. DEA is then an extremal prediction method which estimates the minimum level of resources needed for a DMU to produce a set of required outputs. Figure 1 is a simple graphical example of how DEA works; Table 1 depicts the variables and their values used for the graphical example.

There are five decision-making units: DMU1, DMU2, DMU3, DMU4, and DMU5. Each of these DMUs use an unique combination of two inputs (X_1 and X_2) to produce 1 unit of output

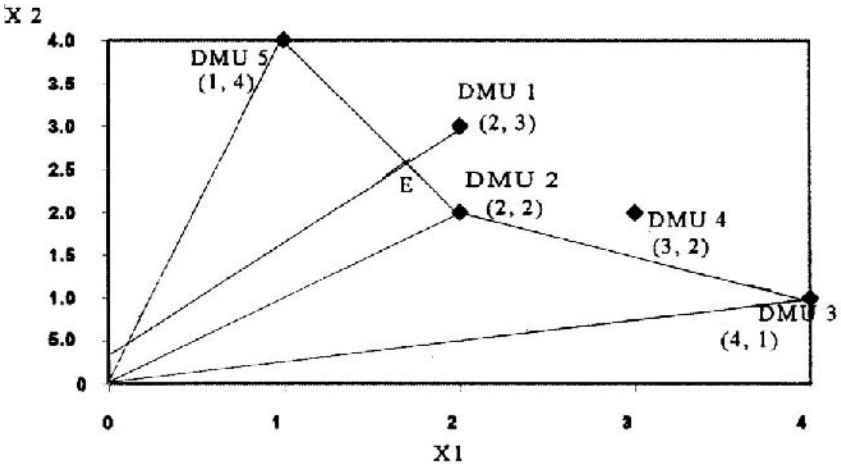


Figure 1 Simplified representation of DEA.

(Y_1). Figure 1 shows how DEA would identify DMU1 and DMU4 as inefficient. That is, both of them could decrease their use of inputs without decreasing the level of outputs. The line that connects DMUs 5, 2, and 3 is the efficiency frontier E.

As can be observed in Figure 1, DMU1 uses the same amount of input X_1 as DMU2 but more of input X_2 than DMU2 to produce the same number of outputs. In order for DMU1 to become efficient, it would have to reduce its use of both inputs. Thus, the distance between E and DMU1 is the reduction in input utilization necessary for this DMU to become efficient. DEA provides the information necessary to determine the amount of this reduction.

Currently, there are four basic DEA models, which have incorporated different interpretive possibilities. Those models are:

1. The Charnes, Cooper, and Rhodes ratio model (CCR) (shown above);
2. The Banker, Charnes, and Cooper model (BCC) which distinguishes between technical and scale inefficiencies;
3. The multiplicative model;
4. The additive model.

The choice of DEA models will depend on whether the analyst assumes constant or variable returns to scale and whether the focus is on input reduction or output augmentation to achieve efficiency. Both, the CCR and the BCC models allow for either an output or input

Table 1 Hypothetical DMUs with their Respective Input Value used and Output

	Inputs		Outputs
DMUs	X_1	X_2	Y_1
DMU 1	2	3	1
DMU 2	2	2	1
DMU 3	4	1	1
DMU 4	3	2	1
DMU 5	1	4	1

orientation in their formulation. As such, one of the main considerations of analysts using DEA is the purpose of their analysis (i.e. is it output augmentation or input reduction?).

II. ADVANTAGES AND LIMITATIONS OF DEA COMPARED TO OTHER MEASURES OF EFFICIENCY

A. A Comparison of Ratio Analysis with DEA

Ratio analysis in efficiency evaluation often takes the form of cost-benefit analysis and cost-effectiveness analysis. These are based on simple ratios or several ratios being compared simultaneously (Tseng, 1990). Ratio analysis requires common measures of inputs and outputs which could pose difficulty because of the need to transform all inputs and/or outputs into a common measure or assigning a value to each input and output (Thompson, 1908: Rossi and Freeman, 1982).

In the event ratio measures use totals, they would be biased because the mix of output and inputs are not recognized (Sherman, 1984). As suggested by Sherman, this deficiency in ratio analysis could be corrected if it was possible to set efficient relative weights or costs. However, the measure would still be biased because ratio weights would be assigned arbitrarily.

One of the most important characteristics of DEA is that it allows for the evaluation of DMU's with multiple outputs and inputs (Lewin et al., 1982). DEA assigns weights which are derived empirically from the DMU's data.

B. A Comparison of Econometrics Methods with DEA

Econometric regression techniques (simple and multiple) are often used to evaluate efficiency by comparing the expected outputs with the actual output assuming that the output level of a DMU is dependent on the level of inputs (Sexton, 1986). Therefore, if the DMU in question produces less outputs than the regression analysis, it is considered less efficient than the average DMU under evaluation.

Econometrics-regression techniques used to measure efficiency, have been criticized because the models do not discriminate between efficient and inefficient units rendering the models weak (Sherman, 1984). Also, unlike DEA, these models do not identify the inefficient unit. The models only compare the unit to the average DMU not to the best one. Lewin et al. (1982), explained that regression models also require the functional form of the production function to be specified. Further, because econometrics least-square models regress one output at a time against the inputs they make strong assumptions of independence among the outputs (Charnes et al., 1981).

The neoclassical models of frontier estimation have also been criticized because they assume the differentiability of the frontier surfaces; they assume that prices for the inputs and outputs are independent of their magnitudes and an absence of capacity constraints for all the relevant inputs (Charnes et al., 1981). In addition, according to Diewert and Mendoza (1995), unlike DEA, limited degrees of freedom will decrease the usefulness of econometrics methods if only quantity data are available.

C. Advantages of DEA

The formulation of DEA allows the analyst to use both categorical and continuous data (Banker and Morey, 1986). Moreover, DEA can be utilized to obtain the relative measures of efficiency,

involving technical, allocative, and scale efficiency (Lewin et al., 1982). DEA helps determine the magnitude of inefficiency (Sherman, 1984).

Furthermore, with DEA the analyst is able to overcome many of the limitations associated with other techniques. Those limitations can be summarized as follows[2]:

- DEA is capable of deriving a single aggregate measure of the relative efficiencies of courts in terms of their utilization of input factors to produce desired outputs.
- Models using DEA are able to handle non-commensurate multiple outputs and multiple input factors. These models do not require for all outputs or inputs in the model to have the same unit of measure.
- DEA allows the analysis to adjust for factors outside the control of the unit being evaluated.
- DEA models are not dependent on a set of priori weights or prices for the inputs or the outputs. Weights in models using DEA are derived empirically which makes the results more objective.
- With DEA one is able to handle quality factors. As long as the quality factors can be quantified or can be given a nominal value, it is possible to include them in these non-parametric models.

Using multiple measures to achieve the above results, is not particularly helpful because of the lack of consensus on the relative importance of outputs and inputs, and the fact that some agencies may seem to be more efficient or effective than the rest under one indicator but fare poor according to another.

D. Limitations of DEA

Several concerns have been raised about the robustness of the DEA models. Sherman (1984), for example, concluded that since DEA only determines ''relative efficiency,'' it can not locate all inefficient DMUs because all DMUs in a data set may be inefficient. Indeed, DEA identifies the ''best'' practice among the group under analysis. Nevertheless, this information can help the analyst decide whether the goals of a DMU are being attained compared to previous years or to other DMUs providing the same services. Management attention is directed toward identifying formal structures, processes, or other organizational factors that account for the observed differences (Charnes et al., 1994).

Another concern expressed by Sherman (1984), is that DEA does not identify the factors that cause the inefficiency. As such, it only directs a managers attention to the units where inefficiency exist. Nonetheless, as explained by Bessent and Bessent (1980), this is useful information because the inputs and outputs that are contributing to this inefficiency are also identified and administrators can decide whether a reallocation of resources is necessary or feasible.

Diewert and Mendoza (1995), have identified other limitations of DEA models. For example, measurement errors may cause the results of a model to be severely biased because the best or most efficient DMU may be best due to the overstatement of output or the understatement of an important input. This limitation, however, is also present in index number models. In econometrics models, however, the presence of such outliers can be detected and the model may be adapted to deal with this situation. As a result of these limitations, Sherman (1984) has suggested that DEA be used to complement other techniques that can address these deficiencies.

An interesting example of using DEA in conjunction with regression analysis is the stratified model used by Lovell et al. (1994) to determine the performance of secondary education in the United States. The authors used a two stage approach of modified DEA scores with

regression analysis to determine effectiveness outcomes and relate them to organizational characteristics and the operating environment.

Scholars in the field of public productivity measurement have criticized DEA based on its mathematical foundations. For example, some of the critics sustain that DEA is too technical making it difficult for the average person to interpret and understand what is going on (Hatry and Fisk, 1992). Along these lines of thoughts, Nyhan and Marlowe (1995) have argued that DEA is not practical because it requires a solid linear programming background.

As with any analytical approach, DEA requires knowledge about the formulation of models, choice of variables, data representation, interpretation of results, and knowledge of limitations (Charnes et al., 1994). For example, as with DEA, using and interpreting econometrics models require a certain level of skill. The analyst needs to understand the concept of slope of a line, r-squares, betas, significance levels, and so on.

A difficulty involved with running DEA problems with standard linear programming packages is the need to calculate as many solutions as there are DMUs. These calculations are prone to inaccurate classification of the improperly efficient and nearly efficient DMUs because of the need to calibrate the appropriate magnitude of the nonarchimedean infitesimal that introduces lower-bound constraints on all variables (Charnes et al., 1994). The non archimedean infitesimal, used in the objective function to ensure the positivity of ur and vi in Model I presented previously, is not a number and can not be approximated by any finite-valued number. However standard LP packages required this to be represented by a small number (usually 10^{-6}). These limitations have greatly been reduced with the development of DEA software packages including the Warwick-DEA, Pioneer, and IDEAS to name a few. These packages also allow the analyst to select the model most appropriate for the problem at hand.

According to Charnes et al. (1994), several extensions to the basic DEA models have been developed, which have greatly enhanced the applicability of DEA. The extensions allow the analyst to treat both nondiscretionary and categorical inputs and outputs to incorporate judgement or ancillary managerial information, and to investigate efficiency change over multiple time periods.

III. EVALUATING THE PERFORMANCE OF EMPLOYEES

The following example will illustrate the application of data envelopment analysis. The example consist of an analysis of employees of a hypothetical organization that provides certain food services to the poor. The performance of five employees will be evaluated by taking into consideration two measures of input and two measures of output. The measures are:

Inputs	*Outputs*
1. The number of hours spent	1. The number of intake applications
2. The amount of supplies used in Dollars	2. The number of visits

The measures are for a one-week period. The supervisor would like to know which of these employees is or are more efficient given the amount of resources used. That is, does the output of these employees justify their use of the resources. The data on each employee's output and amount of inputs used are summarized in Table 2.

TABLE 2 Inputs Used and Outputs Produced by Each Employee

	EMP1	EMP2	EMP3	EMP4	EMP5
Intakes	25	35	45	25	25
Visits	10	10	15	10	10
Hours	40	35	40	35	40
Supplies	105	140	135	125	120

To find the efficiency rating for each one of the employees by using DEA with a standard linear programming package, a linear programming model has to be developed. The first model to be developed will be to evaluate the relative efficiency of employee number 1. In developing this formulation, the model that will be used in Model 2 in Appendix 1. The requirement for the value of the archimedean will be set to zero (0) for simplicity purposes.

Before continuing with the example, some key linear programming terms will be defined:

a. Objective Function: this is necessary to solve any linear program model. For DEA, it specifies whether inputs are being minimized or output are being maximized.
b. Value: provides the values of the decision variables at the optimal solution. In the case of the present DEA model, it shows the weights to be assigned to each input and output.
c. Constraints: help to rule out possible combination of decision variables as feasible solutions.
d. Feasible Solution: a solution that satisfies all the constraints.
c. Reduced Cost: indicates how much the objective function coefficient of each decision variable would have to improve before it would be possible for that variable to assume a positive value in the optimal solution. (For a maximization problem, improve means get bigger; for a minimization problem, it means get smaller)
d. Slack/Surplus: in a problem having $<=$ constraints it is the amount of unused resources. It is added to the left-hand side of a less-than-or equal to constraint to convert the constraint into an inequality.
e. Dual Prices: this is associated with a constraint and represents the improvement in the optimal value of the objective function per unit increase in the right-hand side of the constraint. In Model 2 of the DEA formulation it represents the proportion of the input/output levels of the peer employee going into the composite set.

Consistent with Model 2 in Appendix 1. the formulation for this problem is as follows: MAX $25UA + 10UB$

SUBJECT TO

2. $25UA + 10UB - 40VA - 105VB <= 0$
3. $35UA + 10UB - 35VA - 140VB <= 0$
4. $45UA + 15UB - 40VA - 135VB <= 0$
5. $25UA + 10UB - 35VA - 125VB <= 0$
6. $25UA + 15UB - 40VA - 120VB <= 0$
7. $40VA + 105VB = 1$
8. $UA, UB, VA, VB >= 0$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

where:

VA = weight of input hours;

VB = weight for input supplies;

UA = weight for output intakes;

UB = weight for output visits.

The first line of the formulation above is the objective function. Here the goal is to maximize the weighted sum of the outputs. Since the evaluation is for employee 1, the coefficient of the variables in the objective function are those of employee 1. Then, under the heading ''subject to,'' there are several constraints. By specifying those conditions, DEA allows each DMU to select any possible set of weights that will present its efficiency in the best possible light (Tankesley, 1990). The specified conditions are.

1. no weight can be negative,
2. each DMU must be allowed to use the same set of weights to evaluate its efficiency, and
3. the ratios resulting from each of these separate evaluations must not exceed one.

In this example, DEA asks the employee under analysis to maximize his efficiency subject to the conditions. The efficiency ratio for this employee is forced into comparison with the efficiency ratio for each other employee using the chosen input and output weights. It is in this manner that the relative efficiency of the unit under analysis is calculated. The solution of this DEA model is shown in Table 3.

A. Efficiency Rating

The efficiency rating is the proportion of inputs that a unit (in this case employees) should use in order to achieve its output level and remain efficient as it is compared with other units. The efficiency rating is printed as the value for the objective function. The objective function speci-

Table 3 Computer Solution of the Data Envelopment Analysis Employee 1 Model (Using LINDO)

Objective function value		
1) .8214286		
Variable	Value	Reduced cost
INTAKE	.007143	0.000000
VISITS	.064286	0.000000
HOURS	.000000	6.190476
SUPPLIES	.009524	0.000000
R	.000000	0.000000
Row	Slack or surplus	Dual prices
2)	0.178571	0.000000
3)	0.440476	0.000000
4)	0.000000	0.416667
5)	0.369048	0.000000
6)	0.000000	0.250000
7)	0.000000	0.821429
8)	0.000000	0.000000

fied the maximization of outputs. The results of the DEA analysis for employee number 1 show that he is only 82% efficient. That is given his output level (25 intakes and 10 visits), he should only be using 82% of the amount of his currently available inputs (40 hours and $105 dollars worth of supplies).

B. Weights Value

The value of the weights assigned to each input and output by DEA are shown under the column marked ''Value.'' As can be observed, the weight for the output variable ''Intake'' is .007143 and for the variable ''Visits'' the weight is .064286. If we were to substitute these weights into the objective function, the resulting value, as shown by the computer output, is .82.

The input variables show a different result. The weight of the input variable ''Hours'' is zero and its associated reduced cost is approximately 6.190476. This is explained below.

C. Reduced Cost

Consistent with the definition of reduced cost provided earlier, in order for the variable ''Hours'' to have a positive value, its corresponding coefficient must be reduced by approximately 6.2 units. That is, the amount of hours used must be reduced by at least 6.2 hours before this variable can obtain a positive weight which would result in an improved objective function or higher efficiency rating.

D. Slack/Surplus

In this column DEA shows the percent of inputs that can be considered a surplus. That is, the worker uses approximately 18% more inputs than he should given his output level. Notice that this percentage is the difference between 100%-efficiency rating ($100 - 82 = 18$).

E. Dual Prices

In this column. DEA shows the efficient reference set or peer group against which the particular employee is being compared. This simply means that DEA has identified a combination or composite employee which can obtain the same level of output by having available only a proportion of the inputs available to the employee under evaluation. The numbers under the dual price column represent the proportion of the input/output levels of the peer employee going into the composite set.

According to the computer solution, the composite employee derived by the model is made out of 41.7% of the input level used by employee number 3 and 25% of the input level used by employee number 5. Thus the composite employee uses 26.7 hours: $[(.417) \times (40)] + [(.25) \times (40)]$; and $86.3 dollars worth of supplies: $[(.417) \times (135)] + [(.25) \times (120)]$ to achieve the same level of outputs achieved by employee number 1. The supervisor could use these numbers as performance targets for employee 1.

If it is specified, the computer solution will also print a sensitivity analysis. Under this analysis, the computer shows the amount by which the coefficients of the objective function could increase/decrease without changing the solution. It also provides ranges for the right-hand-sides of the constraints which would not change the solution.

TABLE 4 DEA Efficiency Scores and Reference Set

DMU	Efficiency score	Reference set
EMP1	82.14%	EMP3, EMP5
EMP2	88.89%	EMP3
EMP3	100%	EMP3
EMP4	76.19%	EMP3
EMP5	100%	EMP5

Table 4 shows the efficiency ratings of all the employees along with their peer employees who make up the composite employee against whom the particular employee is being evaluated. The Table is analogous to one of the outputs that can be obtained by using a DEA software package like Warwick-DEA.

Notice that in the case of employees three and five, there are no reference sets of peers against which they are being compared. The reason is that those employees are efficient as shown by their efficiency rating of 100%.

IV. EVALUATING THE PERFORMANCE OF SCHOOLS

This example is an analysis of the 1996 High School Report Cards of a school district in New Jersey (*The New York Times*, 1996). The analysis is only applicable to high schools within the particular district. For purpose of illustration, the analysis in this example was conducted by means of Warwick-DEA software. A summary of the data used for this example is presented in Table 5 along with the efficiency scores and peer groups against which DEA compares each school. Only two inputs and two outputs were selected for inclusion in the analysis.

The input variables are:

1. Student/Faculty Ratio (S/F RATIO)
2. Spending per Pupil (SPENDING)

TABLE 5 Data for the Analysis of High Schools, Efficiency Rating (EFF), and Peer Schools (PEER)

	−SF Ratio	−Spending	+HSPT	+Grad	EFF	Peer
H1	13.5	8489	47.1	111.0	99.22%	H3
H2	13.9	7493	65.6	97.1	98.33%	H3
H3	9.8	7603	81.7	100.2	100%	H3
H4	10.9	12256	73.8	99.7	89.46%	H3
H5	10.1	12256	68.0	96.0	92.96%	H3
H6	11.0	8255	83.1	98.1	93.68%	H3
H7	10.4	10736	91.6	100.0	100%	H7
H8	12.4	87.3	44.2	101.2	88.23%	H3

The output variables are:

1. High School Proficiency Test (HSPT)
2. Graduation Rate (GRAD)

S/F RATIO is the ratio of students to full-time faculty members;
SPENDING is total school spending divided by the average daily enrollment:
HSPT is the percentage of high school juniors who passed all three sections of the high school proficiency test-reading, mathematics and writing; and
GRAD is the ratio of high schools seniors who graduated by August 1996 to enrollment in October 1995. Because some seniors do not graduate with their original class, percentages may add to more than 100.

Warwick-DEA does not require the analyst to elaborate the model. The analyst only has to input the data and give commands when prompted by the program. The data in Table 5 is formatted in a way that can be read by Warwick-DEA. Notice that inputs are preceded by a ''−'' sign and outputs by a ''+'' sign.

The DEA analysis identified 2 efficient high schools and 6 inefficient ones. Not shown in Table 4 are the weights calculated by Warwick-DEA, and the target performance for each school.

V. CONCLUDING REMARKS

When using DEA to evaluate efficiency, you must take into consideration that:

1. DEA is data sensitive. First, the number of outputs and inputs included in the analysis should not be too large in comparison with the number of units being evaluated. Second, DEA does not work well with missing data. Third, the accuracy of the data is instrumental to the analysis. The less accurate the data, the less accurate the results of the analysis. Forth, all relevant inputs and outputs should be specified. Lastly, because DEA is nonparametric, if one of the DMUs is taken out of the analysis or a new DMU is added, the solution to the previous analysis is no longer valid because the reference set has changed. Thus, a new linear programming model must be solved with the new number of DMUs.
2. If standard software is used to solve DEA models, the analyst will have to solve as many models as there are DMUs. That is, if the analysis consist of 54 DMUs, the analyst will have to formulate and solve 54 models. This tedious task can be avoided by using software designed specifically for solving DEA problems.

APPENDIX I MATHEMATICAL MODELS OF DATA ENVELOPMENT ANALYSIS

Model I

$$\max_{u,\,v} \frac{\sum_{r} u_r y_{ro}}{\sum_{i} v_i x_{io}}$$

subject to:

$$\frac{\sum_{r} u_r y_{rj}}{\sum_{i} v_i x_{ij}} \leq 1, \text{ for } j = 0, 1, \ldots, n$$

$$\frac{u_r}{\sum_i v_i x_{io}} \geq \varepsilon, \text{ for } r = 1, \ldots, s$$

$$\frac{v_i}{\sum_i v_i x_{io}} \geq 1, \text{ for } i = 1, \ldots, m$$

where:

ur = weight for output r
vi = weight for input i
Yrj = observed value of output r for DMU j′
xij′ = observed value of input i for DMU j′
ϵ = non-Archimedean infinitesimal appears in the primal objective function and as a lower bound for the multipliers in the dual. This constants value is often set at 10^{-6}

The next model, model 2, is the equivalent linear programming formulation of the fractional programming problem presented in model 1.

Model 2

$$\max_{u,\,v}: \omega_0 \sum_r u_r y_{ro}$$

subject to:

$$\sum_i v_{i\ xio} = 1$$

$$\sum_r u_r y_{rj} - \sum_i v_i x_{\ ij} \leq 0$$

$$u_r \geq \varepsilon$$

$$v_i \geq \varepsilon$$

The dual formulation for Model 2 is presented below:

Model 3

$$\min_{\theta,\,\lambda,\,s+r,\,s-1} = z_0 = \theta - \varepsilon \sum_r s_r^+ - \varepsilon \sum_i s_i^-$$

subject to:

$$\sum_j \lambda_j\ y_j - s^+ = y_0$$

$$\theta X_o = \sum_j \lambda_j X_j - s^- = 0$$

$$\lambda, s_r^+,\ s_i^- \geq 0$$

θ = a scalar variable is the proportional reduction applied to all inputs of DMU_0 (the DMU being evaluated) to improve efficiency
s^+, s^- = slack variables

NOTES

1. For examples of the application of DEA in education. see: A. M. Bessent, and E. Bessent, *supra*: A. M. Bessent, E. Bessent, T. C. Clark, and A. W. Garrett, ''Managerial Efficiency Measurement in School Administration,'' *National Forum of Educational Administration and Supervision Journal*, *3*: 56–66, 1987; A. Desai and A. P. Schinnar, ''Technical Issues in Measuring Scholastic Improvement due to Compensatory Education Programs,'' *Socio-Economic Planning Sciences*, *24*: 143–153, 1990; W. Ludwin and T. Guthrie, *supra*: and, A. Charnes, W. W. Cooper, and E. Rodhes, 1981, *supra*. For examples using DEA in hospital and physician evaluation see: J. A. Chilingeerian, ''Exploring Why Some Physicians' Hospital Practices Are More Efficient: Taking DEA Inside the Hospital,'' in A. Charnes, W. W. Cooper, A. Y. Lewin, L. M. Seiford (eds.), *Data Envelopment Analysis: Theory, Methodology, and Applications*, Kluwer Academic Publishers: Boston, 1995; and H. D. Sherman, *supra*. In court systems see: A. Y. Lewin, R. C. Morey, and T. J. Cook, ''Evaluating the Administrative Efficiency of Courts,'' *Qmega*, *10*: 401–411, 1982. In nursing services see: S. K. Chattopadhyay. ''Economics of Nursing Home Care in Connecticut: Financing, Cost and Efficiency,'' Ph.D. dissertation, University of Connecticut: Stross, CT, 1991; and, T.R. Nunameker, *supra*. For DEA evaluating the efficiency of highway maintenance see the work of: W. Cook, A. Kazakov, and Roll, J., ''On the Measurement and Monitoring of Relative Efficiency of Highway Maintenance Patrols,'' in A. Charnes, W. W. Cooper, A. Y. Lewin, L. M. Seiford (eds.), *Data Envelopment Analysis: Theory, Methodology, and Applications*, Klwer Academic Publishers: Boston, 1995; and W. Cook, Roll, J., and A. Kazakov, ''A DEA Model for Measuring the Relative Efficiency of Highway Maintenance Patrols, *INFOR*, *28*: 113–124, 1990. in *Banking*, see G.D. Ferrier, and C. A. Lovell. ''Measuring Cost Efficiency in Banking: Econometrics and Linear Programming Evidence.'' *Journal of Econometrics*, *46*(1/2) 229–245, 1990; and, D. I. Giokas, ''Bank Branch Operating Efficiency: A Comparative Application of DEA and the Loglinear Modeal,'' *Omega*, *19*: 199.
2. See Charnes, et al., 1995, *supra*: R. Thompson, P.S. Dharmapala, and R. Thrall, ''Sensitivity Analysis of Efficiency Measures with Applications to Kansas Farming and Illinois Coal Mining''; D. Day, A. Y. Lewin, H. Li, and R. Salazar, ''Strategic Leaders in the U.S. Brewing Industry: A Longitudinal Analysis of Outliers''; and A. Charnes, W. W. Cooper, B. Golany, D. B. Learner, F. Y. Phyllips, and J. J. Rousseau, ''A Multiperiod Analysis of Market Segments and Brand Efficiency in the Comparative Carbonated Beverage Industry.''

REFERENCES

Anderson, D., Sweeney, D., and T. Williams (1991). *An Introduction to Management Science*, West Publishing Company: St. Paul

Banker, R. D. and R. Morey (1986). ''The Use of Categorical Variables in Data Envelopment Analysis,'' *Management Science*, *32*(12): 1613–1627.

Bessent, A. M. and E. W. Bessent (1980). ''Determining the Comparative Efficiency of Schools Through Data Envelopment Analysis,'' *Educational Administration Quarterly*, *16*(2): 57–75.

Banker, R. and R. Morey (1986). ''The Use of Categorical Variables in Data Envelopment Analysis,'' *Management Science*, (32): 1613–1627.

Charnes, A., W. Cooper, and E. Rhodes (1978). ''Measuring the Efficiency of Decision Making Units,'' *European Journal of Operational Research*, *2*: 429–444.

Charnes, A., W. Cooper, and E. Rhodes (1981). ''Evaluating Program and Managerial Efficiency: An

Application of Data Envelopment Analysis to Program Follow Through,'' *Management Science*, *27*: 668–697.

Charnes, A., W. Cooper, A. Lewin, and L. Seiford (1994). *Data Envelopment Analysis, Theory, Methodology and Applications*, Kluwer Academic Publishers: Boston.

Diewert, W. and M. Mendoza (1995). ''Data Envelopment Analysis: A Practical Alternative?'' Department of Economics, University of British Columbia. Vancouver. Discussion Paper No: DP 95–30.

Knox Lovell, C. A., L. C. Walters, and L. L. Wood (1994). ''Stratified Models of Education Production Using Modified DEA and Regression Analysis,'' in A. Charnes, W. Cooper, A. Y. Lewin, and L. Seifored (eds), *Data Envelopment Analysis, Theory, Methodology and Applications*, Kluwer Academic Publishers: Boston (pp: 329–352).

Lewin, A. Y., R. C. Morey, and T. J. Cook (1982). ''Evaluating the Administrative Efficiency of Courts,'' *Omega*, *10*: 401–411.

Ludwin, W. G. and T. L. Guthrie (1989). ''Assessing Productivity with Data Envelopment Analysis'' *Public Productivity Review*, *12*(4): 361–372.

The New York Times, New Jersey Section (1996). Comparing the Districts: The '96 High School Report Cards, December, 8. pp. 8, 10.

Nunamaker, T. (1983). ''Measuring Routine Nursing Service Efficiency: A Comparison of Cost Per Patient Day and Data Envelopment Analysis Models,'' *Health Services Research*, *18*: 183–205.

Sherman, H. D. (1988). *Service Organization Productivity Management*, The Society of Management Accountants of Canada: Hamilton, Ontario.

Sherman, H. D. (1984). ''Hospital Efficiency Measurement and Evaluation. Empirical Test of a New Technique,'' *Medical Care*, *22*(10): 922–935.

Tankersley, W. (1990). *The Effects of Organizational Control Structure and Process on Organizational Performance*, Ph.D. Dissertation, Florida State University: Tallahassee.

Tseng, M. (1990). *Efficiency Comparison of Nursing Homes: An Application of Data Envelopment Analysis*, Ph.D. Dissertation, University of Alabama: Birmingham.

Watson, C., P. Billingsley, D. Croft, and D. Huntsberger (1993). *Statistics for Management and Economics*, 5th edition, Allyn and Bacon: Boston.

Warwick-DEA. Warwick Business School: Coventry CV47AL, UK

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

23

Principal Component Analysis, Factor Analysis, and Cluster Analysis

George Julnes
University of Illinois at Springfield, Springfield, Illinois

I. INTRODUCTION

Quantitative analysis is presented in this book as a tool capable of guiding more effective action for managers and policymakers. Such a use presumes that there are relationships that can be appreciated when the available information is organized properly. Recognizing these relationships requires that we first have a way of differentiating the complex world of administration into meaningful elements or dimensions (Rummel, 1970). In this chapter we address three alternative procedures for organizing the phenomena that are of concern to administrators: principal components analysis, factor analysis, and cluster analysis. Unlike techniques such as multiple regression and discriminant analysis that seek to establish the dependence of one set of variables on another, the three multivariate techniques discussed in this chapter seek to reveal interdependencies among variables (Dillon and Goldstein, 1984).

In order to be useful for a broad range of public administration scholars and practitioners, each of these three techniques is presented first in terms of its major conceptual foundations and then applied to data taken from an evaluation of a program for pregnant teens. The goal is to provide an overview for the readers that will allow you to use these techniques and interpret the results. A fuller understanding, however, will require additional readings, and so throughout this chapter recommendations are made regarding particularly useful sources for specific topics. Of general use, however, are many texts on multivariate analysis, with Dillon and Goldstein (1984) requiring less of a background in mathematics and Morrison (1990) building from a foundation of matrix algebra, and scholarly journals that emphasize multivariate techniques (e.g., *Multivariate Behavioral Research*, *Biometrika*, and *Sociological Methods and Research*). The remainder of this introductory section is devoted to: (1) distinguishing relevant approaches to organizing phenomena, (2) describing the Resource Mothers Program for pregnant teens that will be used to exemplify the implications of using the different techniques, and (3) introducing the computer package used in analyzing the data.

A. Choices in Organizing Phenomena

To claim that studying public administration can improve public administration—for example, the claims that restructuring government can lead to more efficient service provision or that examination of best practices can lead to more effective management—is to claim that there

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

are relationships that are sufficiently enduring and reliable as to be useful in guiding action. The quantitative techniques presented below support our understanding of these relationships by helping us organize our worlds in meaningful ways. Choosing a technique that is appropriate for your particular needs requires considering more issues than can be summarized here. We can, however, understand some of the more fundamental distinctions among principal components analysis, factor analysis, and cluster analysis by considering the conceptual framework presented in Table 1. Each of the columns represents a basic issue in methodology with some of the alternative positions listed below. We introduce these issues here as choices in organizing phenomena and return to this framework in the concluding section of this chapter.

1. Focus of Analysis: Objects, Attributes, and Occurrences

One of the first decisions faced by a researcher using quantitative methods concerns just what it is that the researcher wants to analyze. This decision is often between efforts to organize different types of people (types of ''objects'') or different groupings of variables (types of attributes or characteristics); one may, however, choose instead to analyze different occasions according to their similarity. Recognizing that researchers might be interested in any one or in all three of these emphases, Dillon and Goldstein (1984) note that for multivariate analysis, ''the basic input can be visualized in terms of a data cube with entries denoted by X_{ijk}, where *i* refers to objects, *j* refers to attributes, and *k* refers to occasions or time periods'' (p. 3).

The concept of a data cube may be new to some and seem complicated, but the basic idea is fairly simple. We can think of measurement in terms of a cube that has three dimensions, objects, attributes, and occasions, but we cannot organize all three of these dimensions at once. Instead, we take ''slices'' of this three-dimensional cube and organize the information in those slices: (1) relationships among objects; (2) relationships among attributes; or (3) relationships among occasions. If we wish to identify types of managers, then we are interested in classifying people into groups. The managers being classified are examples of individual ''objects''; other examples of objects might include distinguishing different types of organizations or even classifying office supplies into different budget categories. Alternatively, focusing on ''attributes,'' one might be interested in identifying the performance measures that covary and are associated with different long-term outcomes. This emphasis would lead to classifying the various performance measures into groups of variables. Another example of this approach would be to group measures of fiscal stress into categories (e.g., spending obligations versus fiscal capacity). Finally, if we wished to group different times into categories, we would be analyzing ''occasions.''

TABLE 1 Choices in Organizing Phenomena

Focus of analysis	Scale of measurement	Goal of analysis
Objects: Identify similarities among people, agencies, or other concrete entities	Categorical: Discrete groupings that rely on nominal measurement	Nominalism: Derived organization of phenomena is convenient fiction
Attributes: Identify similarities among characteristics being measured	Dimensional continuous phenomena that can be organized using ordinal, interval or ratio scales	Realism: Natural categories and dimensions that can be approximated through analysis
Occasions: Identify similarities among periods of time being studied		

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

With this focus, organizational performance in the public sector might be analyzed by grouping the available information in terms of changes in elected administration. These examples of the three dimensions of the data cube introduce the idea; in the concluding section of this chapter we develop the data cube notion further by depicting some of the slices of this data cube that are particularly relevant in public administration research.

Each of the techniques of this chapter can be used to organize any slice of the data cube described by Dillon and Goldstein (1984). Organizing attributes that are measured across multiple objects, as in organizing personal characteristics of many people, is referred to as R-analysis; grouping objects together based on their attributes is referred to as Q-analysis; and organizing occasions based on multiple attributes of one object is known as O-analysis (for an introduction to these and the three other slices of the data cube, see Rummel, 1970). Although tradition links certain techniques with these different slices of the data cube, and so these techniques are most developed for these slices, it is up to the investigator to understand which facet of the cube is of greatest relevance to the research questions at hand.

2. *Scale of Measurement: Categories Versus Dimensions*

In addition to distinguishing among objects, attributes, and occasions when organizing phenomena, one must also identify the scale of measurement desired for the results. Measurement is often presented in terms of nominal, ordinal, interval, and ratio scales. The first of these scales, nominal, is categorical; the last three, particularly interval and ratio scales, presume dimensions. Although the three techniques to be addressed generally presume interval data as input (though dichotomous and ordinal data can be used for some purposes), the techniques can be differentiated in terms of their output. Cluster analysis provides categorical groupings (nominal scale) as output, while principal components analysis and factor analysis produce interval dimensions as output.

Both categories and dimensions can be useful in organizing the domain of public administration. We employ categories when we contend that specific types of management are most appropriate for particular business environments (Daft and Weick, 1984). Similarly, we might seek to understand the problems of government by first identifying categories to differentiate types of governmental waste (Stanbury and Thompson, 1995). These types, whether viewed as Weberian ideal types or as empirical groupings, represent claims that it is meaningful to classify styles, individuals, and situations into categories (Bailey, 1994).

As an example of ordering administrative phenomena along dimensions, one might talk of a dimension of ''publicness'' and claim that, rather than distinguish public and private organizations as representing two discrete categories, organizations can be ordered along a continuous dimension on which most organizations are more ''public'' than some but less than others (Coursey and Bozeman, 1990). Similarly, we can conceive of most managerial initiatives along a continuous dimension of implementation and claim that the greater the implementation of a particular program, the greater the program effects. Both of these examples of dimensions are presented as if there were a reasonably continuous underlying phenomenon that is represented best by a continuous variable.

3. *Goal of Analysis: Realism Versus Nominalism*

A final issue of analysis to be considered here concerns the beliefs one has about the proper interpretation of observed interdependence. On the one hand are those who believe that the categories and dimensions used in analysis refer to the real structure of the world (see Julnes and Mark, in press); on the other hand are those who believe that the identified categories and dimensions are simply ''useful fictions'' that facilitate discussions and simplify decisions but

do not refer to anything real about the world. Those in the former group, the realists, would believe, for example, that there really are consistent differences among leaders (whether as different types or in terms of different characteristics that vary along meaningful dimensions) that can be captured more or less effectively by the analysis. Those in the second camp, the nominalists, might accept that leaders differ in important ways but would not believe that these differences are in any sense systematic.

As we will see below, factor analysis presumes that there are underlying factors that are responsible for observed regularities. Principal components analysis takes no stance on this issue and so can be used even by those who take a nominalist view of the world. Cluster analysis can be differentiated into two types in terms of this dimension, with most varieties presuming underlying groupings that are to be revealed by analysis but also some that view the categories produced by the analyses as merely useful and nothing more (Bailey, 1994).

The point to be made in thinking about these issues is that using the techniques presented below requires, and presumes, prior decisions reflecting beliefs and intentions that are to guide the analysis—a computer can analyze data for each of these possible choices. In the remainder of this introduction we first describe the Resource Mothers Program that serves as a backdrop for this discussion of multivariate techniques and the computer program being used for analysis.

B. Resource Mothers Program

The Resource Mothers Program is a lay home visiting program that emerged from the Southern Governor's Task Force on Infant Mortality. In Virginia, the Resource Mothers Program began in 1985 in three metropolitan areas and by 1996 had grown to 20 programs throughout the Commonwealth. Focusing on unacceptably high rates of infant deaths, one of the goals of the program has been to reach out to women who are at-risk for negative birth outcomes and who are not reached by traditional prenatal programs (Julnes et al., 1994). The primary program activities involve the lay home visitor, a woman referred to as a Resource Mother, meeting with the client, usually an at-risk pregnant adolescent, and arranging prenatal medical visits, monitoring and advising on nutrition, and acting as a liaison between the client and other agencies.

In an effort to understand the impact of the program, the March of Dimes Birth Defect Foundation sponsored a two-year, multisite evaluation in Virginia. While much of the evaluation involved qualitative case studies and comparisons of five project sites, quantitative analysis was used to estimate program impact (e.g., logistic regression allowed estimation of the reduction in low birthweight deliveries due to the program) and, using information on program costs and estimated benefits, net economic value. These analyses were based on birth certificate data from 34,104 births, including births from 196 program clients. Table 2 presents some of the variables used in the quantitative analysis.

For the purposes of this discussion of classification methodologies, we will focus on the characteristics of the 196 clients. With this focus, the three techniques described in this chapter—principal components analysis, factor analysis, and cluster analysis—will be examined in terms of their ability to address the relationships to be found in 196 observations measured on some combination of the eleven variables. For each technique we will provide a short introduction to the essential concepts of the technique and then use the results of the analysis of the Resource Mothers Program client birth data to introduce the other important points.

C. Computer Analysis of Interdependencies Among Variables

As mentioned previously, the goal of this chapter is to present quantitative techniques that support scholars and practitioners in public administration in their efforts to make sense of the complex realities that they face. In order for the techniques discussed to be useful, we need to

TABLE 2 Selected Variables from Resource Mothers Program

Birthorder	0 for no prior births, 1 for one prior birth, 2 for two prior births, etc.
Ethnicity	1 for African-Americans, 0 for other.
Marital status	0 for single, 1 for married.
Months of prenatal care	Months of prenatal medical care before delivery, ranged from 0 to 9.
Mother's age	Age in years at last birthday before current birth.
Mother's education	0–12 for elementary school through high school, 13–16 for one through four years of college, 17 for more than college degree.
Source of prenatal medical care	1 for private physician, 0 for other sources of primary prenatal care.
Medical prenatal visits	Number of visits for medical prenatal care.
Weight gain	Increase in weight, in pounds, during pregnancy (weight loss coded as 0).
Gestational age	Physician estimate of weeks of pregnancy before giving birth.
Birthweight	Baby's weight, in grams, at birth.

make sure that they are accessible to those with the appropriate computer resources available. Although there are several highly regarded computer programs available for specific multivariate techniques, most scholars and researchers in public administration make use of one or more of the available computer packages (e.g., BMDP, SPSS, and SAS). For the purpose of this chapter, we will use the SAS computer program (SAS Institute, 1988). This program provides a wide array of multivariate analysis techniques and is available for both mainframe computers and for personal computers. The program can be run either interactively or as a batch program (series of commands run together); for simplicity in presentation without the use of computer screens, we will limit ourselves to discussing the batch commands.

When used in the batch mode, a SAS program can be viewed as consisting of two steps. First is the DATA step of the program in which the data to be analyzed are identified and read as observations on specified variables. Also in the DATA step are operations that change the data before they are analyzed. For example, the researcher might wish to transform the data or even create new variables based on the inputted data. The second step is the PROC step in which the data are analyzed using SAS procedure statements. It is these procedure statements, referred to as PROC statements, and the associated supplemental statements that control the data analysis, that constitute the focus of the computer programming statements made in this chapter. For each of the quantitative techniques discussed we will present the SAS commands for the statistical procedure in a table and discuss them in the text. The reader wishing to learn more about the programming options available in SAS is directed to the SAS manuals but also to user-friendly books on such topics as social science inquiry (Spector, 1993), multivariate and univariate analysis (Hatcher and Stepanski, 1994), and factor analysis and principal components analysis (Hatcher, 1994).

II. PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis is a type of factor analysis that emerged from the work of Pearson (1901) and Hotelling (1933). As a result of this early development, principal components analy-

sis has its roots among the earliest attempts to classify phenomena using quantitative methods. ''Its goal is to reduce the dimensionality of the original data set. A small set of uncorrelated variables is much easier to understand and use in further analyses than a larger set of correlated variables'' (Dunteman, 1989, p. 7). The analysis that follows makes use of books by Rummel (1970) and Jolliffe (1986) and the monograph on principal components analysis by Dunteman (1989).

A. Conceptual Foundation

1. Underlying Logic: Data Reduction

The basic logic of principal component analysis is straightforward, and its purpose can be summed up in a word: parsimony. The idea is to account for the information provided by many variables (or observations) by using a more limited set of constructed dimensions that are effective substitutes for the variables (Dillon and Goldstein, 1984). The only way to achieve this goal is to create composites of the many variables that retain much of the information contained in the original variables. For example, if one had 200 organizations measured on 100 variables addressing organizational characteristics, it would help in making sense of the organizations if one could reduce, without losing important information, the 100 variables to something like a dozen or fewer composite variables. Not only would such a reduced set of variables allow clearer informal comparisons across organizations, it also would improve many of the subsequent statistical analyses used to establish relationships between organizational characteristics and various outcome measures.

2. Quantitative Model

''Principal components analysis searches for a few uncorrelated linear combinations of the original variables that capture most of the information in the original variables'' (Dunteman, 1989, p. 10). By ''capturing the most information'' we mean creating a new variable that accounts for the maximal amount of variance of the original variables. Principal components analysis attempts this task of reduction by calculating a linear combination of the original variables as indicated in Formula 1. This formula is presented to emphasize that the principal components, PCs, are viewed as linear combinations, with weights represented as $w_{(i)}$, of the original variables, Xp. The first principal component derived is, as indicated, the one that accounts for the greatest total variance for all variables being analyzed.

$$PC_{(i)} = w_{(i)1} X_1 + w_{(i)2} X_2 + \ldots + w_{(i)j} X_j \quad (1)$$

Once the first principal component is calculated in this way, the second principal component is calculated to maximize the remaining variance (that not accounted for by the first component) with the constraint that it be orthogonal to the first component. By ''orthogonal'' we mean statistically independent such that variation on one principal component is not related to variation on any other principal component. The correlation between orthogonal dimensions is, therefore, zero. This procedure of finding orthogonal dimensions could be continued to generate as many principal components as there are variables; doing so will provide a progressive accounting of the total variance contained in the variables. In interest of parsimony, however, far fewer principal components are typically used.

3. Graphic Representation

The logic and quantitative model described above can be given geometric form and thus displayed graphically. Figure 1 presents the simplest case in which there are two standardized

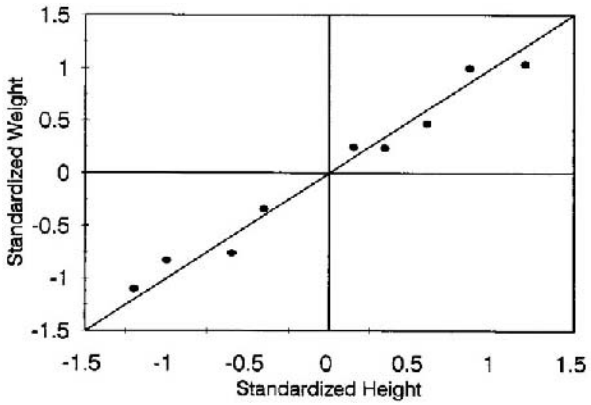


FIGURE 1 Principal components analysis with correlated variables.

variables—height and weight—and one principal component being used to account for the information of the two variables. Note that in this case the scatterplot is narrow and tightly delimited, indicating a high correlation between the two variables. As a result, it is possible to establish a dimension, the first principal component, that accounts for almost all of the information provided by the two variables. Therefore, knowing the position on the first principal component would allow effective prediction of values on both of the original variables. In this case, therefore, the two original variables of height and weight can be replaced by a single ''size'' factor with little loss of information.

In contrast, Figure 2 depicts two variables, ''years of training'' and ''job satisfaction'' that are only moderately correlated. As before, principal components analysis will generate a dimension that accounts for the most variance in the data, but this time the first principal component leaves considerable variation unexplained.

A third point to be made from looking at Figures 1 and 2 refers to the different balance of dimensions represented in the two figures. Whereas the principal component represented in Figure 1 is drawn to indicate an equal weighting of the height and weight variables (allowing for the scaling of the figure axes), the line drawn in Figure 2 is angled to be much closer to the axis representing ''years of training.'' The reason for this is that the variables in Figure 1 are standardized and so have the same variance. In contrast, the variables represented in Figure

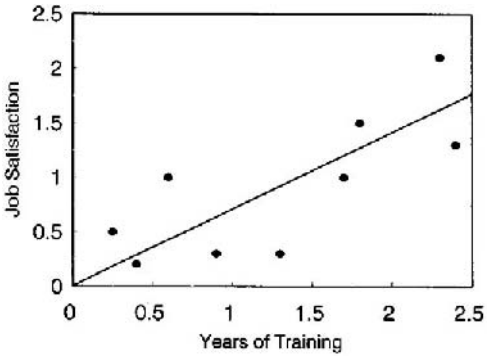


FIGURE 2 Principal components analysis with moderate correlation.

2 are not standardized—the greater variance of the ''years of training'' variable (assuming for the sake of illustration that measured job satisfaction varies little in this case) results in the principal component being weighted to account for this greater variance.

The differences in results due to standardizing the variables can be substantial, but there is no definitive answer to the question of whether or not it is better to standardize variables. In some cases you may want certain variables to carry stronger weight in accordance with their variance; other times it will seem preferable to allow the implicit weighting that results from standardization to replace the variance-based weighting. We will rely below on standardized variables, but the point is that each researcher must make this decision based on the situation being confronted by the analysis.

B. Application to Resource Mothers Program

We have presented the basic concepts of principal components analysis and have introduced its quantitative model. We can now use our example of the Resource Mothers Program to address additional issues that need to be considered by users. The first issue to address involves choosing the number of principal component dimensions to be included in one's analysis; a second point concerns the possible interpretation of the derived dimensions; and a third issue is the use of the results of principal components analysis as new variables used in subsequent analyses.

The results reported below were produced by using the PRINCOMP procedure in SAS as presented in the left side of Table 3 (note, in batch mode it is essential to end commands with a semicolon). As indicated in the right side of Table 3, the commands direct the program to perform a principal components analysis that yields, at most, a two-component solution (N = 2) using the variables listed in the variable statement (VAR).

1. Determining the Number of Principal Components

Choosing the number of principal components to include in the analysis is not objective and represents a tension between parsimony and retention of the information contained in the original variables. The two examples graphed above make clear that only one principal component is needed in Figure 1 while two appear more adequate for the data in Figure 2. Many situations, however, are not that clear, and so more formal procedures have been developed as guides.

To understand these procedures we must first introduce the concept of the eigenvalue, also called the latent root or characteristic root. We mentioned above that principal components are selected to account for the maximal variance of all analyzed variables; the first component will account for the greatest amount of variance, with the second, third, and subsequent components accounting for progressively less variance. The amount of variance captured by a component is conveyed by its eigenvalue (a value defined in terms of matrix algebra and discussed in greater detail when discussing factor analysis), and, taken together, the eigenvalues for the components support several approaches to guide the decision of how many principal components should be retained for subsequent analyses.

TABLE 3 SAS Computer Commands for Principal Components Analysis

Computer commands	Function of commands
PROC PRINCOMP N = 2;	N = 2 specifies retaining two components.
VAR ⟨variables⟩;	Specifies the variables to be analyzed.

In cases where one is using unstandardized data such as raw responses, the number of principal components can be assessed by testing whether the eigenvalues for subsequent principal components are significantly different from each other. The logic of this analysis, based on Bartlett's approximate chi-square statistic, is that the eigenvalues for the meaningful components drop off progressively but then plateau for the subsequent less meaningful principal components (Dillon and Goldstein, 1984). One difficulty with this test is that it typically results in retaining more principal components than researchers would want when pursuing the goal of parsimony. Further, as illustrated above in Figure 2, use of unstandardized variables results in a greater influence of the variables with the greatest variance, a property that means that those variables will be better represented by the derived principal components than will the variables with less variance. In that this influence on the outcomes is generally undesirable, we will focus more attention on determining the number of components when using standardized variables.

In considering standardized variables, recall that the variance of these variables typically is set to equal 1.0. As such, principal components with eigenvalues greater than 1.0 are accounting for more variance than any of the original variables. With this in mind, Kaiser (1958) advocated retaining only those principal components with eigenvalues greater than 1.0. The logic of this procedure being that principal components with eigenvalues less than 1.0 are not contributing to the goal of parsimony. Applying this logic to the analysis of the Resource Mothers Program, we see in Table 4 that three principal components have eigenvalues greater than 1.0.

As an alternative to strict quantitative determination, Cattell (1966) developed the scree test to identify qualitative changes in the ability of principal components to account for variance. The name of the scree test comes from the shape of a cliff or mountainside. At the base of a steep cliff is likely to be considerable rubble of fallen stones, a rubble referred to as scree. This accumulated rubble will slope downward away from the cliff but at a different slope than the cliff itself. Using this analogy, the slope of the meaningful principal components as measured by change in eigenvalues can be differentiated from the slope of the noise factors that might otherwise be retained as principal components. The intent is to be guided by the changes in the eigenvalues rather than their actual values. Figure 3 illustrates this logic by showing that the slope between the first and second components is steep compared to the slopes between subsequent components. Indeed, the slopes are so similar from the second to the fifth principal components that they can be said to define a straight line. This would suggest that the second component and those higher represent the scree at the base of the real structure.

Unfortunately, the scree test often is an ambiguous guide for several reasons. First, Cattell himself had mixed thoughts on the proper interpretation. In the context of factor analysis, ''Cat-

TABLE 4 Eigenvalues and Variance Accounted for By Principal Components

Principal component	Eigenvalues	Proportion of variance	Cumulative variance
1	2.148	0.358	0.358
2	1.336	0.223	0.581
3	1.021	0.170	0.751
4	0.744	0.124	0.875
5	0.465	0.077	0.952
6	0.286	0.048	1.000

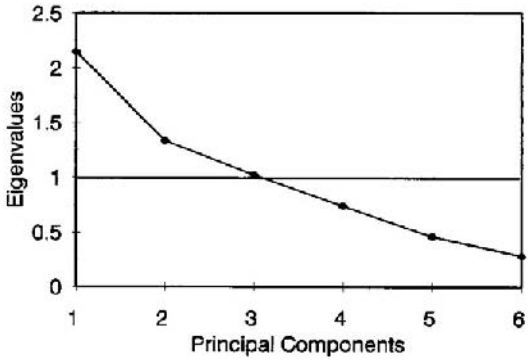


FIGURE 3 Scree test for determining number of components retained.

tell originally suggested taking the first factor on the straight line as the number of factors to be sure that sufficient factors were extracted; he has since (Cattell and Jaspers, 1967) suggested that the number of factors be taken as the number immediately before the straight line begins'' (Gorsuch, 1983, p. 167). In the present case, the first component on the straight line would be the second component, the one immediately before the straight line would be the first principal component. The logic of Cattell's revised interpretation is the desire to maximize the ratio of variance accounted for over the number of components.

A second source of ambiguity for the scree test is the lack of objective standards in what constitutes a break in the straight line of the scree. As a result, small changes in eigenvalues for the principal components (e.g., the eigenvalue for the second principal component being only slightly higher) could change perceptions of where the straight line in Figure 3 had begun. Cattell and Vogelmann (1977) provide greater elaboration for those wishing to use the scree test in accordance with Cattell's intended logic.

Concerns about the scree test, however, go beyond the issue of ambiguity. In particular, as reasonable as the scree test appears as a guide to avoid interpretation of trivial principal components, it does not necessarily serve the needs of particular research situations. Hatcher (1994) describes two other criteria for determining the number of principal components to retain in the analysis. Returning to quantitative assessment, one option is to focus directly on the proportion of the variance accounted for by the principal components, either individually, in which one might choose to retain any component that accounts for at least 10% of the variance of the variables, or cumulatively, in which one might retain enough principal components to account for at least 70%, or perhaps 80%, of the variance of the variables. In the present example, these criteria would argue for retaining three or four principal components (refer back to Table 4). As such, it may be that a two-dimensional solution does not account adequately for one or more of the variables and so a three-dimensional or higher solution might be preferred.

The second option proposed by Hatcher (1994), a more qualitative one, is to determine the number of principal components to retain by using the criterion of interpretability. Analyses with a given number of principal components might yield results that are particularly consistent with the results reported by previous scholars. For the sake of illustration and for comparison with factor analysis, we will present our results in terms of a two-dimensional solution to the principal components analysis. This decision to retain two principal components is seen in the N = 2 option included in the PROC PRINCOMP statement provided in Table 3.

2. Component Interpretation

Having reduced the six variables to two principal components, it is natural to attempt interpretation of these dimensions. In approaching this task, the variables with the strongest relationships with the principal components are important in defining a principal component. Table 5 presents the correlations between the six variables and the two principal components. By looking at each variable row, we can identify the highest correlations (in absolute terms) for each variable. For example, Months of Care is more closely related to the second principal component ($r = 0.52$) than to the first ($r = 0.32$), while Age of Mother is more closely related to the first principal component ($r = 0.58$) than to the second ($r = -0.26$).

Repeating this examination for each variable, we pay particular attention to those coefficients that are large. No set definition exists for what constitutes ''large,'' but many researchers require interpretable loadings to have correlation coefficients of at least 0.35 or 0.40 (absolute value). Thus, we can see that the ''large'' correlations for the first principal component are 0.58 for Age of Mother, 0.53 for Mother's Education, and 0.43 for Marital Status. In that each of these three variables increases with age, this component is concerned with the personal maturation that is associated with increasing age. The second principal component has its highest correlations with the number of Prenatal Visits (0.68) and Months of Medical Care (0.52), suggesting a dimension of health activities. Note that the variable measuring the weight gained during pregnancy loads highest on the health activity component but does not have a large loading on either of the two principal components. As a result, the variance of Weight Gain accounted for by the two factors is low, only 9% (as will be pointed out below, this lack of fit for Weight Gained and, to a lesser extent, Marital Status will help differentiate principal components analysis from factor analysis).

Finally, at the bottom of Table 5 there is a row labeled ''component explained.'' This row is not provided when principal components are reported but is presented here to highlight the logic of principal components analysis: if you square each of the six correlations reported for each component and sum them, you get 1.0. This is another way of saying that all of the variance of the principal components is accounted for by the six variables, something that we know by definition in that a principal component is nothing other than a linear combination of the variables being analyzed.

If one wishes to confirm the interpretations suggested by the high loadings, one can explore the relationships between these derived principal components and other variables not used in the original analysis. To the extent that these subsequent analyses produce results consistent with the interpretations given to the principal components, these interpretations are supported. It should be noted, however, that if the goal of analysis is simply parsimony, interpretation of

TABLE 5 Correlations between Variables and Components

	First principal component	Second principal component	Variable variance accounted for
Months of care	0.32	0.52	37%
Prenatal visits	0.28	0.68	53%
Weight gain	0.14	0.27	9%
Age of mother	0.58	−0.26	40%
Mother's education	0.53	−0.31	38%
Marital status	0.43	−0.18	22%
Component explained	1.00	1.00	

the derived dimensions may not be as important for principal component analysis as we will see it to be for factor analysis.

3. Use of Component Scores

Once the computer has solved for the weights for Formula 1, we can use that formula to calculate the appropriate scores for each observation on the principal components. Doing so allows us to use the derived principal components to predict the birthweights of the babies born to the teen mothers. It turns out that the two principal components as independent variables account for only half of the variation in birthweight that is accounted for by the six original variables, but, because of the fewer explanatory variables in the analysis, the value of the F test statistic for the overall regression model is higher. Of particular note, the use of the two orthogonal principal components as predictors clarifies the greater impact of the health activity dimension on birthweight than was found for the personal maturation dimension, something that the multicollinearity (the high level of multiple correlation among the explanatory variables) of the original variables obscured when they were used as independent variables.

C. Summary Assessment of Principal Components Analysis

We have presented the conceptual foundations of principal components analysis and demonstrated its use with data from the Resource Mothers Project. We now need to offer a summary assessment of its key features. First, using the choices in organizing phenomena that were presented at the beginning of this chapter, principal components analysis is traditionally used to identify dimensions that capture the information contained in variables. That is, the focus is on organizing information about attributes (the variables), measured across objects (such as people), into new continuous dimensions. While such dimensions might refer to the underlying structures posited by realism, principal components analysis makes no such requirement.

1. Strengths

The strengths of principal components analysis follow from the simplicity of the way that it seeks parsimony. By definition, the derived principal components account for a maximal amount of variance in the variables measured using orthogonal dimensions. As mentioned above, the emphasis on accounting for maximal variance is important when attempting to reduce many explanatory variables to a few explanatory factors. This reduction helps avoid the problem of capitalizing on chance associations when using many explanatory variables. A related virtue of simplicity is that principal components analysis does not require the data to have a particular distribution.

The fact that the technique produces orthogonal principal components is valuable in avoiding multicollinearity in subsequent analyses. This strategy for reducing variables and thus minimizing multicollinearity is particularly useful when your explanatory model includes interaction terms based on multiplying the explanatory variables together. An example of this might be when we are interested in how the impact of health activities varies across different client characteristics, such as client age; the interaction variables are almost certainly related to the variables that were multiplied to create them. Principal components analysis does not result in the interaction terms being orthogonal to the principal components, but if the original variables are highly correlated themselves, the problem of multicollinearity with interactions becomes substantially worse.

This use of principal components analysis to generate weighted composite variables for use in subsequent analyses highlights an additional strength of the approach. In contrast to factor

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

analysis as described below, principal components analysis provides an objective rationale for calculating scores based on the original variables. The dimensions that account for the most variance are derived and the weights for the variables appropriate to support these dimensions are used to calculate the scores for each observation.

2. Concerns

The source of the strength of principal components analysis is also the cause for concern in its use. Principal components analysis provides a vehicle for accounting for the total variance of a set of variables. Part of its simplicity is that it takes the variance as it is defined by the inputted data, without performing any transformation of the variance. As such, one concern with principal components analysis is that it is not invariant with regard to scaling decisions involved in measuring the variables to be analyzed. The greater the variance of one variable relative to the others, the more influence it will have on the direction of the principal component. Thus, when using unstandardized data, one could use scaling decisions to influence the results of the analysis.

A second, and more serious, implication of this focus on total variance is that the method does not differentiate between meaningful variance and variance that may be the result of measurement error. This lack of differentiation is appropriate for the goal of principal components analysis but represents a limitation if one wishes the derived principal components to correspond to real phenomena. As we will see when discussing factor analysis, it is the covariance, rather than the variance, that is of interest when we believe that there are underlying factors that are responsible for the patterns observed in the measured variables.

III. FACTOR ANALYSIS

Factor analysis, sometimes more precisely referred to as common factor analysis, was developed by Spearman (1904) in the context of investigating general and specific intellectual abilities and later elaborated by L.L. Thurstone (1935). As such, factor analysis emerged at about the same time as principal components analysis and shares with it many features. Nonetheless, factor analysis differs from principal components analysis in fundamental ways that need to be appreciated by those who use either technique. Numerous books address factor analysis, both in terms of the mathematical foundations (e.g., Morrison, 1990) and in terms of an appreciation of the logic underlying this method (see Gorsuch, 1983; Kim and Mueller, 1978a; Rummel, 1970). We develop this appreciation by examining the conceptual foundations of the technique and then by exploring a hypothetical two-factor model. Once these conceptual issues are established, we develop the concepts of factor analysis further by applying them to the study of the Resource Mothers Program.

A. Conceptual Foundation: Underlying Factors

The primary difference between factor analysis and principal components analysis concerns the meaning of the dimensions: whereas the goal of principal components analysis is parsimony, factor analysis is concerned in addition with what is referred to as underlying structure (Rummel, 1970). ''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 the observed variables'' (Kim and Mueller, 1978a, p. 12). This emphasis on underlying factors follows from a basic philosophical belief that there are real qualities in the world, such as self-esteem (Shevlin et al., 1995), group cohesion (Cota et al., 1995), aggression

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

(Harris, 1995), personality (Digman and Takemoto-Chock, 1981), and life satisfaction (Shevlin and Bunting, 1994). The need for factor analysis comes from the belief that these qualities are not directly measurable but can be revealed through the covariation of related variables. For more background on the use of factor analysis in revealing underlying constructs, see Thompson and Daniel (1996). In this section we use a simplified model of leadership ability to examine the implications of this focus on underlying factors for the conduct and interpretation of factor analysis.

1. Basic Logic: Factor Identification

Beginning with the observation that some people are better than others at leading organizations, we can accept for illustration the otherwise controversial notion that there is something that we can call ''leadership ability.'' Given our common sense notion of leadership, we might hypothesize that the greater one's leadership ability, the more effective that one might be in improving the quality of services provided by one's subordinates. We recognize, however, that the quality of workgroup performance is the result of many factors, only a subset of which involve the leadership abilities of those in positions of responsibility. As such, we might use a variety of other measures to assess an individual's leadership ability.

Figure 4 illustrates the posited leadership ability manifesting itself in three measurement modalities: (1) ratings by subordinates; (2) performance of natural workgroups; and (3) performance in an assessment center that focuses on leadership tasks. The logic of factor analysis is this: If these measures truly reflect an underlying leadership ability, then individuals with high leadership ability should tend to have high scores on each of these measures. Similarly, those with low leadership ability should tend to have low scores on all of the measures. In other words, if the scores on these measures are caused in part by an underlying common factor, the scores will covary, meaning that they will vary together across individuals. This arrangement is depicted in Figure 4; each of these measures of leadership is influenced to some extent by ''unique factors'' that are essentially unrelated to leadership (e.g., leadership ratings by subordi-

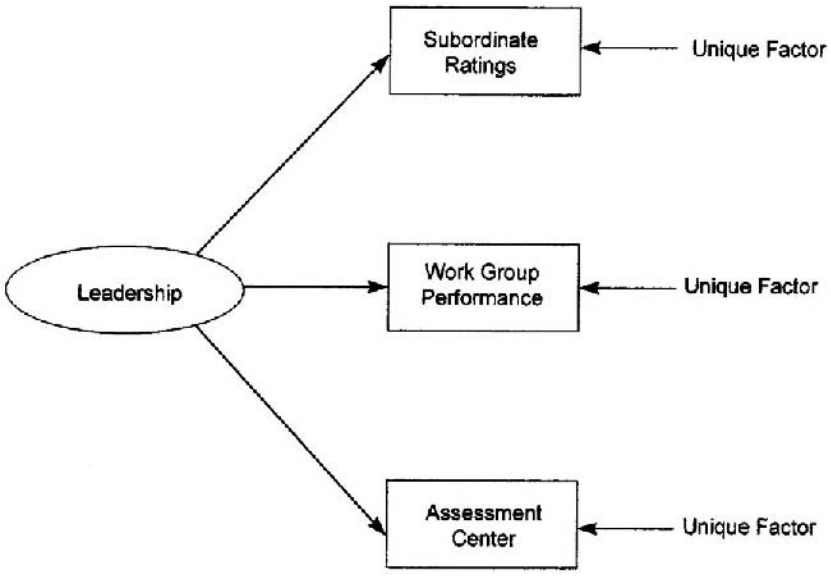


FIGURE 4 Measures of underlying leadership factor.

nates may be affected also by work conditions and even by salaries) and unrelated to each other (there are no arrows between the unique factors), but each also is a function of the common leadership factor.

Because factor analysis is intended to reveal underlying factors, the technique can be contrasted with principal components analysis in terms of the variation of interest. Whereas principal components analysis lumps all variation in the variables into one category, factor analysis ''assumes that data on a variable consists of common and unique parts.... The object of common factor analysis is to define the dimensions of this common vector space'' (Rummel, 1970, p. 104). This means that factor analysis begins by determining the extent to which the variables being examined covary with each other. Only the covariance of the variables is used to determine the underlying dimensions; all other variation among the variables is partitioned out as unique variance. While not illustrated in Figure 4, the unique variance can itself be partitioned into two components, the specific factors (unique underlying factors) and random error.

The concern for factor analysis, therefore, is to find successive factors that account for the covariance of the variables being considered. Parallel to principal components analysis of variance, factor analysis begins by identifying the factor that accounts the greatest amount of covariance and continues by finding additional factors that account for the greatest remaining covariance while subject to the constraint of being orthogonal to the factors already identified. We recall from above that ''orthogonal'' means that the dimensions are unrelated such that the correlation between any two is zero. If one generates as many factors as there are variables, all of the covariance of the variables will be accounted for by the factors. Typically, however, as with principal components analysis, the number of factors solved for and retained is small relative to the number of variables.

2. *Quantitative Model*

We stated above that the goal of factor analysis is, in addition to achieving parsimony, to reveal the underlying factors that produced the patterns observed among the variables. Differentiating between common and unique factors, factor analysis, therefore, seeks a solution to Formula 2, wherein the observed variable, X_j, is a function of both common factors, $CF_{(i)}$, and unique influences, e_j. Formula 3 provides the same information using the symbols of matrix algebra.

$$X_j = \nu_{j(1)}\, CF_{(1)} + \nu_{j(2)}\, CF_{(2)} + \ldots + \nu_{j(i)}\, CF_{(i)} + e_j \tag{2}$$

$$X = \Lambda f = e \tag{3}$$

Notice the difference between Formula 2 and Formula 1; this difference captures much of the contrast between the two techniques. Whereas Formula 1 presented the principal components as functions of the measured variables, Formula 2 highlights the manner in which the observed variables are conceived as functions of underlying factors. Thus, principal components analysis creates a scale based on observed variables; factor analysis estimates factors responsible for observed variables.

The weights in Formula 2 ($\nu_{j(i)}$) and Formula 3 (Λ) are referred to as factor loadings. Those familiar with multiple regression analysis will recognize the form of these two formulas, and, indeed, it is the case that the factor loadings in Formulas 2 and 3 correspond to regression coefficients. ''If all the variables (both hypothetical and observed) are standardized to have unit variance, the linear weights are known as *standardized regression coefficients* (in regression analysis), path coefficients (in causal analysis), or factor loadings (in factor analysis)'' (Kim and Mueller, 1978a, p. 21; emphasis in original).

3. Graphic Representation

As with regression coefficients in traditional multiple regression, factor loadings represent the influence of the common factors on the variables. When the factors are orthogonal (as they are in at least the preliminary stages of factor analysis and principal components analysis), they are independent and the factor loadings are equivalent also to the correlations between the variables and the hypothesized common factors. As correlation coefficients, the factor loadings can be squared to equal r^2 (single variable version of R^2) and thus represent the variance of the observed variable that is accounted for by the hypothesized factor. We can illustrate these relationships by assigning values to the paths depicted in Figure 4.

Using the loadings provided in Figure 5, we can conclude that the ratings by subordinates, the top measure in the figure, has 25% (0.50 squared being equal to 0.25) of its variance explained by the common leadership factor and the remaining 75% (0.87 squared equaling 0.75) explained by unique factors. In contrast, the variable for work group performance, the middle variable in Figure 5, is shown to have 64% (0.80 squared) of its variance accounted for by the leadership factor with the remaining 36% (0.60 squared) of the variance left to be explained by unique factors. Performance in an assessment center is presented as intermediate between the other two measures in terms of explanation by the common factor, with an equal percent of its variance accounted for by the leadership factor as by unique factors (50% for each, as 0.71 times 0.71 is approximately 50%).

4. Extension to Multiple Factors

Extending this analysis to two underlying factors, we can now develop the concepts of factor analysis in the context of a hypothetical two-factor leadership model. In this model we go beyond the view that leadership is a unidimensional capacity and instead elaborate our factor analysis model by using the results of some of the early research on leadership (see Stogdill, 1974) that identified two distinct leadership abilities: (1) task orientation (referred to by Stogdill as ''initiat-

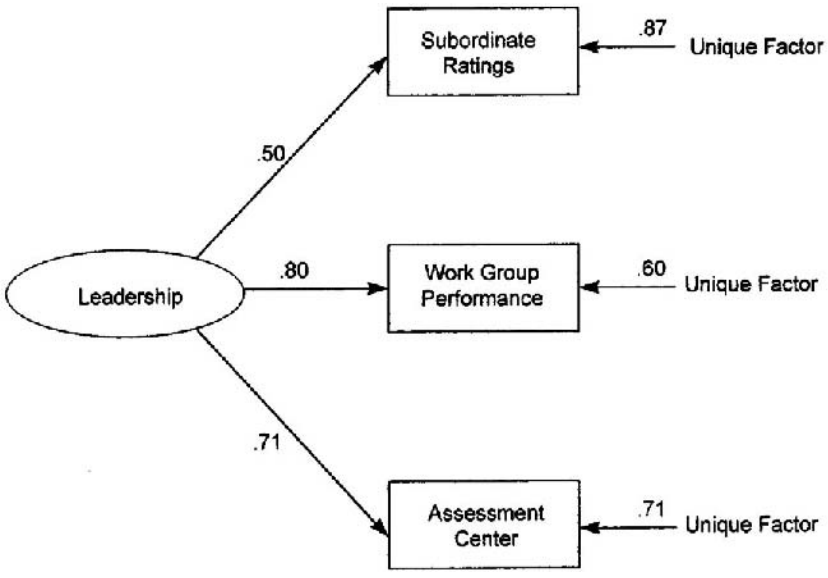


FIGURE 5 Correlations between leadership factor and measures.

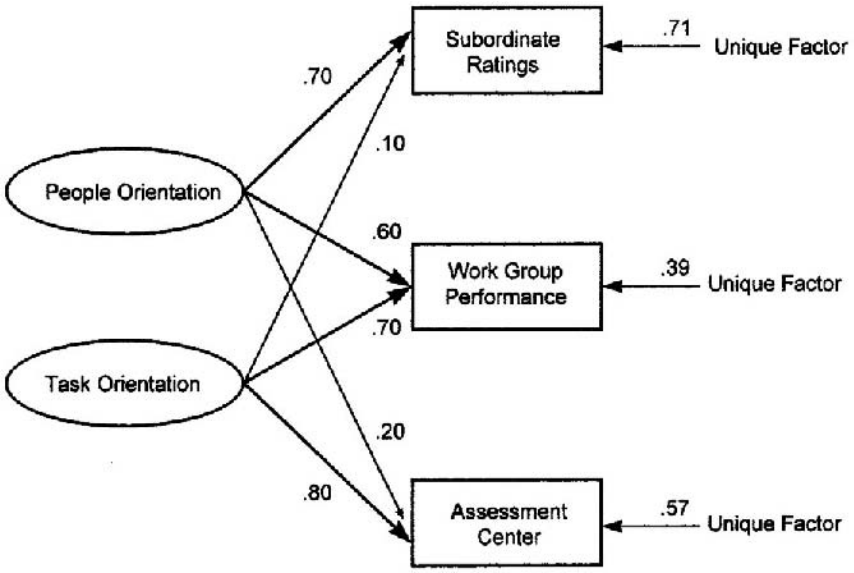


FIGURE 6 Two factor model of leadership.

ing structure'') and (2) people orientation (or ''consideration''). Figure 6 presents this elaborated model, with the three leadership measures conceived now as functions of two common factors. The numbers on the arrows from the two common and the three unique factors are factor loadings and so, as before, when squared they represent the degree to which each of the measures is determined by the underlying factors. We see that the task orientation factor has its greatest influence on the results of the assessment center exercise ($r = 0.80$) and least on the ratings by subordinates ($r = 0.10$). The people orientation factor, on the other hand, has its greatest influence ($r = 0.70$) on the ratings by subordinates and least influence ($r = 0.20$) on assessment center performance (these relationships are for illustration only; research on these two leadership factors has provided little evidence that they influence outcomes in any consistent manner).

The factor loadings displayed in Figure 6 can be presented as the factor pattern matrix presented in Table 6. In addition to summarizing the relationships between factors and variables that were displayed in Figure 6, Table 6 has a column labeled ''communality'' and a row that refers to ''eigenvalues.'' Communality is defined as the proportion of the variance of a variable that is accounted for by the identified common factors (Gorsuch, 1983, p. 29). An extension of the single-factor calculations described above, the communality measure is calculated by squaring and summing the factor loadings for a particular variable, with the result ranging from 0.0,

TABLE 6 Factor Pattern for Two-Factor Solution

	Task orientation	People orientation	Communality
Assessment center	0.80	0.20	0.68
Workgroup performance	0.70	0.60	0.85
Subordinate ratings	0.10	0.70	0.50
Eigenvalues	1.14	0.89	

meaning that the factors explain none of the variance of a variable, to 1.0, which indicates that all of the variance of a variable is in common with the other variables and is accounted for completely by the derived factors (Dillon and Goldstein, 1984, p. 67). In this example with two orthogonal factors, we see that 68% of the variance in assessment center performance, 85% of the variance of the performance of natural workgroups, and 50% of the variance in ratings by subordinates are accounted for by two leadership factors.

Whereas communality refers to the variance of a particular variable explained by all common factors, eigenvalues, introduced above under principal components analysis, represent the complementary concept of the total standardized variance of all variables that is accounted for by a particular factor. As with communalities, eigenvalues can be calculated by squaring the factor loadings, but for this purpose the squared loadings are summed for each factor. The standardized variance of a variable is 1.0, and so, in this example, with three standardized variables, the total variance to be explained is 3.0. Table 6 indicates that the first factor, task orientation, accounts for 1.14 of the 3.0 total, or 38% of variance of the three variables; the second factor accounts for .89 of the 3.0 total, or 27.7% of the variance.

In summary, we have presented the logic of factor analysis and have discussed the quantitative model, stressing the similarity between factor loadings and the regression coefficients and correlation coefficients that many in public administration are familiar with through their use of multiple regression analysis and correlation analysis. There are, however, additional concepts that need to be addressed in applying factor analysis to the example of the Resource Mothers Program. We will see that one important concept to be discussed is factor rotation. The issue of rotation arises because, unlike the independent variables in regression analysis, the predictor variables in Formula 2 are hypothetical common factors which must be estimated as well as the regression coefficients. This need to estimate the common factors will mean that the factors derived by the analysis will not be uniquely suitable, that there will be an inevitable indeterminancy as to the nature of the underlying structure. One consequence of this need for estimation is that the factor scores, values for the factors that are parallel to the composite variables produced by principal components analysis, are likely to have non-zero correlations despite the true factors being orthogonal.

B. Application to Resource Mothers Program

The above sections presented the basic concepts of factor analysis, many of which require users to make a variety of choices when applying the technique to their data. This section will address some of these choices and their implications in the context of the Resource Mothers Program. The analyses reported below were conducted using the FACTOR procedure in the SAS statistical package; the actual commands used are displayed in the left side of Table 7. Following the

TABLE 7 SAS Computer Commands for Factor Analysis Commands

Computer commands	Functions of commands
PROC FACTOR M = PRIN PRIORS = SMC R = PROMAX N = 2;	M = principal components method; setting PRIORS to SMC means that the squared multiple correlations will be used to estimate the communalities of the variables; F = PROMAX specifies an oblique rotation; N = 2 specifies maximum number of factors.
VAR ⟨variables⟩;	Specifies the variables to be analyzed.

PROC FACTOR command we see listed M = PRIN. The M refers to ''method'' and indicates that we will be using what will be described as the principal components method for deriving factors. An alternative is to specify M = ML, which directs the computer to use the maximum likelihood solution for deriving the factors. After the Method option is the command that controls how we are estimating the communalities that will be used to produce the reduced correlation matrix used in factor analysis. By specifying PRIORS = SMC, we are guiding the analysis by estimating the communalities based on the squared multiple correlation of the variables, defined and calculated as the proportion of variance of each variable that can be accounted for by the variance of all other variables. On the following line (moving to the next line is of no consequence in SAS), the command R = PROMAX refers to a variation of what are described below as rotations that are performed on the original results in order to support more meaningful interpretations. And, finally, N = 2 specifies the maximum number of factors to be retained.

1. Determining the Number of Factors

We will see that there are a variety of available quantitative solutions that can be used to derive underlying factors. Before considering these options, however, we will first address the question of the number of factors to be retained in our analysis. Many of the relevant issues were introduced above when considering this question for principal components analysis and so will be mentioned only briefly here. For example, as with principal components analysis, we can use scree tests to decide on the number of factors. In an attempt to improve on the scree test as presented above, Zoski and Jurs (1996) discuss the use of standard errors to provide a more objective scree test. Also noted under principal components analysis, one can decide on the appropriate number of factors through use of criteria such as the proportion of variance of the variables that are accounted by the total set of factors and the proportion of variance accounted for by each additional factor.

Some aspects of factor analysis, however, require a different understanding from that applied for principal components analysis when choosing the number of dimensions to retain. First, because factor analysis attempts to address only the common variance of the variables, there is less variation to be accounted for by the factors and so fewer factors may be required. Second, because factor analysis is explicitly concerned with identifying underlying structure, it is even more important than with principal components analysis that the number of factors to be retained be decided in the context of the relationships between the derived factors and the constructs that are supported by previous experience. The factor pattern that results should have high loadings on factors from variables that are expected to covary. Third, because factor analysis attempts to partition covariance into distinct groups of reliable variation and error variation, it is important that the number of factors retained be such that we can reduce the error variation included in the interpreted factors. The point of this last concern is to minimize the bias in the factors that result from analysis.

Because of these reasons, we have to be particularly concerned with the relative dangers of overinclusion (retaining too many factors) and underinclusion (too few factors). Recall from above when this issue was addressed for principal components analysis that Cattell's use of the scree test had evolved to signify fewer factors than when he began using the test. Nonetheless, ''Thurstone (1947) and Cattell (1978) held that overextraction [too many factors] produces less distortion in factor analytic solutions than underextraction and is therefore to be preferred'' (Wood et al., 1996, p. 154).

Based on their study of principal axis factor analysis and varimax rotation, Wood, Tataryn, and Gorsuch (1996) conclude, ''When underextraction occurs [too few factors retained], the estimated factors are likely to contain considerable error. . . . When overextraction occurs, the

estimated loadings for true (i.e., ''core'') factors generally contain less error than in the case of underextraction'' (p. 359). The implications of their study are that, first, it is important to retain the correct number of factors but that, second, retaining too many factors produces less bias than too few. If one wants to reduce the bias in the core factors (the first several or so that seem strongly supported by high loadings for related variables), one might choose to extract additional factors, recognizing that the additional factors may be false factors that are not to be interpreted. Thus, there may be circumstances in which the users of factor analysis may direct the computer to retain more factors than they intend to use in their subsequent analyses and interpretation.

2. *Solutions*

There are numerous ways to estimate factors in an attempt to fulfill the basic goal of identifying underlying structure. For more information on the options for deriving the initial estimates for factor analysis, see Buley (1995). In brief, early applications used heuristic strategies to develop estimates of the factors (Gorsuch, 1983). More recently the availability of computers has resulted in the use of iterative solutions in which initial estimates allow calculations that result in more refined estimates and, hence, further calculations. We will begin by considering the basic principal factor method and then the maximum likelihood method, one that makes particular use of iterative estimation.

a. Principal Factor Method Though different in concept, this method is identical in calculations to the factor extraction method described above for principal components analysis. One begins by calculating a correlation matrix for the variables to be factor analyzed. This matrix is modified, however, by replacing the diagonal entries (1.0 for the correlation of each variable with itself) with the estimated communalities of the variables. These communalities are estimated in our analysis by squaring the multiple correlation of each variable with all other variables (indicated by the PRIORS = SMC statement in Table 7). This replacement effects the change from an analysis of variance—appropriate for principal components analysis—to an analysis of covariance. This new matrix, with estimated communalities along the diagonal, is referred to as a reduced correlation matrix. Factor analysis then operates on this reduced matrix to reveal its dimensions.

Had we not specified PRIORS = SMC, the SAS program would have, as its default option, used PRIORS = ONE. This default option would have left the 1's along the diagonal of the matrix being analyzed. By putting 1's along the diagonal of the matrix, we are returned to the original, unreduced, correlation matrix and would, therefore, be performing principal components analysis. Thus, we see that we could also perform principal components analysis using the FACTOR procedure in SAS, with METHOD set to be PRINCIPAL and PRIORS set to ONE. This is important to emphasize in that some users might believe that they are performing factor analysis (i.e., common factor analysis) when using the PROC FACTOR program in SAS but, in not changing the default setting of 1.0 for prior communality estimates, are really conducting principal components analysis.

Having raised this concern about possible misuse of principal components analysis as factor analysis, let us also recognize that often it may not matter. That is, if the reduced correlation matrix (with communality estimates along the diagonal) is similar to the original correlation matrix, then principal components analysis and factor analysis will yield similar results. One of the ways in which the two matrices will be similar is if the communality estimates of the reduced matrix are all close to 1.0 (in practical terms, if the communalities are all above 0.70). This is why our Resource Mothers Program example deliberately includes two variables with somewhat low communalities; only by including these lower communalities do we have the

TABLE 8 Factor Pattern for Principal Components Method

	First factor	Second factor	Communality
Months of care	0.34	0.45	31%
Prenatal visits	0.28	0.56	39%
Weight gain	0.13	0.16	4%
Age of mother	0.80	−0.18	66%
Mother's education	0.71	−0.22	55%
Marital status	0.48	−0.06	23%
Eigenvalues	1.57	0.62	

opportunity to notice some differences between principal components analysis and the principal factor method of factor analysis. The other way in which the reduced and original correlation matrices become similar is if the number of variables is large. If there are, say, 20 variables being analyzed, the 20 entries along the diagonal are only a small part of the correlation matrix (190 other cells in the resulting half matrix), and so the reduced and original matrices; alike in every nondiagonal entry, become essentially the same.

Table 8 presents the results of the principal factor method when applied to the data from the Resource Mothers Project. The numbers in the table are factor loadings and so represent the correlations between the variables and the derived factors. As with principal components analysis, these numbers can be used to interpret the meaning of the derived factors. The first three variables have their highest loadings on the second factor, the last three variables load highest on the first factor. From this pattern of loadings we see that each factor is defined primarily by two variables. The largest loadings on Factor 1 are for Age of Mother (0.80) and Mother's Education (0.71). Factor 2 has the highest correlations (absolute value) with the number of Prenatal Visits (0.56) and with the Months of Care (0.49).

We also note in Table 8 that the communalities are fairly low. The highest communalities are only 66% of the variance of Age of Mother and 55% of Mother's Education being accounted for by the two factors. The lowest communalities are 23% of Marital Status and only 4% of Weight Gain explained in this way. Further, we see from the eigenvalues that while the first factor explains a sufficient amount of variance (1.57 out of a possible 6.0), the second factor explains comparatively little (0.62).

In addition to placing the results in a table, it is possible to depict them graphically, as is done in Figure 7. In this figure, the correlations between each of the variables and the two

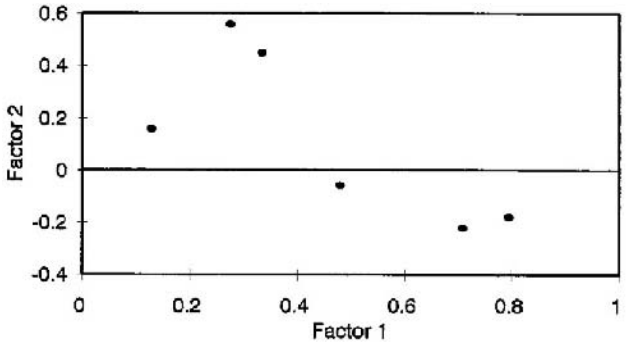


FIGURE 7 Factor analysis with principal components method.

common factors (presented in Table 8) are plotted, with positive correlations being placed above and in front of the axes defined by the two factors, negative correlations being below and behind the axes.

Comparing the factor loadings to the results of the principal components analysis, the most immediate impression is of the similarity of the results—in both approaches the rank ordering of variables in terms of their correlations with the derived dimensions is the same. Looking closer, we can see some meaningful differences in the two sets of results. For example, Factor 1 has higher correlations with its primary variables (age of mother, mother's education, and marital status) than does the first principal component. In contrast, Factor 2 is less distinguished in its associations than the second principal component, with lower positive correlations with its primary variables (number of prenatal visits and months of care) and less negative correlations (closer to zero) with the other three variables. These differences aside, the relative sizes of the loading on the factors are remarkably consistent; even with few variables and some relatively low communalities, principal components analysis and factor analysis would in this example lead to the same substantive conclusions.

b. Maximum Likelihood Solution A more recent approach to estimating factors uses maximum likelihood analysis to solve for the factors. Although this approach is more demanding of computer resources, it is also more consistent with the intended logic of factor analysis. ''What values for the population parameters make the sample observations have the greatest joint likelihood? When we answer this question, we will take such values to be *maximum likelihood estimators* of the population parameters'' (Mulaik, 1972, cited by Dillon and Goldstein, 1984, pp. 80–81).

As indicated in the above quote, the maximum likelihood approach offers an appealing strategy for estimating the factors that produced the observed data. This approach, however, does make additional data requirements over the earlier principal components solution. For a more adequate explanation of these added assumptions, the reader is referred to Dillon and Goldstein (1984) or Morrison (1990). We mention here, however, two areas of potential concern if using the maximum likelihood approach. First, the maximum likelihood solution requires that the variables being analyzed are not linearly dependent on each other (i.e., the matrix is nonsingular). Second, the distribution of the variables is presumed to be multivariate normal, an assumption that is violated in the typical case of responses on survey scales. These assumptions can restrict the use of the maximum likelihood solution, but Fuller and Hemmerle (1966) found that most violations of the normal distribution assumption do not distort the results beyond usefulness, at least not when working with large sample sizes ($n = 200$ in their study).

Applying the maximum likelihood solution to the data from the Resource Mothers Project yields first an assessment of the number of factors needed for an adequate solution and then an interpretation of the derived factors. Table 9 presents the series of hypothesis tests (these tests presume multinormal distributions and are likely more sensitive to violations of this assumption than are the factor loading results mentioned in the previous paragraph; see Gorsuch, 1983, p. 148), based on the chi-squared distribution, that is provided by the maximum likelihood

TABLE 9 Chi-Squared Tests for Number of Factors using Maximum Likelihood Method

Null Hypothesis	Chi-squared	Probability	Decision
No common factors	375.33	0.0001	Reject null
One factor sufficient	60.68	0.0001	Reject null
Two factors sufficient	7.05	0.133	Fail to reject

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

solution. Interpreting Table 9, note that the first null hypothesis, that there are no common factors underlying the three measures, would be rejected based on the relatively large chi-square value and associated low probability ($p < 0.0001$). Similarly, the second null hypothesis contends that no more than one common factor is needed, and, due to the large chi-square score, is also rejected ($p < 0.0001$). In contrast, the third null hypothesis, that no more than two common factors are required, cannot be rejected at traditional levels of significance ($p = 0.133$), suggesting, therefore, that two factors may be adequate.

This statistical evidence of the appropriateness of a two-factor solution is consistent with our expectations, but we reiterate that this use of the chi-squared test is recognized as often suggesting more factors than researchers are willing to accept. While, as argued above, overinclusion is to be preferred to underinclusion, this tendency towards overinclusion of factors leads authors such as Kim and Mueller (1978a) to emphasize the importance of the substantive significance of the factors over and above their statistical significance.

As to the interpretation of the two factors, we see in Table 10 that the rank ordering of the largest correlations again remains consistent with the previous two analyses of these six variables. To the extent that there are meaningful differences in the results for the maximum likelihood solution and the principal components solution, it is that the maximum likelihood solution is inclined to orient the results to emphasize individual variables. This consistency across methods is reassuring for both the reliability and validity of the approach, particularly as the nature of this example works against this consistency. First, as noted earlier, our example has two variables with relatively low communalities. The results of the maximum likelihood method converge with the results of principal factor method as all of the communalities approach 1.0 (Gorsuch, 1983, p. 121). Second, we have used only six variables. ''*As the number of variables increase, communality estimates and the method by which exploratory factors are extracted both become less important*'' (Gorsuch, 1983, p. 123; emphasis in original).

3. Factor Rotation

We have described the goal of factor analysis in terms of identifying underlying structure. The methodology of factor analysis presumes that this identification is most effective when the results are relatively easy to interpret. However, several features of the original factor solution work against meaningful interpretation. In particular, because the various solutions begin with the best-fitting factor and continue by identifying progressively less adequate factors, the first factor tends to be a general one with relatively strong relationships with all of the variables. The remaining factors tend to have complicated and potentially confusing relationships with many of the variables. We attempt to counter this potential confusion and achieve the desired interpretable factors by what is called factor rotation.

TABLE 10 Factor Pattern for Maximum Likelihood Method

	First factor	Second factor	Communality
Months of care	0.17	0.45	23%
Prenatal visits	0.00	1.00	100%
Weight gain	0.09	0.22	6%
Age of mother	0.99	0.10	100%
Mother's education	0.69	0.05	48%
Marital status	0.43	0.14	21%
Eigenvalues	1.69	1.28	

To understand the logic of rotation, remember that the particular dimensions that result from factor analysis are not uniquely capable of representing the interdependence of the variables analyzed. Rather, the derived factors provide a grid to aid interpretation in much the same way as geographic lines of longitude and latitude provide a grid for representing geographical relationships. Whereas we have conventions for the fixed directions of north, south, east, and west, we might find it more meaningful to rotate the grid placed on a specific geographic area so that the horizontal and vertical dimensions highlighted significant variation (e.g., elevation above sea level or political ideology of residents) that we wished to interpret. Similarly, we rotate the grid for the factors so that the results might be more meaningful and more easily interpretable.

One way in which results are easy to interpret is when they suggest factors that are consistent with our prior expectations. Results are also interpretable if groups of variables cluster together around distinct factors, an outcome referred to as ''simple structure.'' With this simple structure as a goal, rotation is judged successful when:

1. Each variable is identified with one or a small proportion of the factors.
2. The number of variables correlated with (loaded on) a factor is minimized.
3. The variance accounted for by the major unrotated factors is spread across all the rotated factors.
4. Each rotated factor is now more or less identified with a distinct cluster of interrelated variances (Rummel, 1970, pp. 380–381).

The basic idea of simple structure is described in this quote by Rummel as requiring each factor to have its own small set of variables with significant loadings on it. ''Significant'' in this context refers not necessarily to statistical significance but to loadings that are moderately large. Recognizing that factor analysis, with its goal of accounting only for common variance, is expected to account for less total variance than principal components analysis, as a rule of thumb one can think of loadings of 0.30 or larger as being significant in factor analysis (Dillon and Goldstein, 1984, p. 69).

Researchers have accepted that the primary goal of rotations is to achieve simple structure. ''Unfortunately, the concept of simplicity itself is not so straightforward as to allow for a formal and undisputed criterion'' (Kim and Mueller, 1978b, p. 30). Given this lack of a straightforward criterion, there are a variety of options. The primary distinction that differentiates types of rotation is between orthogonal and oblique rotations. Orthogonal rotations maintain the constraint that the factors be orthogonal or independent; oblique rotations relax this constraint and allow the factors to be correlated.

Looking first at the orthogonal rotations, each approach is based on a different notion of what constitutes simple structure. Quartimax is the name given to rotation that emphasizes having each variable load on a minimum number of factors; the goal is to avoid having variables load on more than one factor. Varimax rotation emphasizes the other aspect of simplicity, that each factor should have only a few variables loading on it; the goal is to avoid general factors that are associated with many of the variables. Equimax rotation, as its name suggests, takes an intermediate, or equidistant, stance between these two criteria.

Table 11 displays the results of a varimax rotation when applied to the principal components solution for factor analysis presented in Table 8 (varimax is the initial rotation that SAS uses to prepare the factors for the oblique promax rotation and so is reported when R = promax is specified). We apply the rotation to the principal components solution rather than to the maximum likelihood solution because the results provide clearer illustration of desired effects of rotation. Compared with the unrotated results in Table 8, we note that the varimax rotation balanced somewhat the variance explained by the two factors, increasing the variance explained

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 11 Factor Pattern for Orthogonal Rotation (Varimax)

	First factor	Second factor	Communality
Months of care	0.13	0.54	31%
Prenatal visits	0.04	0.62	39%
Weight gain	0.06	0.20	4%
Age of mother	0.80	0.15	66%
Mother's education	0.74	0.08	55%
Marital status	0.46	0.14	23%
Eigenvalues	1.42	0.77	

by the second factor and decreasing the variance explained by the first. This balancing fulfills the primary objective of the varimax rotation, avoiding a general factor with many variables. Also, the structure is simpler in having variables load on only one factor. For example, whereas the variable Months of Care previously loaded on both the first and second factors, after the varimax rotation it loads on only the second factor. Note, however, that, comparing the results with those in Table 8, the communalities do not change; the amount of variance accounted for by the factors does not change because of rotation (the reader is invited to square and sum the loadings to confirm this). Figure 8 presents a graphic depiction of this simpler structure wherein the variables are closer to the factor axes than they were in Figure 7.

Oblique rotation is similar to orthogonal rotation in that the goal is simple structure with easily interpretable factors; the difference is that with oblique rotation the factors are no longer required to be statistically independent. Graphically this means that the factors are no longer required to be at right angles with each other. As an example of one of the more recent approaches, promax rotation begins with an orthogonal rotation and then modifies the factors using what is called a target matrix as a guide.

> The rationale behind the promax rotation is that the orthogonal solutions are usually close to the oblique solution, and by reducing the smaller loadings to near-zero loadings, one can obtain a reasonably good simple structure target matrix. Then by finding the best fitting oblique factors for this target matrix, one obtains the desired oblique solution (Kim and Mueller, 1978b, p. 40).

Table 12 reports the results of the promax rotation, and, as intended, the smaller loadings are generally closer to zero than with the orthogonal varimax rotation. These smaller loadings

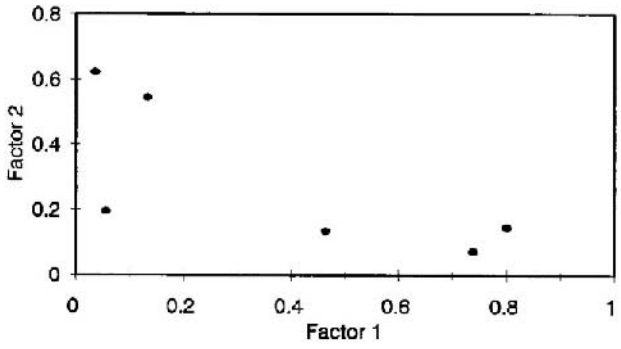


FIGURE 8 Varimax orthogonal rotation.

TABLE 12 Factor Pattern for Oblique Rotation (Promax) Standardized Regression Coefficients; Inter-Factor Correlation = 0.27

	First factor	Second factor
Months of care	0.07	0.54
Prenatal visits	−0.04	0.63
Weight gain	0.03	0.19
Age of mother	0.81	0.03
Mother's education	0.75	−0.04
Marital status	0.46	0.07
Variance accounted for (controlling for other factors)	1.44	0.73

represent what has been described as simple structure and provide some rational foundation for the oblique rotation. Changes in the first factor are minor (with, for example, Months of Care decreasing from 0.13 to 0.07), but the decreased loadings for the last three variables on the second factor result in the factor being more clearly associated with only the first three variables. The communalities for the variables remain the same as for the unrotated and varimax solutions but are not reported in this table because, with the factors correlated, they can no longer be calculated by squaring the loadings for each variable.

One virtue of oblique rotations is that they reveal the degree to which the identified factors are correlated. If the results indicate that the rotated factors are essentially not correlated, despite the relaxing of that requirement, then you can be more confident that the real world factors are independent and can use the orthogonal solution. If, on the other hand, the resulting factors are strongly correlated, then you can compare this result with your understanding of the real world relationships among the constructs that you believe that you identified. Often we expect factors to be related and so want our quantitative methods to allow for this. In our example, Factor 1 has a correlation of 0.27 with Factor 2. This is a moderately high correlation that suggests that the construct addressed by Factor 1 is meaningfully related to the construct for Factor 2. This correlation can be displayed graphically as in Figure 9. We see that the oblique factors are no longer at right angles (instead, they are at the angle that corresponds to a correlation of 0.27) and that the new factors are more closely associated with distinct clusters of variables.

The primary disadvantage of oblique rotations is that they complicate matters. In particu-

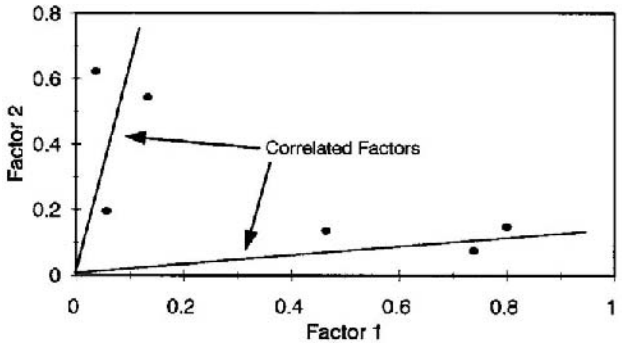


FIGURE 9 Promax oblique rotation.

lar, the relationships between variables and factors are open to contrasting interpretations. For the unrotated and the orthogonal rotations we presented the factor loadings both as correlations between the variables and the factors and as standardized regression coefficients relating the factors to the variables. As long as the dimensions are orthogonal, these two interpretations involve the same numbers. With oblique factors, however, the regression and correlation coefficients represent different relationships, and the tables that report the regression and correlation coefficients are given different names. The regression coefficients reported in Table 12 retain the label ''factor pattern'' that we used for the orthogonal loadings. The table of correlation coefficients is referred to instead as a table of ''factor structure.'' Thus, to provide a complete account of the oblique promax rotation we need to present the correlations between the variables and factors in Table 13. The main thing to note about Table 13 is that the significant correlations tend to be about the same size as the corresponding regression coefficients in Table 12, while the near-zero loadings in Table 12 are noticeably larger in Table 13. For example, the small loading for Months of Care on the first factor, 0.07, increases to a correlation of 0.21 on the first factor in Table 13. The correlation of 0.21 is the overall relationship between the variable and the first factor; the loading of 0.07 is the relationship when the effect of the correlated factor is controlled (partial relationship).

Gorsuch (1983, pp. 206–208) describes the complementary contributions of these two matrices (and also the matrix of reference vectors) but also explains why the factor structure is typically more central to interpretation of factors. A major advantage of interpreting the factor structure is that the loadings are conceptually independent of the other factors derived in the analysis, a property particularly important when comparing the results of different studies. ''Regardless of what other factors occur in the next study, the variables should correlate at the same level with a particular factor'' (Gorsuch, 1983, p. 207). Thus, the correlation between Months of Care and the first factor would be expected to remain at approximately 0.21 in future studies whereas the loading of 0.07 is dependent on having variables that produce a factor similar to the second factor in Table 12.

In summary, because of the emphasis on underlying structure in factor analysis, it is more important with this technique than with principal components analysis that the resulting factors suggest meaningful interpretations. Thus, while rotations can be valuable for principal components analysis, they are more central for factor analysis. Choices among the various rotations can be daunting as the approach chosen might be expected to influence the eventual interpretation of one's results. The available texts provide further guidance in this matter (see Gorsuch, 1983,

TABLE 13 Factor Structure for Oblique Rotation (Promax) Correlation Coefficients; Inter-Factor Correlation = 0.27

	Age of mother	Health behaviors
Months of care	0.21	0.56
Prenatal visits	0.13	0.62
Weight gain	0.09	0.20
Age of mother	0.81	0.24
Mother's education	0.74	0.16
Marital status	0.48	0.19
Variance accounted for (ignoring other factors)	1.51	0.86

pp. 175–238; Kim and Mueller, 1978b, pp. 29–41; or Rummel, 1970, pp. 368–422), but we can offer some crude conclusions.

First, the use of rotations to achieve simple structure presumes that simple structure is desirable. Keep in mind that ''simple structure is a mathematically arbitrary criterion concept that will hopefully lead to greater invariance, greater psychological meaningfulness, and greater parsimony than would leaving the factors unrotated'' (Gorsuch, 1983, p. 231). You may confront situations in which simple structure is not desirable. For example, if you expect to find a general factor that underlies all of your measured variables (as some expect to find a general factor of intelligence underlying a variety of measures or a general leadership factor underlying the range of specific measures), then a varimax rotation would be inappropriate as it would attempt to separate the variables away from this general factor.

As a second point, if you have reason to believe that simple structure is an accurate reflection of the relationships in your data, then many of the available rotations may be adequate: ''If the simple structure is clear, any of the more popular procedures can be expected to lead to the same interpretations. Rotating to the varimax and promax or Harris-Kaiser criteria is currently a recommended procedure'' (Gorsuch, 1983, p. 205). One advantage of this two-step varimax-promax procedure as performed by SAS is that, as explained above, it allows one to decide whether the orthogonal varimax rotation is adequate or whether the oblique promax rotation is necessary. If the factors correlations are negligible in the oblique solution, you have an important argument for sticking with the orthogonal approach. On the other hand, significant inter-factor correlations argue for an oblique approach.

C. Summary Assessment of Factor Analysis

We have discussed factor analysis in terms of its goal of revealing underlying structure. This goal suggests the position of factor analysis on the three distinctions in organizing phenomena as described above. First, factor analysis can be used with equal facility to organize attributes (in particular, R-analysis in which one creates dimensions that account for variables, as we have done here), objects (the Q-factor analysis approach described in Chapter 24 of this book), and occasions (O-analysis when one object has many variables measured at different time periods, or T-analysis, where many objects are measured on one variable at different times). Second, these underlying structures are presumed to be represented better using continuous dimensions rather than discrete categories. And, third, factor analysis is based on a realist viewpoint in which underlying structures are presumed to exist and to be important to understand.

1. Strengths

As with principal components analysis, the strengths of factor analysis need to be understood in the way that the approach pursues its goals. In addition to contributing to parsimony in organizing information, factor analysis does operate on common variance and so does provide a technique for exploring the possibility that underlying factors are responsible for covariation among observed variables. That is, unlike principal components analysis, factor analysis does focus on covariance and so does presume a measurement model wherein some of the variance of a variable is understood as being due to unique factors and to measurement error. Because the technique does analyze covariance rather than variance, to the extent that there are underlying influences that have common effects on variables, factor analysis should find it.

2. Concerns

The main concerns in using factor analysis follow from what is said above but can be summarized in terms of: (1) indeterminancy; (2) instability across methods; and (3) instability from

small changes in the data. The problem of indeterminancy is that factor analysis, unlike principal components analysis, requires estimating unobservable factors. The indeterminacy arises because there will always be alternative conceivable factors that would produce the same observed covariation among the variables measured. While this presumption of unobservable factors, including their unobservable relationships to the variables measured, is responsible for much of the usefulness of factor analysis, it also creates concern that the results of analysis may be misleading. For example, some consider the problem of correctly estimating the communalities for the variables to be so problematic as to argue for the adoption of principal components analysis rather than factor analysis (e.g., Nunnally, 1978).

Adding to the lack of faith that some have in factor analysis are the differences that we saw above when comparing the results of different methods of estimating factors (principal components method versus the maximum likelihood method). While the rank ordering of the relationships between variables and factors remained the same for our two methods, the sizes of loadings changed dramatically, changes that could lead to differing interpretations. In addition to the instability of results across methods is the instability that results from small changes in the data. If a large data set is divided randomly into two parts, analyses of the two separate parts often yield different factor patterns. Similarly, if different samples of a population are taken or data are collected at two points in time, the resulting estimated factors can differ markedly. All of this argues for caution in interpreting the results of factor analysis.

Because of this desire to avoid overinterpretation, some researchers choose, under certain circumstances, to use factor-based scores rather than factor scores in subsequent analyses (see Kim and Mueller, 1978b, pp. 70–72). This means that rather than use the estimated values for factors scores that are based on derived weights provided by computer packages such as SAS, some prefer to use factor analysis to indicate which variables have similar high loadings on a factor (e.g., Months of Care and Prenatal Visits in the promax factor structure of Table 13) and simply take the average of those variables as an equal-weighting scale that is comparable to a factor score.

IV. CLUSTER ANALYSIS

Cluster analysis refers to a quantitative approach to classification that was developed during the 1930s in social science (e.g., Tryon, 1939) and elaborated in the 1960s in biology (Sokal and Sneath, 1963). By classification, we mean ''the ordering of entities into groups or classes on the basis of their similarity'' (Bailey, 1994, p. 1). As with factor analysis and principal components analysis, the development and use of cluster analysis accelerated considerably after modern computers became available to researchers. We address the major issues in cluster analysis by first presenting the conceptual foundation of the approach and then addressing the practical concerns by applying cluster analysis to the example of the Resource Mothers Program. More information on cluster analysis can be found in general multivariate texts (e.g., Dillon and Goldstein, 1984) and in books that focus on cluster analysis (Aldenderfer and Blashfield, 1984; Bailey, 1994).

A. Conceptual Foundation

Cluster analysis, as an empirical approach to classification, seeks to identify not continuous dimensions along which phenomena vary (the focus of principal components analysis and factor analysis) but categories into which phenomena can be placed. The foundation for cluster analysis is based on prior work on conceptual classification schemes (Bailey, 1994). Examples of particu-

lar relevance to administrators include the classification of organizations into mechanistic and organic types (Burns and Stalker, 1961). Scott (1981) used two dimensions, natural versus rational systems and open versus closed systems, to distinguish four variants of organization theory. Daft and Weick (1984), also proposing a two-dimensional model, relate their dimensions to four types of organizations.

Conceptual typologies such as those cited above generally distinguish pure types that define categories, categories into which actual phenomena, such as existing organizations for Daft and Weick (1984), fit to a greater or lesser degree. Cluster analysis, in contrast, is an empirical technique that begins with actual entities, such as individual employees, and then groups them into categories based on measured similarity. In order to provide a conceptual foundation for cluster analysis, we present below the logic of this similarity grouping, the quantitative model used to assess similarity, and a graphical representation of a cluster model.

1. Underlying Logic: Grouping by Similarity

Although the term ''cluster'' is not easily defined for quantitative analysis, the goal of cluster analysis is to establish a taxonomy that is comprised of meaningful categories for classifying the phenomena of interest to investigators. Meaningful categories often are presumed to be those that involve ''clusters of objects that display small *within-cluster* variation relative to the *between-cluster* variation'' (Dillon and Goldstein, 1984, pp. 157–158). Identifying clusters with desired within-cluster and between-cluster variation requires being able to generate overall measures of similarity, or distance as a measure of dissimilarity, for pairs of entities being analyzed.

The logic of cluster analysis is based on the belief that the resulting numerical taxonomies assist researchers and practitioners in describing the phenomena of interest and the relationships among them. This assistance depends, as in the case of principal components analysis and factor analysis, on the value achieved in reducing the complexity of the real world down to more manageable categories. To the extent that the categories derived refer to real distinctions between entities, cluster analysis offers the promise of highlighting real similarities among subsets of phenomena along with real differences between subsets (Bailey, 1994).

The logic as presented so far has been described in terms of developing clusters of objects such as types of organizations or types of leaders, and this is indeed the primary use of cluster analysis (in contrast to the standard use of factor analysis to explicate relationships among variables). The objects for this type of analysis (which we have referred to as Q-analysis) are sometimes called entities, and they are differentiated by virtue of differing characteristics, as measured by variables. Recall, however, that measures can be constructed using other slices of the data cube that was introduced at the beginning of this chapter (Dillon and Goldstein, 1984). For example, variables can be clustered together by using the values of objects on those variables as characteristics of the variables (R-analysis). Similarly, one might cluster occasions together by measuring many characteristics of an entity on multiple occasions (O-analysis).

2. Quantitative Model

The effort to identify meaningful categories based on some notion of similarity assessment requires a quantitative framework that can operationalize similarity judgments and use this sense of overall similarity assessment to make classification decisions. In what follows, we discuss two ways to make similarity judgments and describe a procedure to combine these judgments into an overall schema of classification.

a. Operationalizing Similarity The first task of cluster analysis is to characterize the similarity of the entities being studied. Most approaches to this task depend on calculating a measure of the similarity of each entity to every other entity being studied. This results in a similarity

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

score for each of the possible pairings of entities within the particular population being studied. The two most common approaches to generating these similarity scores are based on distance measures and correlation coefficients.

Distance measures take many forms, but the main points can be made using Euclidean distance as a measure of similarity. In this approach, the differences between two entities on all of the variables measured are combined in the same way that, recalling Pythagorean's theorem, the length of a hypotenuse of a right triangle can be calculated by combining the lengths of the orthogonal legs. Once the differences between the two entities on each of the variables is squared and then summed, the square root is taken to yield the Euclidean distance between the two. Those entities separated by the least Euclidean distance are defined as being most similar. Another type of distance is referred to as city-block distance, represented by the sum of all of the differences between two entities.

Correlation coefficients provide an alternative measure of similarity. When clustering objects, however, the correlation is calculated not between variables measured across objects, as typically done, but between the objects themselves. This may seem counter-intuitive to some, but, as Aldenderfer and Blashfield (1984) point out, it uses the same data matrix as regular correlations; the matrix is simply inverted so that the rows become columns and columns become rows. This inversion of the matrix highlights an important concern for the use of cluster analysis in clustering objects. Just as one would want to have many more observations than variables in traditional factor analysis (R-analysis), so it is reversed in cluster analysis of objects—the desired ratio reverses and one wants many more variables than observations in clustering objects (Bailey, 1994).

Other similarity measures have been developed, such as association measures for dichotomous data; interested readers are encouraged to consult multivariate texts or the more specific works by Sneath and Sokol (1973) and Clifford and Stephenson (1975). The main reason to consult these other works and to think carefully about your measure of similarity is that the different measures can produce quite different results, a point that will be made again below (Aldenderfer and Blashfield, 1984, pp. 26–28).

b. Defining Clusters Once the chosen similarity measure is calculated for all pairs of the entities being considered, we need a method of using those measures to create categories. Three such methods will be described below, but we can introduce the quantitative methods involved by describing one of the simpler procedures, the single linkage method, as it is applied in one of the more common approaches, hierarchical agglomerative clustering. This method begins by identifying the two entities that are most similar (least distant) based on the variables used. These two entities are grouped together as a cluster. Then the pair of entities with the next highest similarity are grouped together as a cluster. If one of these second most-similar entities is part of the first cluster, then the first cluster incorporates also the other similar entity. Otherwise a second cluster of the two similar entities is formed. This process continues, with new clusters being formed and entities and clusters being incorporated in a hierarchical manner into other clusters, until at the last stage all entities are part of a single, completely inclusive cluster.

This process of forming clusters is fairly straightforward, but there are important decisions that affect the nature of the clusters that result. For example, as with principal components analysis, one has to decide whether or not to standardize the variables before analysis. The argument for standardization is that it allows each variable a somewhat equal opportunity to influence the results; without standardization the variables with the greatest variance will dominate the results. The argument against standardization begins by noting that you may wish variables with the greatest variance to have the greatest impacts on the resulting clusters, the logic being that you want the clustering to reflect the meaningful variation found in those variables (Hartigan, 1975). Related, standardization, while providing some equalization among variables,

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

is not neutral but rather represents a particular weighting scheme that highlights the effect of some variables and diminishes others.

2. *Graphic Representation*

The successive, hierarchical results of the quantitative model described above can be illustrated in Figure 10. In this figure, depicting a fairly simple example, individuals 1 and 2 are rated as most similar based on measured variables and so form the first cluster. In the second stage individual 3 is added to the first cluster, but in the third stage individuals 4 and 5 form a new cluster. In the fourth stage, individuals 7 and 8 form a cluster, with individual 6 added to this cluster in the fifth stage. In the sixth stage the first cluster (1, 2, and 3) is combined with the second (4 and 5), and in the final stage the third cluster (6, 7, and 8) is added to form an all-inclusive cluster.

If we could agree that the observations displayed in Figure 10 constitute two clusters and could then display the two clusters in a two-dimensional frame (discriminant analysis provides such a frame), then we might find that the clusters differ from each other in potentially meaningful ways. The depiction in Figure 11 of two clusters calls attention to several concepts used to differentiate the appearance of clusters formed by cluster analysis (Aldenderfer and Blashfield, 1984). First, the *density* of clusters refers to the degree to which the members of a cluster are closely grouped. Second, *variance* is the complementary concept of density, referring to the dispersion of members away from the center of a cluster. Third, clusters differ with regard to *shape*.

Using these definitions, cluster 1 in Figure 11, as a cluster of five variables, displays greater overall variation than cluster 2, which in turn is characterized by greater overall density. As for shape, the shape of a cluster is often spherical, as depicted by cluster #2 in Figure 11 (referred to as hyperspherical if the space in which the clusters are presented involves more than three dimensions; Ward's method, described below, tends to yield spherical clusters). Other

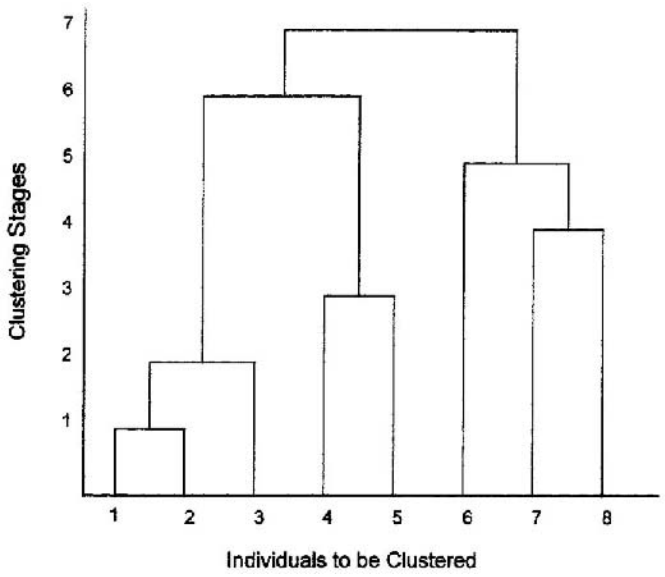


FIGURE 10 Dendrogram for hierarchical clustering.

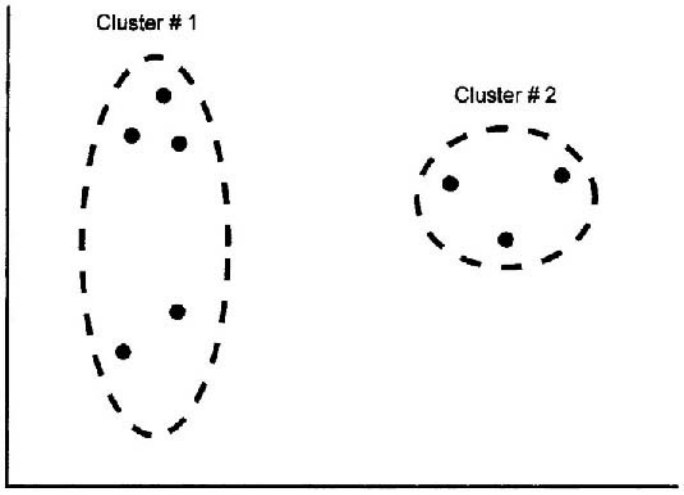


FIGURE 11 Illustration of cluster properties.

common shapes include cylindrical clusters, as depicted for cluster 1 in Figure 11 (these elongated clusters tend to result from the single-linkage approach described in our example).

3. Clustering Solutions

The example above used one particular solution for defining clusters. Without intending to provide a full explanation of all alternatives, we need to introduce other methods of operationalizing our goal of categorizing entities into most appropriate clusters. Addressing the alternative solutions is important for two reasons: (1) none of the available solutions can claim an objective foundation and (2) none of the available solutions appears appropriate for all circumstances. With regard to the first of these reasons, Milligan (1981, p. 380) points out: ''None of the clustering methods currently in use offer any theoretical proof which ensures that the algorithm will recover the correct structure.'' The second reason follows from Monte Carlo simulations that yield conclusions that conflict with other, similar simulations. Milligan (1981) summarizes these sometimes conflicting results, but as examples of his findings, Ward's solution, mentioned below as a hierarchical model, appears to be particularly appropriate when the number of entities per cluster is approximately equal and the clusters overlap with each other.

Given this context of the importance of using solutions that are appropriate for the circumstances, below we describe three of the major approaches to defining clusters: (1) hierarchical methods; (2) iterative partitioning methods; and (3) factor analytic methods. Other methods, such as density searching and graphical methods, are described in Aldenderfer and Blashfield (1984) and Dillon and Goldstein (1984).

Hierarchical methods, as illustrated in Figure 10, create a hierarchy of categories such that entities belong to clusters which, in turn, belong to higher-order clusters. These hierarchies can be created either by (a) agglomerative methods which begin with the separate entities and form clusters by joining at successive steps the entities most similar or by (b) divisive methods that begin with an all-inclusive cluster and successively divide the cluster into distinct subclusters and eventually into individual entities. Divisive methods require substantially more computer resources as the sample size increases and so are used less frequently (Rapkin and Luke, 1993).

Among the agglomerative hierarchical models, the simplest criterion for joining entities and clusters together is the single linkage rule implicitly displayed in Figure 11. Under the single linkage rule, an entity is joined with an existing cluster if the entity is sufficiently similar to any of the cluster members. In contrast, the complete linkage rule ''states that any candidate for inclusion into an existing cluster must be within a certain level of similarity to all members of that cluster'' (Aldenderfer and Blashfield, 1984, p. 40). Intermediate to these two inclusion rules, average linkage joins entities with clusters if the average of the similarities of the entity with the cluster members is sufficiently high. Ward's method produces hierarchical categories by creating clusters that minimize the within-cluster variation for the set of clusters, an approach in line with the definition of meaningful categories that was presented above when discussing the logic of cluster analysis (Dillon and Goldstein, 1984).

Iterative partitioning methods, often referred to as k-means clustering, begin with an initial set of clusters, calculate the multidimensional centers of the clusters (the centroids), and then iteratively reassign entities to clusters so that all entities belong to the cluster whose center is nearest to them. The strength of the iterative partitioning methods is that they allow entities to be reassigned as the analysis proceeds. The disadvantage to this approach, as with most iterative quantitative methods, is that the iterations may lead to convergence on only a local optimum, making the method overly sensitive to the initial set of clusters that is chosen. In the SAS system, FASTCLUS is the procedure that applies the k-means method to clustering objects based on the characteristics measured by the variables.

Factor analytic methods are in effect categorical versions of factor analysis. The goal of these methods is to group things on the basis of dimensions that account for maximal variance. When attempting to group variables, this approach is parallel to using factor-based scores in which one uses factor analysis to identify the variables that load together and then considers each of related variables as equal members of a cluster of variables. In the SAS computer system, the VARCLUS procedure is a factor analytic technique that clusters variables.

This overview of available solutions is sufficient to introduce the techniques; we can elaborate on these concepts by illustrating their application in the context of the Resource Mothers Program. By following the guidance offered here, readers will be able to use cluster analysis in a meaningful way with their own data, but the nature of the data to be clustered has important consequences on the relative usefulness of the various options available. Those wishing to make more informed decisions about which of the various computer options are desirable for their specific requirements need to consult the texts and articles cited above.

B. Application of Cluster Analysis to the Resource Mothers Program

We provided a graphical example of the hierarchical solution in Figure 10. In what follows we apply the other two described approaches, iterative k-means and factor analysis approaches, to the data from the evaluation of the Resource Mothers Program. The SAS computer commands for the iterative and factor analytic variants are listed in Table 14. The command PROC FASTCLUS is for k-means clustering of observations, and the PROC VARCLUS command is for the factor analytic approach to clustering variables. The N = 4 statement limits the number of clusters to a maximum of four. The ITER = 10 specifies a maximum of ten iterative passes in search of a solution that converges on what is at least a local optimal solution.

1. Types of Clients

One of the most important questions for a program such as the Resource Mothers Program is whether it is being used effectively and ethically for all clients. One way to address this question

TABLE 14 SAS Computer Commands for Cluster Analysis (for Objects and Variables)

Computer commands	Functions of commands
PROC FASTCLUS N = 4 ITER = 10; VAR ⟨variables⟩;	FASTCLUS clusters objects; N = 4 specifies the number of clusters of objects to be retained; ITER = 10 refers to the maximal number of iterations to be used; VAR refers to the variables used.
PROC VARCLUS N = 4 ITER = 10; VAR ⟨variables⟩;	VARCLUS instructs the computer to cluster variables rather than objects; VAR refers to the variables being clustered.

is first to identify empirical clusters of clients so that one might then understand the differing needs of these identified groups and evaluate the program effect on them. Recall that grouping objects, people in this case, based on variables is the reverse of the traditional factor analysis problem of grouping variables based on objects. This reversal raises the question of the number of variables to include in a cluster analysis. For example, some scholars have been concerned with having too many variables included in the analysis, particularly when the variables are highly correlated, and have noted the use of principal components analysis prior to cluster analysis as a way to reduce the number of variables and to make sure that the resulting factors are uncorrelated (see Rapkin and Luke, 1993). Others, however, have emphasized that variables assume the role of observations when clustering objects (Bailey, 1994). Because of this use of variables in the role of observations, we would like in this variant of cluster analysis to have many more variables than objects observed. Not having enough variables available in this example, we move in that direction by using all of the eleven variables presented in Table 2.

Using the iterative k-means approach with this expanded set of variables, we can identify any number of nonhierarchical groups of clients. Decisions on the appropriate number of clusters parallel decisions regarding the number of factors or principal components to retain. Rapkin and Luke (1993) outline the primary methods used in supporting this decision. For example, Lathrop and Williams (1990) discuss the use of inverse scree tests. Alternatively, believing that clusters should represent distinct categories of entities, ANOVA could be used to confirm that clusters do entail significant differences on key variables (Rapkin and Luke, 1993). If the differences were not significant, one would try solutions with more clusters.

The SAS program produces several quantitative indicators to aid in this decision. One of the indicators is an estimate of variance accounted for (the R-Squared) by the cluster solution, ranging from 0.0 to 1.0. Two other indicators, pseudo F statistic and cubic clustering criterion, use the information on variance accounted for and the number of clusters to provide information on the relative adequacy of the cluster solution. These two criteria are used by selecting the number of clusters that maximize the values of these statistics. Milligan and Cooper (1985) found that the cubic clustering criterion, developed for SAS, was the sixth best of the thirty criteria that they tested for accuracy in replicating a known cluster structure with well-defined clusters. In addition to these quantitative methods, however, it remains important to consider, as we did with factor analysis, the interpretability of the clusters: the clusters derived by quantitative analysis should correspond to some sense that we have about the natural distinctions among the phenomena being studied.

Having tried different numbers of clusters, a four-cluster solution was chosen for our data from the Resource Mothers Program. It turned out that the three-cluster solution maximized the

pseudo F and cubic clustering criteria that we just described (with the four-cluster solution being second best), but the four-cluster solution was better at producing recognizable groups with significant differences. Before describing the resulting four clusters, Table 15 presents the iterations of the cluster analysis, demonstrating the logic of reassigning individuals to the groups with the closest center until the solution converges and minimal reassignments that no longer influence the group centers.

The numbers in Table 15 represent the changes in the cluster means that result from each round of the iteration process: after individuals are assigned to clusters, the means of the clusters change, resulting in some individuals being reassigned among these adjusted clusters, resulting in additional changes in cluster means, then resulting in additional reassignments of individuals, et cetera. For example, in the second row of Table 15 we see that the mean of Cluster 1 changed by 0.067 as a result of the reassignments of people to clusters during the second iteration; the mean of Cluster 2 changed by 0.098 during this iteration, Cluster 3 by 0.079, and Cluster 4 by 0.047. Note that each of these changes is larger than 0.02, the default criterion of an insignificant change. In contrast, during the fourth iteration only one of the cluster means change by as much as 0.02 (the change of 0.026 for Cluster 4). Iterations continue until all changes in the cluster means are below 0.02 (or whatever criterion is chosen). Because none of the changes is as large as 0.02 during the eighth iteration (the largest being 0.018 for Cluster 2), the analysis is said to have converged on a stable solution (we set the maximum number of iterations at 10 in the computer program displayed in Table 14, but this could have been increased easily had more iterations been required for convergence). These recalculations of cluster means and reassignments can occur in two ways. The default option is to enact these adjustments after each iteration (the option used here for illustration); the alternative in SAS is to make the necessary adjustments after each member is assigned to a cluster, specified in SAS by adding DRIFT after the PROC FASTCLUS command.

Once it is confirmed that the model converges properly, we can go about interpreting the attributes that define the identified groups. Table 16 presents the means of the eleven variables for the four clusters produced by the analysis of the 196 clients (a more thorough examination would require including also the standard deviations of the variables). Looking for attributes that distinguish the four groups, Cluster 1 has the most members (89 clients in this group) and seems to reflect a fairly typical program client: 80% African-American, with somewhat average scores on variables measuring the age of the mother, prenatal weight gain of mother, gestational age, and birthweight. Most distinguishing about this group is that it has the highest percentage of mothers giving birth to their first child (only 13% of the mothers having had prior births). Cluster 2, on the other hand, consists of 27 clients who are distinguishable as primarily the

TABLE 15 Iterative Convergence for k-Means Cluster Analysis (Criterion = 0.02)

	Relative change in cluster seeds			
Iteration	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	.664	.462	.584	.613
2	.067	.098	.079	.047
3	.045	.052	.027	.046
4	.019	.019	.000	.026
5	.016	.000	.010	.023
6	.003	.000	.021	.011
7	.010	.000	.000	.024
8	.011	.018	.000	.014

TABLE 16 Variable Means for Identified Clusters

	Cluster 1: average clients	Cluster 2: older clients	Cluster 3: younger clients	Cluster 4: prior births
Months of prenatal care	7.3 months	7.4 months	6.8 months	4.8 months
Age of mother	17.2 years	24.1 years	16.9 years	17.9 years
Ethnicity	80% African-American	22% African-American	93% African-American	85% African-American
Birth order	13% prior births	66% prior births	57% prior births	73% prior births
Weight gain	28.9 pounds	29.3 pounds	17.9 pounds	22.1 pounds
Education of mother	10.11 years	12.4 years	9.7 years	10.5 years
Medical prenatal visits	10.1 visits	12.1 visits	7.9 visits	5.6 visits
Marital status	100% single	18% single	93% single	90% single
Source of prenatal care	24% private	63% private	36% private	29% private
Birthweight	3384 grams	3388 grams	2376 grams	3110 grams
Gestational age	39.7 weeks	39.6 weeks	36.2 weeks	36.2 weeks
Cluster size (# of clients)	89	27	26	52

older clients who were seen at one of the program sites. This site, prosperous and suburban, had the lowest percent of African-American clients and the highest percent of clients receiving prenatal care from private physicians and other specialists. Associated with their increased age, this grouping is distinguished also by the related attributes—more education and more likely married.

Cluster 3 represents the group of greatest concern. This group, with 26 clients, is the youngest (average age under 17) and has somewhat low averages on health activities such as number of prenatal visits (7.86) and months of prenatal care (6.82). But most disturbing is the relatively low average gestational age of this group (36.18 weeks) and the low average birthweight (2376 grams; babies less than 2500 grams are classified as ''low birthweight'' deliveries). Cluster 4, the second largest cluster with 52 clients, also raises concerns, but different concerns from those of the third cluster. Gestational age is average for program participants and average birthweight is higher than the third cluster, but this fourth cluster is distinguished by the lowest averages on the two health activities, only 4.85 months of care and 5.56 medical visits prenatal visits. Associated with this poor attention to the health needs of the developing baby is not a lack of experience but rather the prior experience of motherhood—73% of the clients in this cluster have had previous babies, the highest percent of all four clusters. It appears that, consistent with previous research, second-time mothers (and third-time, etc.) in this at-risk population are less concerned than first-time mothers about obtaining proper prenatal care (perhaps because their prior births were fairly successful with minimal effort) and so have the weakest statistics for the health activities being monitored.

Because we wanted to use more variables than was used above for factor analysis and because we were interested in clusters that emphasized the groups with the greatest needs, we clustered together the demographic variables with the birth outcome variables of birthweight and gestational age. Alternatively, one could cluster the variables that would typically be used in regression analysis to predict outcomes, in this case the demographic variables, and relate the resulting clusters to the outcomes, in this case the birth outcomes (Rapkin and Luke, 1993).

2. Clusters of Variables

Just as types of clients can be identified, we may also group together variables into clusters. As described above, this approach is most similar to the use of principal components analysis

TABLE 17 Determining the Number of Clusters

	Number of variables in each cluster	Total variation explained	Percent explained
One cluster	6 variables	2.15	36%
Two clusters	3 and 3 variables	3.47	58%
Three clusters	3, 2, and 1 variables	4.43	74%
Four clusters	2, 2, 1, and 1 variables	5.13	86%

and factor analysis in organizing variables. The procedure used for this clustering purpose, VARCLUS in SAS, uses an R-squared analysis to group variables with other variables and clusters of variables. Beginning with a single grouping of all variables, VARCLUS successively separates those variables that fit least with the existing clusters.

Table 17 displays the R-squared information for the solutions involving one, two, three, and four cluster solutions. We see that the one-cluster solution accounts for 36% of the variable variance (2.15 explained out of a total variance of 6.0), with 58% explained by the two-cluster solution (3.47 out of 6.0), 74% for the three-factor solution (4.43 out of 6.0), and 86% for the four-cluster solution (5.13 out of 6.0). As with the other techniques reviewed in this chapter, this approach to cluster analysis requires us to choose the appropriate balance between parsimony (few clusters in this case) and fidelity to complexity (many clusters). If parsimony were paramount and we were satisfied with explaining less than 60% of the variance, we might choose the two-cluster solution displayed in Table 18. Note from the table that this analysis results in two groups of variables, groups that correspond to the results of principal components and factor analysis. Just as factor analysis differentiated variables that related to the age and maturation of the client (mother's age, mother's education, and marital status) from those that involved client health behaviors (month care began, number of medical prenatal visits, and weight gain during pregnancy), so too, does cluster analysis.

Table 19 presents the results of the four-cluster solution. We saw in Table 17 that R-squared for the clusters increases to over 85% when the four-cluster solution isolates into new clusters the two variables that least fit the previous two clusters. As such, the four-cluster solution accounts for considerable variance but at a cost of parsimony. Looking more carefully, we see that the four clusters also correspond to the results of the principal components and factor analyses. In both of those dimensional analyses of the six variables, two variables formed the core of each of the two derived dimensions with each dimension associated with a third, less closely related, variable. The four-cluster solution replicates this pattern with the core variables forming

TABLE 18 Two-Cluster Solution for Factor Analytic Clustering

Cluster and variables	R-squared with own cluster	R-squared with next closest cluster
Cluster 1		
age of mother	.81	.04
education of mother	.72	.02
marital status of mother	.46	.02
Cluster 2		
months of prenatal care	.60	.04
medical prenatal visits	.76	.01
weight gain	.13	.01

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 19 Four-Cluster Solution for Factor Analytic Clustering

Cluster and variables	R-squared with own cluster	R-squared with next closest cluster
Cluster 1		
age of mother	.85	.20
education of mother	.85	.10
Cluster 2		
months of prenatal care	.72	.04
medical prenatal visits	.72	.05
Cluster 3		
weight gain	1.00	.01
Cluster 4		
marital status	1.00	.20

two-variable clusters and the two less-related variables forming one-variable clusters. In this sense the two-cluster and four-cluster solutions support the reliability of each other, and of the principal components and factor analyses, by yielding corresponding structures.

C. Summary Assessment of Cluster Analysis

We have presented cluster analysis as a flexible technique for identifying groups within data. As with principal components analysis and factor analysis, we now want to provide a summary that reiterates the stances taken by cluster analysis in organizing phenomena and reviews its strengths and areas of concern.

The most obvious contrast between cluster analysis and the other two methods in this chapter is that it results in categories rather than dimensions. This is an important distinction in that it reflects a belief that discrete categories are at least as useful as continuous dimensions in making sense of a particular domain (one can, however, subsequently use discriminant analysis to derive dimensions that serve to differentiate the clusters).

With regard to the goal of analysis, cluster analysis is somewhat intermediate to principal components analysis and factor analysis in its stance on realism. Most approaches to cluster analysis are based on realism and the associated beliefs that there are real categories among the phenomena of interest and it is, therefore, the task of cluster analysis to reveal those real categories. Bailey (1994), however, notes that this realist stance is not universal and that some approaches attempt only to yield clusters that simplify variations among phenomena.

Finally, cluster analysis is similar to the other two methods in that it can be used to organize phenomena in terms of any of the three dimensions of the data cube—entities, attributes, or occurrences (though typically employed for clustering entities). We illustrated clustering people and variables, and clustering occurrences can be approached in the same way (e.g., a particular Resource Mothers Program could be assessed on a number of variables that were measured quarterly over a period of ten years).

1. Strengths

One of the primary strengths of cluster analysis is that it is simple. This simplicity is of value not only because it requires less computational time (of less concern these days) but also because it requires few assumptions. Not only is the assumption of a multinormal distribution not necessary, one need not even presume a specific measurement model. A second strength of cluster

analysis is that it will find a structure for a data set. This is to say that to the extent that there are clusters of variables or of entities, cluster analysis will likely detect it. A third strength of cluster analysis is its diversity. We have seen that there are a variety of approaches, each designed to fulfill a somewhat different purpose. As such, there are variations of cluster analysis available for different disciplines and that are appropriate for different presumptions about the underlying clusters that are being estimated.

To appreciate this variety of approaches to cluster analysis, more reading is required. A particularly useful overview is the monograph on classification by Bailey (1994) that we have cited repeatedly. This resource is written in a non-mathematical manner but is analytical in the sense of providing frameworks for understanding the many choices required in choosing an approach to cluster analysis. For example, Bailey (1994) presents 15 criteria, several of which are addressed above but many are not, to consider in selecting a clustering technique (pp. 40–48) and a typology of clustering techniques (pp. 48–50).

2. *Concerns*

We have seen in each of the earlier techniques that the strengths of the method tend to entail particular weaknesses, and so it is also with cluster analysis. One of the main concerns about cluster analysis, following from a strength, is that most of the available approaches ''are relatively simple procedures that in most cases, are not supported by an extensive body of statistical reasoning'' (Aldenderfer and Blashfield, 1984, p. 14). Thus, we cannot rely on formal theory to ensure that our choices in using the available techniques are warranted. A second concern with cluster analysis is that its intended logic is more structure-seeking than structure-imposing, but its quantitative implementation tends to be more structure-imposing. This means that cluster analysis will result in clusters whether there is any real basis for the derived clusters or not.

A third concern is that, as with factor analysis, the diversity of the approach means that some of the techniques are quite different and will result in quite different notions of the appropriate clusters to be derived from the data. For example, had we reported the results of an analysis of the Resource Mothers Program data with the DRIFT option of the SAS FASTCLUS procedure (recall that this results in recalculating cluster means after each individual is assigned to a cluster), we would have seen different clusters. In part these differences across methods result from the different disciplines that came together to form the domain of cluster analysis: (1) biology for the hierarchical methods and (2) social sciences for the k-means and iterative approaches. But even within the social sciences there are conceptual barriers such that developers of alternative techniques and their followers rarely cite those outside their group (Blashfield, 1980)

These concerns—lack of a foundational theory of cluster analysis, possibility of imposing artificial structure, and observed differences in the results produced by the available methods—make it essential that some effort is made to validate the clusters that result from one's analysis. We have presented some evidence of validation, but more could have presented that would call into question the validity of the clusters. Not having devoted the pages necessary to illustrate the variations in results that follow from method choices, it is important to outline what can be done to strengthen our confidence in the proper use of cluster analysis.

Most of the texts on cluster analysis describe validation procedures (e.g., Aldenderfer and Blashfield, 1984), but Humphreys and Rosenheck (1995) provide a particularly sophisticated example of cluster validation. In their approach, which they refer to as sequential validation, one first assesses replicability with subsets of your data, ''If the same clustering procedure generates completely different structures on random subsamples of the data, this may be an indication that no 'real' subgroups exist in the sample'' (Humphreys and Rosenheck, 1995, p. 79).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

If replications are consistent, one can have some faith in reliability and begin assessing validity by comparing clusters on external variables, variables not used in the clustering but on which the clusters should differ. Once it is established that the clusters differ on external variables, the generalizability or external validity of the structure is then assessed by applying cluster analysis to samples of other populations. The logic of sequential validation is that this sequence of assessing reliability and validity is repeated for several of the major options in cluster analysis, with the method that provides the best results being used for the final analysis.

V. CONCLUSIONS

We have described three techniques for organizing phenomena: principal components analysis, factor analysis, and cluster analysis. One of the central themes of our analysis is that the three techniques were developed to serve distinct needs. We want to emphasize, therefore, a framework that will help users choose the techniques best suited to addressing their needs. Figure 12 presents the three techniques in terms of the three choices in organizing phenomena that were discussed at the beginning of this chapter. These broad choices do not address the many operational decisions that must be made when using any of the quantitative techniques, but the set of choices is presented as framework that can orient those becoming acquainted with these methods. We turn now to discuss these three choices, not in a formal order that represents a sequential logic of research but rather in an order that simplifies presentation of some key distinctions among principal components analysis, factor analysis, and cluster analysis.

A. Measurement Scale: Dimensional Versus Categorical Analysis

Whether or not it is the first decision confronting those interested in using these quantitative techniques, the most obvious distinction among the three procedures discussed in this chapter is the difference between techniques that yield categories or groups and techniques that yield

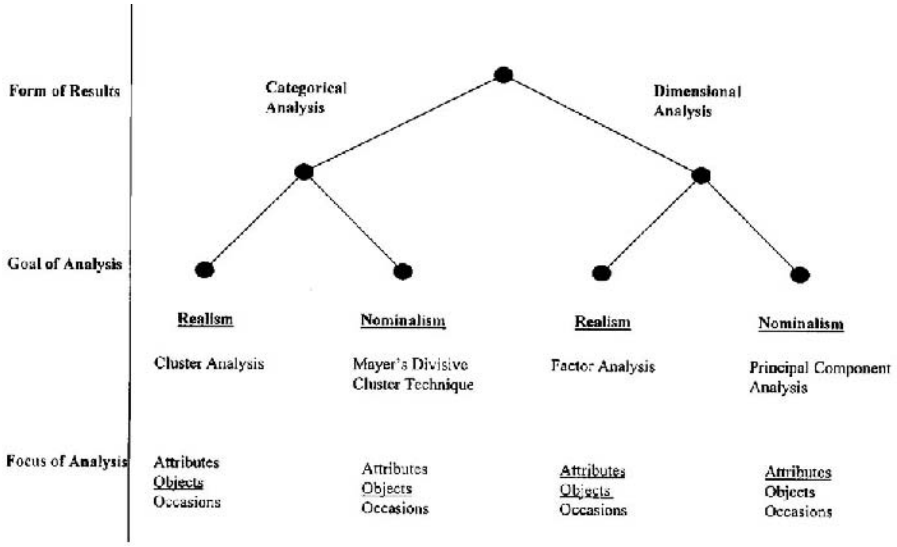


FIGURE 12 Choices in organizing phenomena.

dimensions. Both principal components analysis and factor analysis produce dimensions that order phenomena; cluster analysis groups phenomena into categories. In choosing between categorical and dimensional organizations, two major considerations should influence the decision: (1) your view of the nature of the phenomena of interest and (2) your view of the value of simplicity in your analyses.

The debate between those who see the world in terms of categories and those who see the world as varying along continuous dimensions is persistent and pervasive. In psychology, one form of the debate was between those who studied *personality types* (e.g., ''Type A personality'') and those who identified *personality dimensions* along which people differ (Digman and Takemoto-Chock, 1981). An example of the debate in public administration is the research cited earlier by Coursey and Bozeman (1990). Rather than accept a categorical notion of ''public organization'' as a discrete type, they proposed a continuous dimension of ''publicness'' along which organizations differ (based on a more contextual assessment of organizational attributes).

Some may contend that dimensional analyses in public administration are always superior to categorical ones, based on the belief that our social reality is fundamentally non-categorical (or may argue that categorical analyses are always to be preferred for other reasons). A more realistic alternative, however, would seem to be to respect the two traditions, categorical and dimensional, as having emerged in response to particular needs of analysis, each being more appropriate for certain circumstances. Indeed, many now call for hybrid approaches that combine the strengths of each stance (Skinner, 1979) This perspective places upon you, as the person directing the inquiry, the responsibility of understanding the nature of the phenomena in public administration that you wish to understand. The point of the decision depicted in Figure 12 is that you have a choice in any analysis that you conduct as to whether a categorical or dimensional approach is more appropriate and, therefore, should choose accordingly between cluster analysis and its dimensional alternatives of principal components analysis and factor analysis.

Complicating the categorical-dimensional decision is the value you place on the simplicity of the analyses. That is, you might believe that your phenomena of interest is better described by continuous concepts than categorical ones and still choose a categorical analysis because the greater simplicity compensates for the decreased fidelity. The assumptions of cluster analysis are simpler than those of factor analysis and many find interpretation of clusters easier than interpretation of factors. As such, your decision to use categorical or dimensional techniques can be influenced by your assessment of whether the added information in dimensional analyses is justified given the decreased simplicity. For example, Bailey (1994) points out that some people use factor analysis to identify variables that covary and then group the variables together to form a category. In such uses, little information is gained to justify the added complications of factor analysis.

B. Goal of Analysis: Nominalism Versus Realism

Even more fundamental a decision, though less often addressed explicitly, is the choice between nominalism and realism. Realism refers to the position that there are underlying constructs, such as intelligence, leadership ability, and job satisfaction, that exist in some meaningful way but cannot be measured directly (see Julnes and Mark, 1998). Nominalism, in contrast, resists what is seen as reification; what are constructs for the realist are for the nominalist merely convenient labels. As with the prior choice, there are two major considerations that should influence whether one chooses to use a quantitative technique that seeks to identify underlying patterns: (1) one's view of the nature of the phenomena and (2) the value one places on simplicity of analysis.

As for the nature of phenomena (in the sense of our experience of the world), the contro-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

versy involved is longstanding. During the middle of the twentieth century ''empiricism'' was understood by some to mean that we should rely on direct observation and not presume unobservable constructs. This interpretation led to a nominalist view of the categories and dimensions that we use to organize phenomena—that it is useful, for example, to talk of ''leadership ability,'' but all we really see are a variety of behaviors performed by persons in leadership positions, the ''ability'' construct being just our way of creating parsimony out of chaos.

This view of empiricism was gradually abandoned as it became clear even to proponents that direct observation, not mediated by constructs, was a mythical ideal and, further, was insufficient for making sense of our world in a meaningful way (Meehl, 1986). As a result, construct validity, the extent to which a measure in some way reflects the underlying, unobservable construct that it is intended to measure, became an established part of social inquiry. As a result, most research methods texts with a quantitative orientation include a section on measurement theory and the role of construct validity (including coverage of Cronbach's alpha for scale construction). In recent years, however, some have again questioned the meaningfulness of measures of underlying constructs, believing that the constructs that we choose to use in organizing phenomena are fundamentally arbitrary and reflect only our projection upon the world (Goodman, 1996).

Choosing a position in this controversy leads to choices in methods. One way to frame this decision is to ask whether you want your quantitative methods to identify underlying patterns that are not directly observable. If you opt for the nominalist stance, you will want your quantitative methods to analyze observed variation, believing that to provide the most faithful information about the world. If, on the other hand, you view constructs such as ''fiscal stress'' and ''organizational commitment'' as meaningful in making sense of the world, then you will choose quantitative methods that presume that there are natural patterns in the social world and attempt to reveal them. Factor analysis supports the realist stance by distinguishing between common variation, and unique variation, as a result focusing on covariance rather than variance. In contrast, the lack of realist presumptions is characteristic of principal components analysis, a technique that can be consistent with a nominal world.

As indicated in Figure 12, factor analysis and most variations of cluster analysis presume, in their own way, an underlying structure that is to be revealed. In contrast, principal components analysis makes no assumption of real, underlying patterns. The case of cluster analysis, mostly based on realism, can be addressed with the distinction Bailey (1994, p. 41) makes between natural (realist) and artificial (nominalist) clustering.

> [V]irtually all agglomerative and most divisive methods have the goal of seeking natural, underlying clusters. Few numerical taxonomists would claim that they are seeking artificial clusters. One salient exception is Mayer's (1971) divisive method designed to create artificial clusters.

Mayer's view is that the value of cluster analysis does not depend on it being applied to real clusters, even if identifying real clusters remains a valued goal. ''The goal is to create a taxonomy although it may be clear that no obvious natural clustering exists. If an obvious natural clustering does appear then, of course, the researcher is advised to use that clustering'' (Mayer, 1971, p. 146).

The second consideration, as before, is about simplicity. Related to our first consideration but more of a tactical rather than strategic issue in research methods, one might believe that there are underlying patterns that we wish to understand but have concerns about the assumptions required in factor analysis and so prefer to use principal components analysis. The logic of this reluctance is that the assumptions required for the realist project of factor analysis might introduce so much ill-considered bias that the realist goal becomes less attainable than it would be

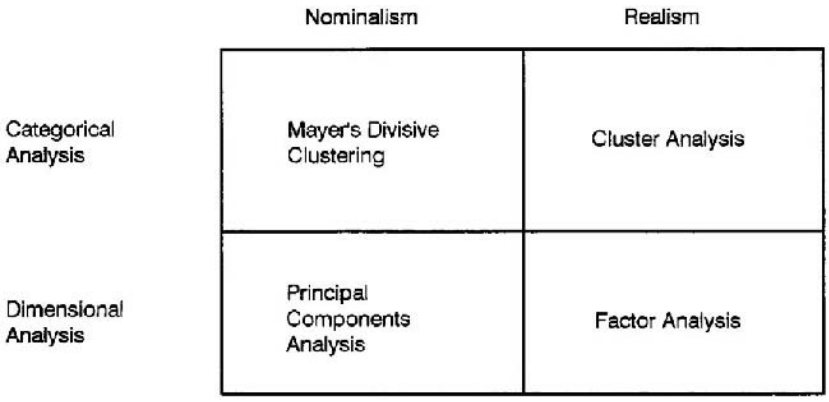


FIGURE 13 Two-dimensional model of organizing choices.

with principal components analysis and its more neutral assumptions. In support of this position, we noted that Nunnally (1978) preferred principal components analysis over factor analysis because he was skeptical about such things as the communality estimates that are required as initial assumptions in factor analysis.

Combining the nominalism-realism decision with the categorical-dimensional decision results in the 2×2 matrix depicted in Figure 13. We see that each of the three techniques is appropriate for different positions in the matrix. As discussed when addressing these two dimensions, you have the opportunity and responsibility as someone familiar with the phenomena that you are studying to make the necessary decisions when choosing the quantitative techniques that are most appropriate for your particular tasks.

C. Phenomena of Interest: Attributes, Objects, and Occurrences

The last decision depicted in Figure 12 relates to the type of administrative phenomena that are of greatest interest for those designing the studies. We introduced this topic at the beginning of this chapter in terms of a data cube comprised of three dimensions: attributes, objects, and occurrences. We now consider the relationships between these three potential dimensions of data and the three techniques of principal components analysis, factor analysis, and cluster analysis. The simple answer to the question about relationships is that each of the techniques can be used to study objects, attributes, and occasions. A longer answer is that there is greater use of some techniques for particular uses and that these differences reveal our inclinations in organizing phenomena. In the context of the Resource Mothers Program, we can think of the data cube in terms of three dimensions composed of (1) selected demographic and health variables (attributes or characteristics) measured for (2) numerous program clients (objects) at (3) specified intervals (occasions; assuming for now that the Resource Mothers Program was studied over many years). Each of these three dimensions can be the focus of analysis in one of two ways. The two alternate approaches for each of the three dimensions yields six possible types of analysis (labeled R-analysis, Q-analysis, O-analysis, P-analysis, S-analysis, and T-analysis by Cattell, 1952). Below, beginning with the attribute dimension, we illustrate one of these alternate analyses for each of the dimensions of the data cube.

1. Attributes

The predominant way in which attributes, or characteristics, have been studied is in terms of the data matrix presented in Table 20. This table represents the way that most people code

TABLE 20 R-Analysis: Organizing Attributes Using Multiple Individuals

	Prenatal visits	Months of care	Weight gained	Mother's age	Mother's education	Marital status
Client 1	6	4	19	16	10	0
Client 2	12	7	27	19	12	1
Client 3	8	6	33	17	11	0
Other clients	. . .	. . .	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .	. . .	. . .
Client 196	9	7	25	16	10	0

data for computer analysis—columns representing different variables measured for the rows of individuals in the matrix. Here the matrix is presented as 196 program clients measured on six demographic variables. Each of the techniques discussed is capable of using this data set to identify relationships among the variables. Principal components analysis and factor analysis use the variables measured to create a new variable, a linear composite for one and a measurement of a construct in the other. This type of analysis, organizing the attributes of individuals into clusters or dimensions, is referred to as R-analysis. While we will not discuss it here, the other approach to organizing attributes, referred to as P-analysis, is to measure the attributes for one organization on many time occasions.

We discussed at length the use of factor analysis in service of R-analysis; less was said about R-analysis for cluster analysis, limiting ourselves to describing the use of the VARCLUS procedure in SAS for grouping variables into either two or four clusters. The reason for this is that cluster analysis is used less often in relating variables, perhaps because we are less inclined to presume ''types'' of variables. The issue here is the categorical-dimensional distinction and the value we place on the information provided by the two approaches. Although the SAS VARCLUS computer output reports the R-squared between each variable and its assigned cluster, this information is lost if one considers only the category assignment as the final result (for an exception to the idea of discrete categories, see the work on overlapping object categories see Bailey, 1994, p. 42). In contrast, a major theme of the construct validity concept in measurement is based on the view that different measures differ in their adequacy in reflecting the construct and need to be treated accordingly. This additional information provided by factor analysis is valuable, however, only if we use it. To the extent that factor analysis is used to group variables into what are viewed as homogenous categories, factor analysis will offer little information beyond cluster analysis and, so, cluster analysis will be equally appropriate.

> Thus, although the factor analysis may yield more information, many times users in effect throw away this extra information and attempt to get their cases or variables to have principal loadings on only a single factor, thus in effect transforming them into a sort of de facto cluster analysis (Bailey, 1994, p. 70).

2. Objects

The typical data matrix for analyzing relationships among objects is as represented in Table 21. In this table, the 196 program clients are arranged as columns with the 11 measured characteristics of the clients presented as rows. The task with this arrangement of the data cube is to identify patterns of clients that share some inherent similarities, an analysis that is referred to as Q-analysis (the alternative approach to organizing objects, referred to as T-analysis, involves measuring one characteristic over multiple occasions).

TABLE 21 Q-Analysis: Organizing Individuals Using Multiple Attributes

	Client 1	Client 2	Client 3	Other clients		Client 196
Prenatal visits	6	12	8	. . .	. . .	9
Months of care	4	7	6	. . .	. . .	7
Weight gained	19	27	33	. . .	. . .	25
Mother's age	16	19	17	. . .	. . .	16
Mother's education	10	12	11	. . .	. . .	10
Other attributes	. . .	. . .	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .	. . .	. . .
Gestational age	36	40	39	. . .	. . .	39
Birthweight	2580	3650	3325	. . .	. . .	3460

In that the matrix associated with Q-analysis is the simple inverse, or transpose, of the matrix presented in Table 20, we mentioned above that Q-analysis reverses the data requirements that we commonly associate with quantitative analysis. Though this leads to the natural conclusion that we need many measures of attributes to conduct Q-analysis, training in research methods can produce the dogma that it is always better to have large samples of people (or other objects) for our analyses, a dogma sufficiently ingrained as to warrant a countering quotation: ''Here [Q-analysis], one should have several times as many variables as objects. Thus we might wish to have a sample of 100 persons, each measured on 400 variables.'' (Bailey, 1994, p. 70). As Bailey notes, social scientists rarely have this many variables available and so regularly perform Q-analysis with data sets more appropriate for R-analysis, as was done in this chapter with only eleven variables relevant for the Resource Mothers Program.

The last point to be made about organizing objects is to reiterate that the focus of analysis is a separate choice from the choice of categorical or dimensional output. Because cluster analysis has been used primarily for Q-analysis, some might equate the two and think first of cluster analysis whenever they want to organize objects. This would be unfortunate as it would limit the organization of objects to categorical groupings. For circumstances where continuous organization of objects is preferable, Q-technique factor analysis is an established approach and is described in Chapter 24 of this volume.

3. Occasions

Whether or not it represents a natural bias, much less work in public administration and in other areas has focused on relationships among occasions. This neglect is understandable (it does, after all, often take many years to gather the required data), but many of the issues confronting public administrators make more sense when we recognize the time-based patterns of the profession, be they yearly budgeting cycles or longer trends in the changing nature of federalism. In addition, we are increasingly aware of the importance of a ''process'' orientation in administration. Not only are managers expected to guide the processes of their organizations for such things as continuous quality improvement, they are also expected to support a training process that prepares their subordinates to guide the organization (Julnes et al., 1987). It turns out that focusing on process is supported when we emphasize research that organizes changes over time. The two relevant alternatives from the data cube are to organize occasions using multiple measures of one entity, O-analysis, or to organize occasions by measuring one attribute of many entities, T-analysis.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 22 O-Analysis: Organizing Occasions (Years) Using Multiple Attributes of One Entity

	1993	1994	1995	1996	1997	1998
Average prenatal visits	7.7	10.7	11.5	12.3	10.5	10.7
Average months of care	5.4	6.5	7.3	7.4	6.7	6.5
Average weight gained	23.7	24.4	26.3	27.2	25.0	24.6
Other attributes	. . .	. . .	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .	. . .	. . .
Average birthweight	3130	3250	3380	3390	3270	3240

Table 22 presents the data matrix for O-analysis, in which the rows are the average characteristics of the clients of the Resource Mothers Programs and the columns are periodic occasions of measurement. The result of organizing occasions in this way would be clusters or dimensions of time periods with different profiles on the measured variables (e.g., periods in which healthy activities and birth outcomes predominate and other periods with unhealthy activities and outcomes). Another example for public administration might be an examination of patterns of urban development and change over the past century (with various measures of urban characteristics); an O-analysis could produce categories of periods in U.S. history with different profiles of urbanization.

VI. SUMMARY

This chapter provides an introduction to three quantitative techniques with examples of use and interpretation. These three techniques were described as important for public administrators because of the complexity that we confront in this field and the resultant necessity to organize the complexity in a meaningful way. This last section of the chapter has attempted to provide a framework to help users select the technique that is appropriate for their particular needs. Underlying this framework is the belief that people tend to use the quantitative techniques with which they are most familiar, even if other techniques are better suited to the task at hand (as the old saying goes, ''for someone who has only a hammer, all the world's a nail''). Asking questions about the three sets of choices described in this chapter (categorical versus dimensional output, realism versus nominalism, and attributes, objects, and occasions) may seem secondary when one has a data set that needs to be analyzed, but differing positions on these issues have real implications for the proper selection of quantitative techniques (implications preferably considered before data are collected). Fortunately, as we have tried to convey, there is a sufficient variety of techniques available to serve most needs.

ACKNOWLEDGMENTS

Primary support for the evaluation described in this chapter was provided by a March of Dimes Birth Defects Foundation Innovations and Research in Health Education and Service Delivery Grant (#3FY92-0265).

REFERENCES

Aldenderfer, M.E., and R.K. Blashfield (1984). *Cluster Analysis*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-044. Newbury Park, CA, Sage.

Bailey, K.D. (1994). *Typologies and Taxonomies*: *An Introduction to Classification Techniques*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-102. Thousand Oaks, CA, Sage.

Blashfield, R.K. (1980). ''The Growth of Cluster Analysis: Tryon, Ward, and Johnson,'' *Multivariate Behavioral Research*, *15*: 439–458.

Buley, J.L. (1995). ''Evaluating Exploratory Factor Analysis: Which Initial-Extraction Techniques Provide the Best Factor fidelity?'' *Human Communication Research*, *21*: 478–494.

Burns, T. and G.M. Stalker (1961). *The Management of Innovation*, London: Tavistock.

Cattell, R.B. (1952). *Factor Analysis*: *An Introduction and Manual for the Psychologist and Social Scientist*, New York: Harper and Row.

Cattell, R.B. (1966). ''The Scree Test for the Number of Factors,'' *Multivariate Behavioral Research*, *1*: 140–161.

Cattell, R.B. (1978). *The Scientific use of Factor Analysis in Behavioral and Life Sciences*, New York: Plenum Press.

Cattell, R.B. and J. Jaspers (1967). ''A General Plamode (No. 30-10-5-2) for Factor Analytic Exercises and Research,'' *Multivariate Behavioral Research Monographs*, *67*(3): 1–212.

Cattell, R.B. and S. Vogelmann (1977). ''A Comprehensive Trial of the Scree and KG Criteria for Determining the Number of Clusters,'' *Multivariate Behavioral Research*, *12*: 289–325.

Clifford, H. and W. Stephenson (1975). *An Introduction to Numerical Taxonomy*, New York: Academic Press.

Coursey, D. and B. Bozeman (1990). ''Decision Making in Public and Private Organizations: A Test of Alternative Concepts of 'publicness,''' *Public Administration Review*, *50*: 525–535.

Cota, A.A., C.R. Evans, R.S. Longman, K.L. Dion, and L. Kilik (1995). ''Using and Misusing Factor Analysis to Explore Group Cohesion,'' *Journal of Clinical Psychology*, *51*: 308–317.

Daft, R.L. and K.E. Weick (1984). ''Toward a Model of Organizations as Interpretative Systems,'' *Academy of Management Review*, *9*: 284–295.

Digman, J. and N. Takemoto-Chock (1981). ''Factors in the Natural Language of Personality: Reanalysis, Comparison, and Interpretation of Six Major Studies,'' *Multivariate Behavioral Research*, *16*: 149–170.

Dillon, W.R. and M. Goldstein (1984). *Multivariate Analysis*: *Methods and Applications*, New York: Wiley.

Dunteman, G.H. (1989). *Principal Components Analysis*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-069, Newbury Park, CA: Sage.

Fuller, E.L., Jr. and W.J. Hemmerle (1966). ''Robustness of the Maximum Likelihood Estimation Procedure in Factor Analysis, *Psychometrika*, *31*: 255–266.

Goodman, N. (1996). ''Words, Works, Worlds,'' in P.J. McCormick (ed.), *Starmaking*: *Realism, Antirealism, and Irrealism*, Cambridge, MA: MIT Press.

Gorsuch, R.L. (1983). *Factor Analysis*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Harris, J.A. (1995). ''Confirmatory Factor Analysis of the Aggression Questionnaire,'' *Behaviour Research and Therapy*, *33*: 991–994.

Hartigan, J.A. (1975). *Clustering Algorithms*, New York: John Wiley and Sons.

Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, Cary, NC: SAS Institute.

Hatcher L. and E.J. Stepanski (1994). *A Step-by-Step Approach to Using the SAS System for Univariate and Multivariate statistics*, Cary, NC: SAS Institute.

Hotelling, H. (1933). ''Analysis of a Complex of Statistical Variables into Principal Components,'' *Journal of Education Psychology*, *24*: 417–441, 498–520.

Humphreys, K. and R. Rosenheck (1995). ''Sequential Validation of Cluster Analytic Subtypes of Homeless Veterans,'' *American Journal of Community Psychology*, *23*: 75–98.

Jolliffe, I.T. (1986). *Principal Component Analysis*, New York: Springer-Verlag.

Julnes, G., M. Konefal, W. Pindur, and P. Kim (1994). ''Community-Based Perinatal Care for Disadvantaged Adolescents: Evaluation of the Resource Mothers Program,'' *Journal of Community Health*, *19*: 41–53.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Julnes, G. and M.M. Mark (1998). ''Evaluation as Sensemaking: Knowledge Construction in a Realist World,'' in G.T. Henry, G. Julnes, and M.M. Mark (eds.), *Realist Evaluation*: *An Emerging Theory in Support of Practice*, New Directions for Evaluation, no. 78, San Francisco: Jossey-Bass.

Julnes, G.D. Pang, N. Takemoto-Chock, G. Speidel, and R. Tharp (1987). ''The Process of Training in Processes,'' *Journal of Community Psychology*, *15*: 387–396.

Kaiser, H.F. (1958). ''The Varimax Criterion for Analytic Rotation in Factor Analysis,'' *Psychometrika*, *23*: 187–200.

Kim, J. and C.W. Mueller (1978a). *Introduction to Factor Analysis*: *What it is and How to Do it*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-013, Newbury Park, CA, Sage.

Kim, J. and C.W. Mueller (1978b). *Factor Analysis*: *Statistical Methods and Practical Issues*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-014, Newbury Park, CA, Sage.

Lathrop, R.G. and J.E. Williams (1990). ''The Reliability of Inverse Scree Tests for Cluster Analysis,'' *Educational and Psychological Measurement*, *47*: 952–959.

Mayer, L.S. (1971). ''A Theory of Cluster Analysis When There Exist Multiple Indicators of a Theoretic Concept,'' *Biometrics*, *27*: 143–155.

Meehl, P. (1986). ''What Social Scientists don't Understand,'' in D.W. Fiske and R.A. Schweder (eds.), *Metatheory in Social Science*: *Pluralisms and Subjectivities*, Chicago: University of Chicago Press.

Milligan, G.W. (1981). ''A Review of Monte Carlo Tests of Cluster Analysis,'' *Multivariate Behavioral Research*, *16*: 379–407.

Milligan, G.W. and M.C. Cooper (1985). ''An Examination of Procedures for Determining the Number of Clusters in a Data Set,'' *Psychometrika*, *50*: 159–179.

Morrison, D.F. (1990). *Multivariate Statistical Methods*, New York: McGraw-Hill.

Mulaik, S.A. (1972). *The Foundations of Factor Analysis*, New York: McGraw-Hill.

Nunnally, J.C. (1972). *Educational Measurement and Evaluation*, 2nd ed., New York: McGraw-Hill.

Pearson, K. (1901). ''On Lines and Planes of Closest Fit to Systems of Points in Space,'' *Philosophical Magazine*, *2*: 559–572.

Rapkin, B.D. and D.A. Luke (1993). ''Cluster Analysis in Community Research: Epistemology and Practice,'' *American Journal of Community Psychology*, *21*: 247–277.

Rummel, R.J. (1970). *Applied Factor Analysis*, Evanston: Northwestern University Press.

SAS Institute, Inc. (1988). *SAS/STAT User's Guide*: *Version 6.03 Edition*, Cary, NC: Author.

Scott, W.R. (1981). *Organizations*: *Rational, Natural, and Open Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Shevlin, M.E. and B.P. Bunting (1994). ''Confirmatory Factor Analysis of the Satisfaction with Life Scale,'' *Perceptual and Motor Skills*, *79*: 1316–1318.

Shevlin, M.E., B.P. Bunting, and C.A. Lewis (1995). ''Confirmatory Analysis of the Rosenberg Self-Esteem Scale,'' *Psychological Reports*, *76*: 707–711.

Skinner, H.A. (1979). ''Dimensions and Clusters: A Hybrid Approach to Classification,'' *Applied Psychological Measurement*, *3*: 327–341.

Sneath, P.H.A. and R.R. Sokol (1973). *Numerical Taxonomy*, San Francisco: Freeman.

Sokol, R.R. and P.H.A. Sneath (1963). *Principles of Numerical Taxonomy*, San Francisco: Freeman.

Spearman, C. (1904). '''General Intelligence' Objectively Determined and Measured,'' *American Journal of Psychology*, *15*: 201–293.

Stanbury, W. and F. Thompson (1995). ''Toward a Political Economy of Government Waste: First Step, Definitions,'' *Public Administration Review*, *55*: 418–427.

Stogdill, R.M. (1974). *Handbook of Leadership*, New York: Free Press.

Spector, P.E. (1993). *SAS Programming for Researchers and Social Scientists*, Newbury Park, CA: Sage.

Thompson, B. and L.G. Daniel (1996). ''Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines,'' *Educational and Psychological Measurement*, *56*: 197–109.

Thurstone, L.L. (1935). *The Vectors of Mind*, Chicago: University of Chicago Press.

Thurstone, L.L. (1947). *Multiple Factor Analysis*, Chicago: University of Chicago Press.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Tyron, R.C. (1939). *Cluster Analysis*: *Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Ann Arbor, MI: Edwards Brothers.

Wood, J.M., D.J. Tataryn, and R.L. Gorsuch (1996). ''Effects of Under- and Overextraction on Principal Axis Factor Analysis with Varimax Rotation,'' *Psychological Methods, 1*: 354–365.

Zoski, K.W. and S. Jurs. (1996). ''An Objective Counterpart to the Visual Scree Test for Factor Analysis,'' *Educational and Psychological Measurement, 56*: 443–452.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

24
Q Methodology

Steven R. Brown
Kent State University, Kent, Ohio

Dan W. Durning
University of Georgia, Athens, Georgia

Sally Selden
Syracuse University, Syracuse, New York

I. INTRODUCTION

Like the other public administration research methods described in the various chapters of this book, Q methodology requires the collection and manipulation of data, and the analysis of these data using a sophisticated statistical technique (factor analysis). And, as with other methods, Q method can be used to explore a phenomenon of interest to gain insight into it, to generate hypotheses, and to test hypotheses.

Despite these traits that Q methodology has in common with the usual body of statistical techniques employed in public administration research, it differs from them in ways that have profound implications for its use. In fact, the designation of this method as ''Q'' is intended to differentiate it from ''R'' methodology, which comprises the statistical methods used for ''objective'' or ''scientific'' research in the social sciences. As discussed in more detail below, the differences between Q and R methods are not simply a matter of technique, they reflect very different philosophies of inquiry that encompass competing epistemologies and understandings of what constitutes sound scientific practice.

Although many Q methodologists justify its use on philosophical bases, rejecting R methods as tools of discredited positivism, other researchers can turn to Q methodology for pragmatic reasons. For the more pragmatic researcher or public administration professional, Q methodology can be viewed as a research or investigatory tool that offers insights often different from those that can be obtained through R methods. The differences in insights provided by Q methodology are illustrated by the example that follows in the next subsection.

In this chapter, we provide an introduction to Q methodology for researchers who know little about it but might like to employ it in their work. We introduce the topic by providing a brief comparison of research using Q and R methodologies and noting the key differences in them. Then, in the next section we provide an overview of the reasons for using Q methodology in public administration research and practice, and we point out some examples of how the method might be employed by researchers and administrators. The third section of the chapter explains in detail, using a case study, how a Q method study is carried out. Also, this section

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

addresses many of the questions that are raised about Q methodology. In the fourth section, we summarize public administration-related studies that have been done using Q methodology and suggest other topics that are good candidates for research using Q method.

A. Comparing Q and R Methods: A Practical Example

Suppose a researcher is interested in how managers of public agencies in another country, say Ukraine, view their jobs, specifically the organization of their work, the methods of control and discipline they use, their attitudes toward change, and their relationships with subordinates and superiors. As part of this study, the researcher would like to investigate whether younger managers have perspectives different from those of older managers. Also, he or she thinks it would be interesting to compare the attitudes of Ukrainian public managers with those of public managers in the United States.

1. The R Method Approach

The common ''scientific'' approach to this type of research would be to formulate hypotheses about the different types of managers, and how the types of managers would vary by location and age. Then, the hypotheses would be tested with data collected through a survey instrument. This survey instrument would contain questions or statements with scales allowing the respondents to indicate their degree of agreement or disagreement. The questions and statements selected for inclusion in the survey would be intended to measure the dimensions of the public manager's job addressed in the hypotheses. For example, to test the hypothesis that older managers in Ukraine are more likely than younger managers to have an autocratic managerial style, the researcher would create a set of questions and statements intended to measure how the manager relates to subordinates (e.g., does he or she order them around or invite participation in decisions).

This survey might be sent to a random sample of public managers in Ukraine, or perhaps a stratified random sample to insure that both younger and older managers are fully represented in the survey. Most likely, however, because of the difficulties in identifying the whole population of public administrators and extracting a random sample, the target population would be the public employees in a particular city or region, and the sample would be drawn from that population. To make it a cross national comparison, a random sample of public managers in the United States, or more likely, managers in a comparable government, would be selected and sent the survey.

After receiving the completed surveys, the data would be entered into some statistical spreadsheet, creating a huge matrix of variable responses by observation. Then, different statistical analyses would be performed. Likely, the responses to different questions or statements would be used to construct dependent variables, which would represent different views of the elements of public management under study (for example, more or less autocratic managerial style). Then, these dependent variables could be included in regression analyses to test hypotheses (for example, with managerial style as the dependent variable, the independent variables might be the age of the respondent, years of experience, educational background, nationality).

2. Using Q Methodology

The researcher could also use Q methodology to investigate this research question and to explore many of the same hypotheses. However, the process of research and the results would look much different, and we would argue, the results would more accurately reflect the richness and complexity of the views of the different managers.[1]

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The process would begin by identifying the communication (what is being said) about the topic of interest, the jobs of public managers. This communication could be identified by interviewing several public managers about their jobs, focusing particularly on the job dimensions of research interest, or it might be done by reading accounts of managers in the two countries about their jobs. From this concourse of communication about the jobs of public managers, a sample would be selected that represented, as far as possible, the diversity of communication in all of its important dimensions. This sample, in the form of 30 to 60 statements, would make up the Q sort to be administered to managers.

The researcher would then ask selected managers to complete the Q sorts. The selection of managers would be intended to insure that those most likely to have different views would be included. A key goal in selecting managers to complete the Q sorts would be to get the largest possible diversity of views. Also, the researcher would purposely select a group of younger and older managers to insure that information would be obtained that would help explore whether the two groups have different views toward their jobs. For a comparative perspective, the sorts would be administered to groups of Ukrainian and American public administrators.

The managers would complete the Q sort by placing the statements in a quasi-normal distribution from ''most agree'' to ''most disagree.'' The sort would be forced, in that the number of cards to be placed in each category, say from +4 (most agree) to −4 (most disagree), would be specified. In this case, there would be nine categories, with the largest number of cards to be placed in the 0 category and the fewest to be placed in the +4 and −4 categories.

As the sort was being completed, the researcher could engage in a dialogue with the sorter, noting questions that he or she raised, comments that accompanied the placement of statements, and reactions to certain statements. Then, the sort could be followed up by asking the manager his or her reasoning in placing statements in the most extreme positions. This discussion would aid the investigator to understand more fully the perspectives of the managers completing the sorts.

When all of the Q sorts were completed, they would be analyzed by first correlating them, then using the correlation matrix for factor analysis. However, because the factor analysis would treat the sorters as variables and the statements as observations, the resulting factors would represent the cluster of managers whose views of public management are quite similar. Using information about how the different clusters of managers completed their sorts, the researcher would then identify and discuss the different views about public management among the managers who completed the sorts.

The analysis of the Q sorts would provide insight into how public managers understand their job. These views would not necessarily conform to any models that were specified a priori, nor would they be forced into categories based on responses to any particular statements whose meaning was specified in advance by the researcher. In fact, before carrying out the analysis of the sorts, the number of different perspectives and their nature would be unknown. Thus, the managers participating in the Q study would determine the number and nature of the perspectives of public management by revealing through their Q sorts their operant subjectivities.

As this discussion indicates, the use of Q methodology serves primarily to identify the common and different subjectivities of the people selected to complete a Q sort. By systematically comparing and contrasting the different common subjectivities (factors), the research can add to knowledge about managerial attitudes and behaviors.

In some cases, a researcher might want to use the Q sort results to explore the extent to which sorters with specific characteristics are clustered in the same factors or are spread out among different factors. For example, with this study of Ukrainian and American managers it is possible to examine the characteristics (e.g., age, experience, nationality) of the managers

in different factors to see if managers with specific characteristics have similar or different understandings of the job of public manager. In this way, it is possible to explore hypotheses, such as the one that older managers would have more autocratic management styles than younger managers. To the extent that older managers sorted together in a factor that could be characterized as more autocratic, and younger managers clustered together in another factor with a less autocratic orientation, the Q study would provide support for acceptance of the hypothesis. (Of course it is possible that a factor is not systematically associated with any demographic variables; such a factor could therefore never be predicted in advance, yet through Q methodology it could be demonstrated to exist.)

It should be emphasized that researchers must be very cautious when using Q sort results to investigate the distribution of (R-type) populations among different factors. This type of investigation should be acknowledged as only suggesting a pattern of common or different viewpoints related to certain demographic characteristics because Q methodology is intended to identify subjectivities that exist, not to determine how those subjectivities are distributed across a population.

B. Summary: Key Differences in Q and R Methodologies

This comparison of how Q and R methodological studies would be carried out on the same topic provides some insight into the key differences in the methodologies.[2] These include:

- Q methodology seeks to understand how individuals think (i.e., the structure of their thoughts) about the research topic of interest. R methodology identifies the structure of opinion or attitudes in a population. Thus, the results of Q method will identify how an individual, or individuals with common views, understand an issue; the results of R methods describe the characteristics of a population that are associated statistically with opinions, attitudes, or behavior (e.g., voting) being investigated.
- While R methods are intended for the ''objective'' analysis of research issues, Q methodology is designed to study subjectivity. R methodology is founded on logical positivism in which the researcher is an outside objective observer. In contrast, Q methodology is more closely related to post-positivist ideas that reject the possibility of observer objectivity and question the assumption that the observer's vantage point is, if not objective, then is in some sense superior to that of any other observer, including the person being observed. Thus, Q methodology is in tune with phenomenological, hermeneutic (McKeown, 1990, in press), and quantum theories (Stephenson, 1983).[3]
- Q methodology is an intensive method that seeks in-depth understanding of how at least one person thinks about the topic of investigation. As an intensive method, Q methodology requires a small number of well selected subjects to complete the Q sort, which is a sample of the communications about the topic of interest. In comparison, R methods are extensive methodologies designed to extract understandings of populations through representative samples of them; thus, they require—depending on the population size and sampling techniques—data from a certain percentage of the population of interest.

II. Q METHOD IN PUBLIC ADMINISTRATION RESEARCH AND PRACTICE

The above example shows how Q method could be used to study a topic that also could be addressed by R methods. The pragmatic justification for using Q methodology, instead of R

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

methods, would be that it could provide more in-depth insights into how managers think, though it could say little about the distribution of views across the population of managers. We think that on this pragmatic basis alone, Q methodology should be in the arsenal of public administrators, planners, and policy analysts. However, the use of Q methodology has a stronger justification as a post-positivist methodology that avoids the fallacies built into the prevalent positivist research methodologies. According to critics of the ''scientific'' method now widely used in social sciences and professions, these methodologies, at best, result in poor information for decisions and, at worst, reinforce power relationships that oppress some groups in society. In this section, we briefly discuss both justifications—pragmatic and philosophical—for using Q method.

A. Pragmatic Uses of Q Method in Public Administration Research and Professional Tasks

Rather than worrying too much about epistemological questions, researchers and practitioners may be satisfied knowing that Q method can supplement R methods, providing additional and different insights into issues and policies of interest. Instead of being forced to choose between positivism and post-positivist perspectives, it may be comfortable to use both of them, reflecting the strategy of methodological triangulation that has been advocated to bring together the use of quantitative and qualitative methods.

This perspective is likely to appeal most to public administration professionals, including managers, planners, and policy analysts, who spend little time worrying about the nature of knowledge but much time worrying about how well they are performing their jobs. They are more likely to find Q method of value if it provides them, efficiently, with insights not available through other methods. Although we do not expect that all professionals are oblivious to the larger philosophical questions, we do believe their interest in the topic is overshadowed by the need to meet client and public expectations.

Public administration professionals are likely to be good candidates for use of Q method for several reasons. First, they are not strong adherents to the scientific method; though they may believe theoretically in the objective investigation of hypotheses to find general laws, they have little time for such endeavors and little expectation—based on experience—that such laws would help them do their jobs. Second, in practice, it is likely that a substantial portion of these professionals are practicing phenomenologists, though some might punch you in the nose if you accused them of being such (Forester, 1990). In the hurly-burly of management, planning, and analysis, professionals understand that events are viewed differently by different people, that the context of their actions is usually more important than general laws of behavior, and that much of what they do is make sense of actions and context so that their behavior is appropriate and helpful. Third, they might appreciate a method that would help them to identify competing realities and values as much as they appreciate methods that enable them to sample opinion. Just as businesses must discover and mold the tastes of consumers who are targets for goods to be sold, public administrators require knowledge of the views, attitudes, and opinions of stakeholders and the attentive public. Finally, professionals may find Q method attractive because of its relative efficiency: it is usually faster and easier to create and conduct a Q method study than undertake random sample survey research.

When public administrators discover Q methodology, they are apt to identify many uses to which it can be put. Following are some possible uses of Q method in public administration practice.

Identifying groups with conflicting values, preferences, and opinions to understand the context of actions. For example, Q methodology might be used in budgeting to determine if

groups of employees in an organization have similar or different perspectives on budget expenditure or taxing priorities (see Cooper and Dantico, 1990). Relatedly, Q method can be used in strategic planning to explore different understandings of the organization's past, its present problems, and its paths to the future (see Gargan and Brown, 1993; and Section III of this chapter).

This understanding of different viewpoints is important because, as McCool (1989) wrote, ''Public administration is shaped by its context. This is true not only for the discipline as a whole, but for the individuals who study, teach, and practice public administration.'' This context includes the interpretation of history by the relevant actors, the construction of present realities, an identification of the nature of the problems or issues being addressed, and the meaning of symbols and myths affecting an organization.[4]

Q method can contribute significantly to understanding and clarifying the perspectives that decision makers, stakeholders, and the public bring to policy issues. For example, if an analyst were studying the issue of whether Georgia's flag should be changed to eliminate the Confederate stars and bars, a Q method study could be used to understand in depth the views of groups opposing and supporting the change. Q methodology has been used in this way to understand more fully the perspectives of stakeholders toward land use planning (Coke and Brown, 1976), city-county consolidation (Durning and Edwards, 1992) and animal experimentation (Brown, 1993).

Also, Q method could be used by an analyst to help structure the criteria to be used to evaluate competing alternatives by insuring that the values to be used in the decision are clarified (see Brown, 1994). For example, the concept of fairness has many competing definitions, and Q method could be used to identify the definition that stakeholders think should be used in evaluating alternatives.

Assisting with the democratization of management and policymaking. Q method can be used in research to understand better and more fully how different coalitions in an organization or a policy arena perceive an issue. In this way, Q methodology can insure that the perspectives of a broad range of stakeholders are known to managers and decision-makers (Coke and Brown, 1976). With Q studies, the voices of different people can be more fully articulated with their differences and similarities noted. Also with the use of Q methodology, it is possible to let clients speak for themselves in regard to management and policy issues (Radin and Cooper, 1989).

Assisting with key public administration tasks and research. As discussed in more detail in section IV of this chapter, Q methodology has the potential for providing useful and important insights into key public administration topics, including elements of a job that most strongly affect morale, work motivation, and job performance; factors affecting employee trust; different concepts of bureaucratic responsibility; and different perceptions of managerial roles.

Use for evaluations. Q methodology can be used to help identify different understandings about the value and efficacy of policies that have been implemented. With Q method, evaluators can systematically collect views of the way that a policy is working or not working (for example, see Kelly and Maynard-Moody, 1993).

B. Q Methodology in Non-Positivist and Post-Positivist Public Administration

Arguments against the positivist foundations of traditional public administration have been made for decades, and they have influenced many researchers in public administration, especially those working in public policy and evaluation (see Kelly and Maynard-Moody, 1993; Torgerson, 1986; Jennings, 1987). Despite growing doubts about positivist public administration methods, their use is still dominant among researchers and practitioners.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Some non-positivists and post-positivists make strong arguments that positivist methods not only lead to mistaken information, but also that they contribute to inequality in society (Dryzek, 1990; Fischer, 1990). Some view the method as reinforcing liberal democracy, which integrates citizens into politics only through group influence. In this type of democracy, citizens have only votes and influence through lobbying groups. These critics are proposing alternatives to liberal democracy, typically discursive democracy, in which disputes over political issues are resolved in favor of better arguments rather than more power.

The post-positivist perspectives of public administration have been around for a while, dating back at least to the New Public Administration of the late 1960s and early 1970s (Kirkhart, 1971), and non-positivist perspectives predate the New Public Administration.[5] The most popular non-positivist and post-positivist perspectives in public administration and public policy have been hermeneutical (Dryzek, 1982; Mesaros and Balfour, 1993; Balfour and Mesaros, 1994), phenomenological (Forester, 1990; McGee, 1971; Hummel, 1987), interpretative (Jennings, 1987; Torgerson, 1986), action (Harmon, 1989), critical (Denhardt, 1981; Forester, 1993), and post-modern theories (Guba, 1985).

These approaches have in common a rejection of the basic ideas of positivism, summarized by Guba (1985, p. 12) as:

> the positivist approach asserts a realist ontology. It rests on the assumptions (and please note that they are assumptions, neither provable or disprovable) that there exists an objective reality ''out there'' going on about its business irrespective of the interest that anyone may display in it; that reality is regulated by certain natural laws (the generalizations of which positivists are so fond); and that many of those laws take the form of cause-effect linkages. Furthermore, positivism asserts an objective epistemology on the assumption that the inquirer (would-be knower) can maintain an objective (non-interactive) posture in relation to the knowable.

In the place of positivism, the non-positivist and post-positivist frameworks emphasize subjectivity. As Guba (1985, p. 13) summarizes:

> [Post-positivism] asserts a relativist ontology on the assumption that all reality is mentally constructed and that there are as many realities as there are persons to contemplate them; that there are no general or universal laws that can be counted on in every situation but that the action or behavior noted in any context is uniquely determined therein; and that all elements of a context are continuously involved in ''mutual simultaneous shaping'' in ways that render the concept of cause-effect meaningless.[6] Further, the emergent paradigm assumes a subjective . . . epistemology, so that inquirer and respondents mutually shape their constructions in a hermeneutic circle throughout the inquiry and thus create the ''reality'' which the inquiry may finally mirror.

While R methods are tools of positivist inquiry, Q methodology is more attuned to non-positivist and post-positivist inquiry. Dryzek (1990) explained why:

> The hallmark of Q methodology is that it takes the subjective, self-referential opinions of respondents seriously in seeking to model the whole subject as he or she apprehends a particular situation (p. 176).
>
> Q is essentially interpretive in its philosophy of social science. As such, it abjures both objectivism and causal explanation (thus departing substantially from opinion research). Instead, Q seeks a ''feeling for the organism'' (Brown, 1989). It engages in intensive analysis of particular individuals or collectivities in order to apprehend the fullness of their subjectivity in the subjects' own terms. . . . [i]t does not (and cannot) seek causal explanation of individual actions. That is, Q interprets the actions of individuals in terms of their consistency (or otherwise) with the subjective orientations it uncovers (pp. 178–179).

> The encounter Q contrives is a thoroughly egalitarian one and the roles of observer and respondent are interchangeable (p. 184).

Not all Q methodologist agree with all of the details of Dryzek's statements above, but they do agree that Q method overcomes many of the shortcomings of positivistic research by providing a ''technique for the objective study of human subjectivity'' (McKeown, 1990). And the unifying element of the non-positivist and post-positivist perspectives is the focus on subjectivity and understanding. Thus, Q methodology, as a tool to investigate operant subjectivity, is well suited for empirical work of non-positivists and post-positivists.

III. UNDERSTANDING AND USING Q METHODOLOGY: A DETAILED EXPLANATION

In this section of the chapter, we discuss in more detail the history and some of the intellectual foundations of Q methodology, then we present a detailed case study of the use of Q methodology. This case study is intended to be a methodological guide for researchers and practitioners who would like to conduct their own Q method research. The section concludes with answers to ''frequently asked questions'' about Q methodology, which we hope will address the issues and concerns of potential users of Q methodology.

A. History and Intellectual Foundations of Q Methodology

1. Brief History

William Stephenson, the inventor of Q methodology, was the last graduate assistant to Charles Spearman, who in turn is best remembered as the inventor of factor analysis. Spearman's main interest, however, was in unlocking the creative potential of the mind, and factor analysis was merely his way of mathematically modeling the processes of thinking in which he had interest. Spearman once referred to Stephenson as the most creative statistician in psychology, but like his mentor, Stephenson was likewise interested in the mind's potential, and the mathematics of his Q methodology are to a large extent secondary to that interest.

Stephenson's innovation can be traced to an August 24, 1935 letter to the editor of the British science journal *Nature* (Brown, 1980: pp. 9–10) in which he drew attention to the possibility of ''inverting'' conventional factor analysis. In R factor analysis, as the conventional approach is often called, traits are correlated across a sample of persons, where ''trait'' is taken to mean any quantitatively measurable characteristic: the factor analysis of the trait correlations points to families of similarly covarying traits. In Q factor analysis, by way of contrast, persons are correlated across a sample of statements which they have rank-ordered, the ranking being called a Q sort: the correlations reflect the degree of similarity in the way the statements have been sorted, and factor analysis of the person correlations points to families of like-minded individuals. A detailed illustration of what is technically involved is presented later in this part of the chapter.

Stephenson's innovation was misunderstood practically from the start and by such eminent University of London colleagues as Sir Cyril Burt, R.B. Cattell, and Hans Eysenck, and at the University of Chicago by L.L. Thurstone, so that even today his name is often associated with a statistical development which was not only not his, but was one which he strongly opposed. Virtually without exception, texts which address Q and R factor analysis regard the two as simply the transposed equivalents of one another—that R consists of correlating and factoring

the columns of a data matrix, and that Q consists of correlating and factoring the rows of the same matrix. In fact, Stephenson's most sustained articulation of Q methodology, his book *The Study of Behavior* (Stephenson, 1953: p. 15), is typically cited in support of this position despite his clear assertion, repeated often, that ''there never was a single matrix of scores to which *both* R and Q apply.'' In this connection, Miller and Friesen (1984, pp. 47–48) are far from alone when, in their factor-analytic studies of organizations, they confidently assert that ''Q-technique is merely R-technique using a transposed raw-data matrix. It treats similarities between companies, rather than between variables. . . . Discussion of Q-technique in particular can be found in Stephenson. . . .''

2. The Quantum Connection

Miller and Friesen's (1984) book is entitled *Organizations: A Quantum View*, which is fortuitous inasmuch as factor analysis and quantum mechanics (as elaborated by Werner Heisenberg and Max Born in particular) are virtually identical mathematically, both relying on matrix algebra (see Peat, 1990) and sharing much of the same nomenclature. Originally trained as a physicist, Stephenson was aware of this parallel in the 1930s, as was Cyril Burt (1940), who wrote extensively about it in his *The Factors of the Mind*, which is as fundamental for R methodology as Stephenson's book is for Q.

But Burt and Stephenson (1939) parted company over what it was that was to be measured. Burt was locked into the study of variables, such as intelligence, assertiveness, temperament, and the thousands of others with which social and psychological science is nowadays familiar. Variables have known properties, but they are largely categorical and owe much to logic and to ''operational definitions,'' hence their quantum character is nothing more than an analogy: Quantum theory is not about variables as such, but about *states* (of energy). Although they may enter into dynamic relations with other variables, the variables of R methodology are not in themselves dynamic and are typically expressed in a single number, usually an average score across a number of responses.

The matter is quite different in Q methodology. Suppose that a person ranks a set of statements (say, from agree to disagree) to represent his or her own point of view about the organization. The statements do not measure anything a priori, i.e., their meaning is indeterminate; they are simply assertions that have been made about the organization (e.g., that ''it is a pleasant place to work''). Meaning and significance are imposed on the statements by the person in the course of Q sorting them, hence the inseparability of measurement and meaning. The process reflects a dynamic *state* (of mind) in a relationship of *complementarity* to other states (i.e., to Q sorts by others in the organization). The final product (the completed Q sort) is not an average, nor is it subject to any external norms; rather, it is a dynamic pattern of interrelationships. The self referentiality of the Q sorter is obviously central to the situation, and the way in which the statements will be understood and scored is solely in the hands of the Q sorter; it can never be known in advance, therefore, how many factors will emerge, nor what their form and content will be. Everything is indeterminant, and the parallel with quantum theory is made the more remarkable by virtue of the fact that it is a function not of analogy, but of the ''sovereignty of measurement'' (Stephenson, 1989).

In sum, there is considerably more to the difference between R method and Q method than a simple decision whether to analyze the relationships between the columns of a data matrix or the rows of the same data matrix. A much more fundamental difference is between the study of the variables of R methodology, conceived as objective and subject to classical conceptions of cause and effect; and the study of a data matrix of a wholly different kind, one thoroughly saturated with self referentiality and probabilism.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

3. Concourse Theory

As was noted previously, Stephenson, like his mentor Charles Spearman, was interested in the creative potential of the mind. What is the source of creativity and how do we liberate it for the social good—more specifically, for administrative advantage?

In an instructive book on word origins, C.S. Lewis (1960) devotes a chapter to ''conscience and conscious,'' both of which derive from a common Latin antecedent in which the prefix *con* means *with*—hence *conscio* means sharing what one knows with someone, which includes sharing with oneself as in musings, daydreams, and mulling things over.

But there was also a weaker sense in which *conscientia* connoted simply awareness, as in being conscious *of* something, epitomized in Descartes' awareness of his own thinking (*cogito ergo sum*) and in introspectionism. Needless to say, this weaker sense of *conscientia*, upon which modern cognitive psychology is based, has virtually replaced the stronger sense of knowledge as shared, thereby elevating the individual thinker while removing almost all traces of the social context within which thinking takes place.

Yet most of ordinary life is based on shared knowledge, and it was on this account that Lewis (Stephenson, 1980) introduced the term *consciring*. Administrators and policymakers spend much of their time exchanging views in committee meetings or around the drinking fountain, and in reading and responding to one another's memos and reports. Under more managed conditions, ideas are shared via Delphi or nominal group techniques in which each person's idea-sharing may release new thoughts in others, the total being greater than could be produced by the same participants working in isolation. The opposite is equally true, that a single individual can often produce ideas which elude groups due to the isolate's relative freedom from conformity and the sometimes suffocating effects of collegiality.

Administration in general has been characterized as ''the activities of groups cooperating to accomplish common goals'' (Simon et al., 1950, p. 3), and as a group assembles and begins to cooperate, a vast array of possibilities begins to appear, typically in linguistic form. We may assume that the impetus for an assemblage is the existence of a task or problem—e.g., how to pare the budget, what to do about a councilman's complaints concerning the streets in his ward, when to mount a campaign for the school levy, determining who would make the best executive director, etc.—and that the fruits of the group's cooperative efforts take the form of proposed solutions. Hence, using the budget as illustrative, we might hear proposals such as ''The easiest thing to do is to slash 3.2% across the board,'' or ''We can't really cut back further on social services,'' or ''It's been a mild winter so there should be some fat in the street maintenance line,'' which may prompt the service director's warning that ''We haven't even completed all the repairs from the freeze of '93,'' and so forth.

All such communicability is inherently contestable, infinite in principle, ubiquitous in character, and inescapably subjective. In quantum theoretical terms, it is also unpredictable, paradoxical, and erratic: No one knows in advance what someone else is going to say or suggest, or how what one person says is going to impact on what others say or think. In Q methodology, such communicability is referred to as a *concourse* (Stephenson, 1980, 1978, 1986), a concept traceable to Cicero, but which takes its most modern form in Peirce's (1955) ''Law of Mind''—that ideas spread and affect other ideas and eventually combine into a system, or schema. Concourse is therefore at the foundation of a society and provides lubrication for all its parts, and it constitutes the very stuff of which decisions are made and problems solved. And it is concourse that supplies the elements of Q methodology.

Concourse is present in the loftiest of philosophical discourse to the simplest coos and gurgles of the nursery, as almost any random examples will easily show. In a recent posting

on an Internet list devoted to quantum theory and consciousness, for instance, the following assertions were made:

> The universe is simple at fundamental levels. . . . A unified approach requires complementary modes of description. . . . Non-locality is firmly established. . . . Self-organization and complexity are prevalent at all scales in the universe. . . . All human experience is connected to the universe.

Of such things volumes have been written, many by Nobel laureates and others of repute, each idea midwifing yet other ideas in endless profusion. And although supported in part by facts, communicability of this kind is thoroughly subjective and typically goes beyond known facts, with the facts as such colloidally suspended in the subjectivity. Such is the character of concourse.

Or consider Sasson's (1995) study of the way in which citizens construct crime stories, based on the shared communicability found on op-ed pages of newspapers; or Finkel's (1995) study of jurors' common sense understandings of justice. Roe's (1994) volume is full of narratives on such diverse social problems as budgets, global warming, animal experimentation, and so forth; as is Ellis and Flaherty's (1992) volume on topics such as feminism, film, dance, and other features of social life. Apart from their narrative richness, the common limitation of efforts such as these is methodological in character—i.e., once Sasson and the others have gathered their mountains of discourse material, they typically end up sorting it into various more-or-less logical categories. What began on a sound footing of naturalism, therefore, ends up being sectioned according to the categories of the analyst.

Perhaps of more direct pertinence to decision-making settings of the kind more familiar to administrators is the concourse which emerged from a 1987 meeting of a commission established to assist in encouraging Central American development (Brown, 1988). The organizing question concerned the commission's role, and the following observations were among those which were added to the concourse as each commissioner stepped up to the microphone:

> The Commission must attend not only to economic growth, but to social and cultural growth. . . . We must begin with those elements which bring us together, and only later address those issues which divide us. . . . It is incumbent upon the Commission to recognize the central role of agriculture in the development of the region. . . . It must be adopted as an operating principle that the problems to be considered are autonomous to the region, and not reflections of East-West security issues. . . . The process which the Commission works out must, through its structure, recommend that aid be conditioned on measurable and swift progress toward genuine democracy.

The commission spent the better part of a day and a half in this vein, which continued during meals, breaks, over drinks, and in walks on the hotel grounds—perhaps even into some members' dreams. Such is the nature of ideational spreading and affectability characteristic of Peirce's ''Law of Mind.''

Concourse proliferates not only around problems, but in terms of the interests of problem-solvers: Ask a manager and a worker how best to improve the company and different suggestions will be offered. This can be seen concretely in the comments obtained from a group of sixth grade students when asked what might be done to improve their school:

> Have more plays or assemblies. . . . Show more movies. . . . Make the halls more colorful and interesting by decorating them with students' work. . . . Have the PTA hold more activities like the carnival. . . . Plant more flowers, bushes, and trees around the building.

Aesthetics, desire for pleasureful activities, and something of a communal spirit predominate, as does the kind of dependency upon adults still characteristic of students at this age. When the same question is asked of a group of young eleventh grade policymakers, however, some of the same issues remain at the surface, but private interest and a desire for autonomy are now more prominent:

> Increase the variety and amount of food for lunch. . . . Do away with assigned seats in classes. . . . Don't assign homework on Friday or the day before a vacation. . . . Add vending machines to the cafeteria (such as candy and pop machines), and put a juke box in the cafeteria. . . . Add a course for fourth year French and Spanish. (Ad infinitum.)

And in equal volume were the solutions proffered by a group with different interests—graduate students and faculty who were queried as to how best to improve their graduate program:

> Establish a rule or procedure whereby faculty are required to specify clearly and precisely the criteria for grading. . . . Structure course offerings so that students can choose between quantitative and non-quantitative approaches. . . . Increase the minimum grade-point requirement for newly admitted graduates. . . . Place more pressure on the faculty to do research and to publish.

Assembling a concourse relative to a particular decisional situation is part of what Simon (1960) designated as the *design* phase of the process, in which alternative courses of action are collected. Mere archiving is not the goal, of course, but a new starting point from which alternatives are then appraised, developed, and eventually adopted or discarded, and it is at this point in the process that some of the qualitative methods (e.g., narrative, discourse, and ethnographic analysis) often falter and sometimes never fully regain balance. Given a welter of textual material, the qualitative analyst must find some method of categorization so as to bring order to the enterprise, and this almost inevitably means the superimposition onto verbal protocols of a logical scheme of one kind or another. The conscientious analyst will of course exercise as much care as possible in an effort to assure that the categories used are also functional and not logical only, but there is no cure for the nagging doubt that the categories belong to a greater or lesser extent to the observer rather than the observed.

Q methodology alleviates these doubts to a considerable extent by revealing the participants' own categories and establishing these as the basis for meaning and understanding. How this is accomplished is best grasped in the context of a more extended example, which will also serve to illustrate the quantitative procedures which are involved.

B. Strategic Planning with Q Methodology: A Case Study

The problem in this illustration has been briefly reported elsewhere (Gargan and Brown, 1993) and involved the formulation of a strategic plan for a Private Industry Council (PIC), a local nonprofit agency primarily responsible for implementation of the Federal Job Training Partnership Act. Specifically, the agency's task was to enhance employment opportunities for the ''hard to serve'' by providing training and skill development, and in light of federal cutbacks in resources the PIC Board of Trustees sought assistance in priority-setting.

The process was begun by inviting the dozen or so assembled Board members silently to contemplate ''what issues and problems should be given priority during the next two to four years if the employment needs of the hard to serve are to be effectively dealt with?'' Each person jotted down freely-associated ideas until it became apparent that few if any new ideas would be forthcoming, at which point the facilitator guided the group through a round-robin

process in which each person in turn nominated one of the solutions on his or her list. Each nominated solution was discussed, modified through group discussion, and finally added to a list preserved on a blackboard for all to see. Group members then copied each item on a 3 × 5 card which had been provided, using the same wording as on the board. The items were numbered serially, and the item numbers were also recorded on each of the cards. Eventually, the Board members collectively generated N = 33 ''issues and problems,'' and each member was in possession of a pack of 33 cards on which those problems were written.

Before proceeding to technicalities, it is important to note that Board appointments had been purposely made so as to represent the business, labor, and political segments of the surrounding communities, and that the Board members were knowledgeable about the agency's role and about the dwindling resources available for fulfilling it. Hence the 33 propositions generated were of wide scope, and there was not a single one among the 33 that all Board members did not immediately understand as a matter of shared knowledge. All of the items are reported in a later Table, but a small sampling will give a sense of the issues confronting the agency:

1. Need to encourage new industry to respond to the manufacture of the reuse of recyclable materials.
4. How to effectively use the newly created Economic Development Office to create new job opportunities in the county.
8. Get back to helping displaced workers—even if other policies are pushed on us.
13. How to recruit new participants from the private sector on the PIC Board of Trustees and create more involvement by the community.

And so forth. Item (1) was in response to countywide recycling initiatives, (4) reflected a need to network with other agencies, (8) was proposed by a Board member sensitive to his labor union constituency, just as (13) was nominated by a representative of the business community. Tributaries into a concourse of communicability emanate from values, political commitments, and other social forces such as these, and it is a virtue of Q methodology that it sharpens and clarifies the form and substance of such forces.

It should be noted in passing that the wording of some of the items may sound odd, but group members were given ample time in which to suggest editorial amendments and to clarify meaning, and many items underwent alteration before the final version was collectively approved. However unusual or ambiguous the phrasings might appear to an outsider, therefore, there is little reason to doubt that the insiders themselves understood each and every statement.

The purpose of the initial phase of item generation was, in this case, to gather as comprehensive a set of agency problems as possible, i.e., to render manifest the topics on the group's agenda. The next phase involved distinguishing important from relatively less important problems, and this was accomplished by instructing each participant to Q sort the 33 items along a scale from +4 (important) to −4 (unimportant), as shown in Table 1 for one of the Board members. It was first recommended to participants that they divide the items into three groups—of high, medium, and least importance—and that from the highly-important items they then select those two deemed of greatest importance: these were placed under the +4 label, with the three next-most important being placed under the +3 label. (The labels were arrayed across the tabletop in front of each participant, to serve as an aid in the sorting.) The participants then examined the stack of relatively unimportant problems and selected the two most unimportant of all (for −4), and then the three next-most unimportant (−3). Eventually all 33 items were arrayed in front of the participant, from those most important on the right down to those most unimportant on the left. The statement numbers were then entered on score sheets so as to preserve the way in which each of the participants prioritized the problems.

TABLE 1 Q Sort for Board Member No. 1

Unimportant								Important
−4	−3	−2	−1	0	+1	+2	+3	+4
2	12	3	13	7	15	14	1	4
5	21	18	19	8	22	16	10	6
	26	27	20	9	24	17	11	
		32	25	23	28	29		
				30				
				31				
				33				

It deserves mention in passing that the so-called forced distribution pictured in Table 1, although somewhat arbitrary in shape and range, is nevertheless recommended for theoretical and practical reasons. Theoretically, a quasi-normal distribution models the Law of Error and is backed by a hundred years of psychometric research indicating that Q sorting and other quasi ranking procedures typically result in distributions of this kind (Brown, 1985). From a practical standpoint, a standard symmetrical distribution of this kind constrains responses and forces participants to make decisions they might otherwise conceal (e.g., by placing all statements under +4 and −4), thereby increasing the likelihood that implicit values and priorities will be rendered explicit and open to view. However, since the shape of the sorting distribution has little impact on the subsequent correlation and factor analysis, recalcitrant respondents (who might otherwise refuse to cooperate) can be permitted to follow their own inclinations while being encouraged to adhere as closely as possible to the distribution specified.

Given 33 possible problems to consider, there are literally billions of different ways—in fact, more than a billion billion—in which they could be prioritized, and yet the participants individually work through the complexities involved within just a few minutes, which is testimony to the adaptability and efficiency of the human mind. Each person is guided in this enterprise by the values, interests, and principles which are brought to the task, and what these values and interests are can usually be inferred through inspection of the person's Q sort.

Consider, for instance, the Q sort provided by Board member no. 1 (Table 1), who singled out the following problems as most important (+4):

4. How to effectively use the newly created Economic Development Office to create new job opportunities in the county.
6. How to coordinate PIC, OBES, and Department of Human Services. All have performance standards to meet. Better understanding of those performance standards for the good of the County.

In order to maximize candidness, participants were not required to place names on their score sheets, and so we are not in a position to associate this specific response with a particular individual. This is unimportant in a study such as this, however, for what is of interest are the subjective views themselves—the perspectives that exist within the group—and not the associated characteristics of those who espouse them. In this connection, Board member no. 1's choice of important issues indicates an awareness of the PIC agency in the context of other county agencies with which this agency's activities could be coordinated.

This person's most unimportant issues (−4) are also illuminating:

2. With training programs now, not just job specific but overcoming barriers of employment.

5. Work ethic and attitudes need to be developed to make individuals more employable; need to increase time to do that.

This Q sorter appears relatively disinterested in (if not downright antagonistic to) removing employment barriers or in helping to improve the skills of the unemployed, and this, coupled with the positive attitude toward other agency elites (shown under +4), suggests a certain pro-institutional attitude on the part of Board member no. 1.

The Q sort is part of the technical accouterment of Q methodology, to which it bears a relationship analogous to the telescope to astronomy or the galvanometer to electricity, i.e., it brings into view for direct and sustained inspection those currents of subjectivity and preference which suffuse political and administrative life. The decisional situation facing the PIC agency contains certain objective realities—e.g., the existence of recyclable materials, barriers to employment, and the existence of other agencies (such as OBES, Human Services, and Economic Development), as referred to in the statements above—but what is felt to constitute an important as distinguished from an unimportant problem is a phenomenon of another kind: it is the subjective medium within which the facts as known by each Board member are suspended, and in terms of which they are given meaning and salience, as rendered manifest by the +4/−4 subjective scale. The concourse of ''issues and problems'' which the PIC Board produced is fed from such subjective currents, and yet these sources remain obscure until transformed by the mechanics of Q sorting.

Among the features of Q methodology which distinguish it from many other quantitative procedures is that the elements comprising the Q sort, unlike those in a rating scale, are not constrained by prior meaning; consequently, there can be no correct way to do a Q sort in the same sense as there is a right way to respond to an IQ test or to the Graduate Record Examination. What a particular PIC Board member considers the most important problems facing the agency is simply that person's judgment, which may or may not be in agreement with others' appraisals. It is this absence of norms—not just in practice, but in principle—that renders the issue of validity of such little concern in Q methodology (Brown, 1992–1993). Validity aside, however, we can nevertheless proceed to compare subjective appraisals, and to determine the varieties of appraisal in existence and the shape and character of each, and it is at this point that quantification is brought to bear.

Q sorts are conventionally intercorrelated and factor analyzed, and the correlation phase is fairly elementary, as shown in Table 2 for the Q sorts provided by PIC Board members 1 and 2. The scores in column 1 are the same as those shown in the previous table (for the 33 issues and problems which were sorted), whereas those in column 2 were given during the same sitting by Board member no. 2. The numbers in column d^2 are the squared differences in score between the two performances. Were the two Board members' views absolutely identical, there would of course be no differences in score between the two for any of the statements, in which case the squared differences would also be zero, as would the sum of squared differences: In this extremely rare instance, the correlation would be $r = 1.00$; were the two views diametrically opposite, the correlation would be $r = -1.00$.

In the instant case, there are differences of a greater or lesser extent among the statements, as recorded in column d^2, the sum amounting to 234. When Q sorts follow the same distribution (hence have the same mean and variance), a convenient formula for correlation is as shown in Table 2, where 234 is the sum of squared differences and 316 is the sum of squares of the scores in the two Q sorts. The correlation between these two Q sorts is therefore $r = .26$.

Q-sort correlations are rarely of any interest in and of themselves and typically represent only a phase through which the data pass on the way to being factor analyzed. It is worth noting in passing, however, that the correlation coefficients are subject to standard error formulae. For

TABLE 2 Correlation Between Q Sorts

Item	1	2	d^2	Item	1	2	d^2
1	3	−2	25	21	−3	2	25
2	−4	−1	9	22	1	−2	9
3	−2	−3	1	23	0	3	9
4	4	0	16	24	1	−2	9
5	−4	0	16	25	−1	2	9
6	4	4	0	26	−3	−4	1
7	0	1	1	27	−2	−1	1
8	0	0	0	28	1	3	4
9	0	0	0	29	2	2	0
10	3	−1	16	30	0	−1	1
11	3	2	1	31	0	−4	16
12	−3	−3	0	32	−2	1	9
13	−1	0	1	33	0	1	1
14	2	−2	16				
15	1	4	9				
16	2	1	1	Sum			234
17	2	0	4				
18	−2	0	4				
19	−1	3	16	$r = 1 - (234/316) = .26$			
20	−1	−3	4				

example, we can assume pro tem that two Board members' views are substantially related if they exceed $2.58(1/\sqrt{(N)}) = .45$ (for N = 33 Q statements), where $SE = 1/\sqrt{(N)}$ is the standard error of a zero-order coefficient, and z = 2.58 is the number of standard errors required to incorporate 99% of the area under the normal curve. The above correlation of .26 for Board members 1 and 2 (which is less than the requisite .45) therefore indicates that their respective appraisals of problems facing the agency share little in common.

Of the Board members originally involved in this strategic planning process, only seven have been included in the following analysis so as to keep calculations and tabular displays within manageable limits for illustrative purposes; the intercorrelations among the seven participants are displayed in Table 3. Note that the correlation between Board members 1 and 2 is r = .26, as calculated above.

The correlation matrix has a certain dynamic to it, just as did the group from which the Q sorts were taken: hence the old-line labor unionist may have had his eye out for ways to re-

TABLE 3 Correlations Among Seven Q Sorts

	1	2	3	4	5	6	7
1	—	26	16	21	−07	13	26
2	26	—	15	−20	61	13	13
3	16	15	—	−02	18	43	11
4	21	−20	−02	—	−27	23	22
5	−07	61	18	−27	—	13	22
6	13	13	43	23	13	—	−08
7	26	13	11	22	22	−08	—

Note: Decimals to two places omitted.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

employ the formerly-employed (his constituency), and might therefore have been less sympathetic to other Board members' expressed concerns about the ''hard to serve,'' which could easily have come to be a symbol in the group for mentally-challenged, unskilled, and other rivals for scarce positions at the bottom of the job chain. Which issues each person places at the top, middle, or bottom of the Q sort can therefore be affected by a myriad of forces, from personal motivation to the climate created by the condition of the national economy, each force being explicitly or sometimes only implicitly weighed by the Q sorter and ultimately having its impact on the final statement arrangement. The correlation matrix is therefore a thick soup of dynamic forces of chaotic proportions which nevertheless summarizes the balance of influences and the way in which the various participants have worked their way to separate conclusions about what are important vs. unimportant issues facing the agency.

Q methodology was conceived in the context of factor-analytic developments as they were unfolding in the 1930s, which was late in the hey-day of Charles Spearman and the ''London School.'' Stephenson's (1935) psychometric innovation of ''correlating persons instead of tests'' consisted of applying the mathematics of factor analysis to correlation matrices of the above kind, in which person-responses were correlated with other person-responses. The result was typically a typology of response, with one subgroup of similar Q sorts constituting one factor, another group constituting another factor, and so forth. Q factors therefore have the status of separate attitudes, perspectives, or understandings, or, in the extant case, of issue priorities.

By whatever substantive terminology (attitudes, perspectives, value orientations, issue priorities, etc.), the factors in Q methodology consist of conglomerates of convergent subjectivity as determined by the concrete operations of the persons themselves as they perform the Q sorts—hence the term *operant subjectivity* (Stephenson, 1977). The number and content of the factors, despite their thoroughly subjective character, are therefore emergent and purely empirical features of the thinking and feeling of the persons who provided the Q sorts: Had the group members felt differently about the issues, their Q sorts would have been different, and so would the structural features of the correlation matrix—and so, as a consequence, would the factors, which in their turn summarize those structural features. The role of factor analysis in this instance is to document the current state of thinking within this particular strategic planning group with respect to the issues at the group's focus of attention.

The technicalities of factor analysis are addressed elsewhere in this volume (De Lancer, 1998), and relatively simplified introductions are available for those wishing to achieve a conceptual grasp (e.g., Adcock, 1954; Brown, 1980: pp. 208–247; Rust and Golombok, 1989); we will therefore bypass the detailed calculations involved in extracting the factor loadings shown in Table 4.

Suffice it to say that from a statistical standpoint, the seven unrotated factors represent a partial decomposition of the previous correlation matrix. This can be illustrated in terms of any two Q sorts—say, nos. 2 and 5, which are correlated in the amount .61 (see Table 3). The sum of the cross-products of the unrotated factor loadings for these two Q sorts is $(.50)(.35) + (-.42)(-.72) + \ldots + (-.17)(-.31) = .58$, which indicates that virtually all of the original correlation of .61 can be composed from these seven factors. The factor loadings indicate the correlation of each Q sort with the factor, hence Q-sort 2 correlates with the first factor in the amount $f = .50$; factor loadings are therefore subject to the same standard error estimates as noted previously for correlation coefficients, i.e., $SE = 1/\sqrt{(33)} = .17$, where $N = 33$ statements. Factor loadings in excess of $2.58(.17) = .45$ are significant ($p < .01$), which means that person 2's Q sort is significantly associated with the first factor.

Had there been only a single viewpoint shared by all members of the PIC Board, then all of the correlations would have been large and positive, only one significant factor would have been in evidence, and there would have been no trace of significant loadings on the other factors.

TABLE 4 Unrotated and Rotated Factor Loadings

	Unrotated factors							Rotated factors		
	a	b	c	d	e	f	g	A	B	C
1	43	08	−26	14	−28	21	13	11	00	(57)
2	50	−42	13	06	−07	01	−17	10	(63)	21
3	47	13	25	01	07	00	01	(48)	18	17
4	06	37	−27	36	−09	02	09	07	−38	32
5	35	−72	36	38	29	15	−31	00	(94)	09
6	44	42	51	28	−16	06	−04	(79)	06	10
7	38	−02	−34	19	29	16	25	−03	00	(61)
								( ) $p < .01$		

As Table 4 shows, however, at least the first three of the unrotated factors contain significant loadings, and some of the loadings on the fourth factor are also substantial; we would therefore anticipate that there are probably three and perhaps four separate points of view within the PIC Board.

Although there are some occasions in which the factor analyst might rest content with the unrotated loadings as these have been extracted from the correlation matrix, in the overwhelming number of cases unrotated loadings do not give the best view of what is transpiring; it is typically the case, therefore, that the unrotated factor loadings are superseded by an alternative set of loadings which give a more focused view. This transformation process—from the unrotated to an alternative set of loadings—is accomplished through the process of factor rotation.

The most conventional scheme for factor rotation is to rely on the Varimax routine found in all software packages (such as SPSS) containing factor analysis, and it is the statistical goal of Varimax to rotate the factors in such a way that each variable (or Q sort) is maximized on a single factor and minimized on all other factors, a solution referred to as ''simple structure.'' If a researcher is totally in the dark about the topic under examination, as is sometimes the case, then leaving factor rotation to Varimax or some other algorithm may be as good a strategy as any other.

However, it is unlikely that there is a single set of mathematical rules, such as Varimax, which is apt to provide the best solution to problems under any and all conditions. In particular, when an investigator has some knowledge or even vague hunches about a situation, then it is often wise to permit that information to play some role in the experimental setting. It is for situations such as these that room was made in Q methodology for ''theoretical rotation,'' a judgmental procedure which is explicitly built in to the QMethod freeware program (Atkinson, 1992), available in both mainframe and PC versions.

Space precludes going into great detail concerning theoretical rotation, but what is essentially at issue can be demonstrated in terms of Figure 1, which graphically displays the location of each Q sort in terms of unrotated factors (a) and (c) in Table 4. The pairs of loadings for each of the seven Q sorts are duplicated in Figure 1 (significant loadings in parentheses) where it is shown that Q sort no. 1 has a loading on factor (a) in the amount .43 and on factor (c) in the amount −.26, and these two loadings serve to locate Q sort 1 in the two-dimensional space in Figure 1. Similarly, Q sort 6 is located .44 on factor (a) and .51 on (c). The relative proximity of each of the Q sorts is a spatial expression of their degree of similarity with respect to these two factors (the nature of which are undefined at this point)—hence, Q sorts 1 and 7 are similar

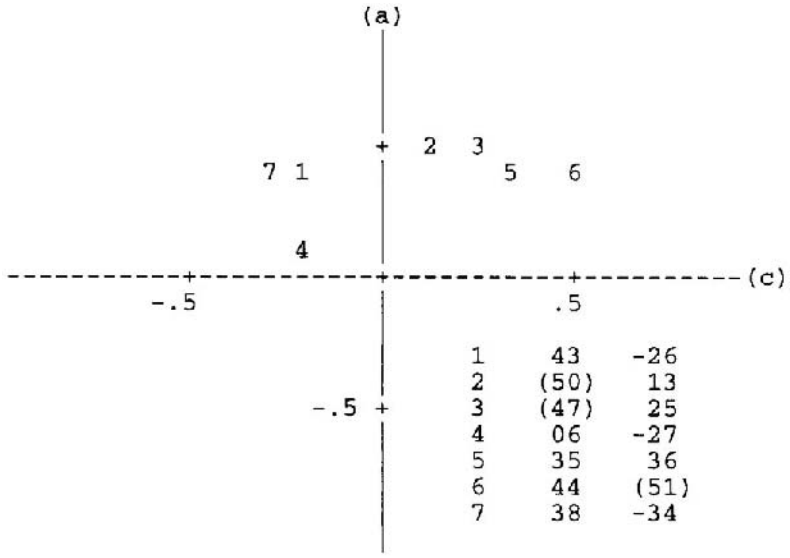


FIGURE 1 Displays each Q sort in terms of unrotated factors (a) and (c). Illustration used to spatially demonstrate the degree of similarity of the Q sorts with respect to factors (a) and (c).

to one another, as are 5 and 6, although these two pairs are different from one another (as reflected in their spatial distance).

Seven unrotated factors were originally extracted, which is the default in the QMethod program. Of these seven, factors (a) and (c) were chosen due to an interest in Q sort 1 (see Table 1). Recall that Q sort 1 displayed an appreciation for the PIC agency's location in the context of related community agencies, such as the Economic Development Office and the Department of Human Services, and it was this inter-agency perspective on PIC's status that attracted interest and subsequently led to a decision to focus one of the factors on Q sort 1. As indicated in Table 4, Q sort 1 has substantial loading (.43) on factor (a) but almost none (.08) on (b); there is a degree of variability (−.26) on factor (c), however, and relocation of reference vectors (a) and (c) could serve to isolate all of Q sort 1's variability on a single factor.

Figure 2 displays the relationships among the Q sorts when the original vectors are rotated counter-clockwise by 40-degrees, to new locations designated a′ and c′. Also shown are the new factor loadings which indicate that Q sort 1's loading on a′ is now .50 (up from .43), and on c′ is .08 (down in magnitude from −.26). These two sets of coefficients, unrotated and rotated, are mathematically equivalent, as can be seen when the respective loadings are squared and summed, as follows:

$$\text{Unrotated: } .43^2 + (-.26)^2 = .2525$$

$$\text{Rotated: } .50^2 + .08^2 = .2564$$

which are identical save for rounding error. Another method of verification is to examine the cross-products of factor loadings for any two Q sorts, e.g., nos. 1 and 2:

$$\text{Unrotated: } (.43)(.50) + (-.26)(.13) = .1812$$

$$\text{Rotated: } (.50)(.30) + (.08)(.42) = .1836$$

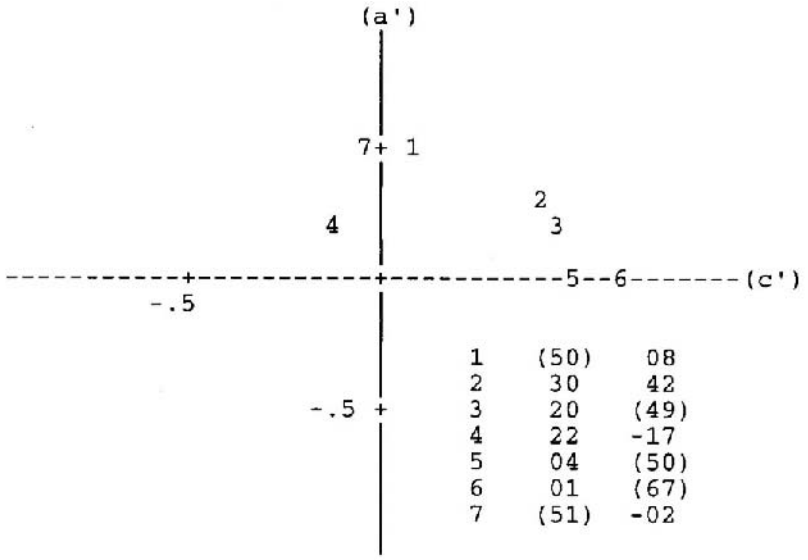


FIGURE 2 Displays each Q sort after rotating the original vector counter-clockwise by 40-degrees.

which, again, are identical with allowances for rounding. The cross-product sums of .18 indicate that of the original correlation of .26 between Q sorts 1 and 2 (see Table 3), factors (a) and (c), or rotated factors a′ and c′, account for .18 of that amount.

It is scarcely necessary for an investigator to comprehend the mathematics underlying the rotational phase of a Q study since the QMethod program (Atkinson, 1992) displays the data configuration pictorially, as in Figure 1: the investigator then simply directs the vectors to a new location based on theoretical considerations, e.g., by specifying that factors (a) and (c) be rotated 40-degrees counter-clockwise so as to locate Q sort no. 1 on a single factor. The selection of any particular Q-sort response as the focus for factor rotation can be due to any considerations that might arise in the course of a study—for instance the person's status (e.g., as PIC Board Chair, or as the only labor representative on the Board, etc.), something a person said in the course of an interview, others' attitudes toward a specific Board member, and so forth. It is at the stage of factor rotation that an investigator's hunches, guesswork, and hypothetical interests enter into the analysis.

The rotation described above was only the first of six rotations that resulted in the final three-factor solution shown in Table 4. Space precludes detailing the thinking that went into these rotations, but the three factors (A, B, and C) indicate that the Board members were, like Gaul, divided three ways in their appraisal of the importance of the problems facing the agency. The figures show that Board members 3 and 6 define factor A, which means that these two participants share a common view about the problems before the agency; that members 2 and 5 likewise share a common view (factor B), but one that is uncorrelated with the factor A view; and that members 1 and 7 define factor C, a third view orthogonal to the other two.

Before moving on to the interpretative phase of Q methodology, it is important to emphasize once again the operant character of these three factors: The Q sample of agency problems was generated solely by the Board members themselves, the Q-technique rankings of the problems emanated from their own subjective perspectives, and the three factors are therefore natural categories of thought and sentiment within the group itself. This is not to assert methodological

individualism, nor is it to deny that the views expressed in the group could have been socially constructed or in other ways a function of forces in the social and economic order. It is simply to acknowledge that the factors are a function of the lived experiences of individuals, and that they are purely inductive in the sense that their number and character have been induced from those individuals who produced them.

As noted above, Board members 3 and 6 share a common outlook, and what is common to it can be approximated by melding the two. This is accomplished by calculating *factor scores*, which are the scores (from +4 to −4) associated with each statement in each of the factors. The two responses first have to be weighted to take into account that Q sort 3 (with a loading of .48) is a lesser approximation to factor A than is Q sort 6 (with a loading of .79). Weighting proceeds according to the formula $w = f/(1 - f^2)$, where f is the factor loading. Q sort 3's weight is therefore $w(3) = .48/(1 - .48^2) = .62$, and by the same calculations $w(6) = 2.10$, hence the latter is magnified more than three times the former ($2.10/.62 = 3.39$) when the two responses are merged. The same process is repeated for factors B and C. Computational details are to be found in Brown (1980: pp. 239–247).

The end products of the above calculations, as shown in Table 5, are three composite Q sorts (one each for factors A, B, and C), with each factor Q sort being composed of the weighted individual Q sorts which define the factor. (All computations are built into the QMethod software package) (Atkinson, 1992). In the case of factor A, for instance (see Table 5), the two highest scores (+4) are associated with statements 4 and 22, the three next-highest (+3) with statements 25, 27, and 33, and so forth in the same distribution as employed in the original Q sorts (see Table 1). What started off as seven Q sorts is therefore reduced to three, which subsume the others, and it is these three Q sorts which provide the basis for factor interpretation.

In most circumstances, factor interpretation proceeds best by (a) physically laying out the factor Q sort and describing its manifest content, giving special attention to those statements at the positive and negative extremes of the distribution, (b) establishing the underlying theme, which is the glue holding the Q sort together, and (c) comparing and contrasting the factor with whatever other factors have also been revealed. Factor interpretation is more hermeneutical art than science, and necessarily so, but the interpreter is not given free rein in this respect since an interpretation is constrained by the factor scores from which it cannot stray too far. Interpretation is further complicated by the fact that the statements in a Q sort can assume various meanings depending on their position relative to other statements, and this places Q methodology in the same league with literary theory in the sense that it has to address ''the ways in which the form that transmits a text . . . constrains [or facilitates] the production of meaning'' (Chartier, 1995: p. 1).

Space again precludes going into detail, but something of what is involved can be seen in reference to a cursory look at factor A which, as Table 5 reveals, gave highly positive and negative scores to the following statements:

> Most Positive (+4, +3): (22) As PIC we are unique—only PIC that is a corporation. Need to consider other sources of funding beyond federal and state governments. . . . (25) Need to develop own Board and staff attitudes towards those we serve—to overcome stereotypes of those we serve. . . . (27) How do we best restructure PIC given potential changes in committee structure of board, management staff, recruit new personnel, etc. . . . (33) PIC has reacted to mandates of state and federal governments. How do we shape and affect state and federal policies rather than simply reacting to them.
>
> Most Negative (−4, −3): (7) Effectively test for aptitude and skills—specifically train where they have the best potential. . . . (8) Get back to helping displaced workers—even if other policies are pushed on us.

TABLE 5 Factor Score Arrays

Factors			Q-sample statements
A	B	C	
0	−4	0	1. Need to encourage new industry to respond to the manufacture of the reuse of recyclable materials.
2	3	−4	2. With training programs now, not just job specific but overcoming barriers of employment.
−2	−3	−2	3. If balancing the federal budget is indeed a top priority as presidential candidates verbalize, will PIC or many other governmental agencies even be in existence?
4	1	4	4. How to effectively use the newly created Economic Development Office to create new job opportunities in the county.
2	2	−4	5. Work ethic and attitudes need to be developed to make individuals more employable—need to increase time to do that.
1	0	2	6. How to coordinate in every way the PIC, OBES, Department of Human Services. All have performance standards to meet. Better understanding of those performance standards for good of the county.
−3	3	1	7. Effectively test for aptitude and skills—specifically train where they have the best potential.
−4	−1	2	8. Get back to helping displaced workers—even if other policies are pushed on us.
−1	−1	3	9. Increase EDWA funding and also the possibility of apprenticeships.
−1	−4	1	10. Encourage cooperation with the University and Liquid Crystal Institute research and private entrepreneurs to establish small industries.
−1	0	0	11. Develop a mentoring program with the employer to serve as a support system for the newly employed.
−2	−1	−2	12. What societal changes will be different in the next few years and how will they be addressed?—e.g., decrease in manufacturing jobs and increase in information-related jobs: How will PIC respond?
−1	0	1	13. How to recruit new participants from the private sector on the PIC Board of Trustees and create more involvement by the community.
2	1	3	14. Cultivating more job sites.
0	4	1	15. Be prepared for the idea of limited resources. Prepare: (a) Plan A, with no improvement in funding; (b) Plan B, worsening of the economy; (c) Plan C, serious cuts in federal and state resources.
−2	3	2	16. Survey of all area businesses to determine need for the low-skilled and the non-skilled.
0	0	3	17. More aggressive in economic development—e.g., get with building trades/realtors to put together packages of what we could do for industries.
0	−2	0	18. Much process in place, needs additional support from private sector in addition to what it is getting from the Commissioners.
1	2	0	19. Some thought to prioritizing services—doing a lot of things, sort of driven by funding. Need to be sure we are taking right route to objectives. Services provided as well as outreach to business.
0	−3	−3	20. Funding available for infrastructure replacement and repair—roads, bridges, public buildings—much in order of WPA.
−2	2	−2	21. Development of some type of survey on five levels: (1) what general public perceives PIC to be, (2) perception by businesses which are using PIC, (3) perception by businesses not using PIC, i.e., potential businesses, (4) perception by clients currently using PIC, (5) perception by potential clients currently not using PIC.
4	0	2	22. As PIC we are unique—only PIC that is a corporation. Need to consider other sources of funding beyond federal and state governments.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

TABLE 5 Continued

Factors			Q-sample statements
A	B	C	
2	2	0	23. Marketing issue—every eligible person in the county should know about PIC services—adults and youths. Every business in the county should be a member of PIC and should know of benefits to business.
0	−2	0	24. Plan not only to make jobs available but to search out companies that will invest in the county—labor intensive companies.
3	−2	−3	25. Need to develop own Board and staff attitudes towards those we serve—to overcome stereotypes of those we serve.
−4	−3	−3	26. Will the definition of ''hard to serve'' change again and how will we respond to it?
3	0	−2	27. How do we best restructure PIC given potential changes in committee structure of board, management staff, recruit new personnel, etc.
0	4	−1	28. An income support standard so that once we have a placement we do not lose transportation, health care, child care, etc.
1	−1	4	29. How do we prioritize our programs so those programs best relate to the Economic Development Office programs.
1	1	−1	30. What will incentive be, negative or positive, for encouraging people to participate and successfully complete a PIC training program?
−3	−2	0	31. How will PIC programs be evaluated for success?
−3	1	−2	32. To avoid spreading ourselves too thin—better to concentrate on limited number of programs and succeed rather than trying to be everything to everybody.
3	0	−1	33. PIC has reacted to mandates of state and federal governments. How do we shape and affect state and federal policies rather than simply reacting to them.

There were other statements receiving high positive and negative scores, but the particular statements above were selected because they also serve to distinguish factor A from B and C; i.e., the factor scores associated with the above statements in A are significantly different from the scores for the same statements in B and C, as Table 5 shows (as a rule of thumb, differences in score of 2 or more are significant, $p < .01$; for details, consult Brown, 1980: pp. 244–246).

The statements given emphasis suggest that the persons comprising factor A are prepared to advocate on behalf of their agency and what they understand to be their clientele group—the ''hard to serve,'' which are conceived not as the ''previously employed'' (note the score assigned statement 8 above), but mainly as those lacking employable skills. A key to factor A's emerging theme is seen in statement 25, where the factor acknowledges stereotypes and the need to overcome them; in giving prominence to this statement, however, factor A is implicitly chastising the other two factors on the PIC Board (which score this same statement −2 and −3, respectively). Factor A's benevolent attitude toward the hard to serve is further revealed in statement 7, in which the factor rejects the idea of training the unskilled for dead end jobs that match their current aptitudes (cf. statement 16, Table 5); rather, the factor is interested in job training and in getting the hard to serve into positions that have some future. The high positive scores given statements 22 and 27 above are a measure of factor A's activist vision for the PIC agency.

Factor A's theme (of advocacy on behalf of the hard to serve) is thrown into sharper relief when it is contrasted with the different commitments of the other two factors. Consider first a few of those statements to which factor B gave high scores and which distinguish B from factors A and C (see scores in Table 5):

> Distinguishing Statements for Factor B (+4, +3): (7) Effectively test for aptitude and skills—specifically train where they have the best potential. . . . (15) Be prepared for the idea of limited resources. Prepare: (a) Plan A, with no improvement in funding; (b) Plan B, worsening of the economy; (c) Plan C, serious cuts in federal and state resources. . . . (28) An income support standard so that once we have a placement we do not lose transportation, health care, child care, etc.

From these and other statements it becomes apparent that factor B's orientation is not based first and foremost on empathy with the clientele group, but is concerned with the management of scarce resources: these are agency managers rather than advocates for the downtrodden. As for factor C, the concern is mainly with economic development as opposed to client improvement:

> Distinguishing Statements for Factor C (+4, +3): (9) Increase EDWA funding and also the possibility of apprenticeships. . . . (17) More aggressive in economic development—e.g., get with building trades/realtors to put together packages of what we could do for industries. . . . (29) How do we prioritize our programs so those programs best relate to the Economic Development Office programs.

A more detailed examination of the factor C array in Table 5 reveals that its commitment to economic development goes hand in hand with an opposition to training programs for the unskilled (statements 2, 5) and an emphasis instead on the previously employed (statement 8). It was in part this struggle over scarce resources between the unskilled and the skilled unemployed that contributed to tensions among their advocates on the PIC Board and that ultimately led to the consultation resulting in the above analysis.

Before concluding, it is important to note that just as there are statements which can serve to distinguish a factor from all others, so are there often statements on which all factors converge as a matter of consensus. In the instant case, there is precious little common ground on which the three factors might stand together:

> (14) Cultivating more job sites. . . . (23) Marketing issue—every eligible person in the county should know about PIC services—adults and youths. Every business in the county should be a member of PIC and should know of benefits to business.

The scores for these two statements range from 0 to +3 (see Table 5), but the fact that none has a negative score gives rise to the possibility that factors A, B, and C could begin to cooperate by endeavoring to effect outcomes in the directions in which these statements point. In this regard, Harold Lasswell (1963) once remarked that ''a continually redirected flow of events can be progressively subordinated to the goals conceived in central decision structures'' (p. 221), and those structures which he conceived are not unlike factors A, B, and C, which are clearly aimed at directing the flow of events. Whether any one or all three of them can succeed in bending future events to their preferences depends at least in part on the kind of self-clarification which Q methodology can provide, as the preceding illustration demonstrates.

C. Questions Frequently Asked About Q Methodology

As researchers and practitioners employ Q methodology in their work, they often encounter questions about how to carry out their work and challenges from R-oriented researchers about the appropriateness of using Q methodology. In this subsection of the chapter, we present some of the questions most frequently asked about Q methodology and our responses.

1. How can Q methodology be useful if it does not include a random sample of the population of interest? Without such a random sample of people completing Q sorts, can Q really reach generalizations?

There are a number of ways to respond to this issue. Consider for illustrative purposes two recent interlocking studies by Maxwell (1996a,b), in the first of which she asked the staff of a

middle school to list all of the problems which had led them to seek her consultation. The N = 44 problems were varied (e.g., ''Some teachers enforce the rules and some don't,'' ''There's no support from the Administration,'' etc.), and a Q sorting of the problems by n = 30 of the staff produced two factors. These factors represented two uncorrelated perceptions of what constituted important problems, but within the breach were also points of consensus which revolved around student discipline. An immediate follow-up study then asked what steps might be taken to deal with this commonly agreed-upon problem, and the N = 35 proposals which the staff generated were then Q sorted by the same staff members. The three resulting factors from this second study pointed to three different understandings about how to attack the discipline problem effectively. Again, consensus emerged, this time around the desirability of consistently enforcing already existing rules, of involving parents, and of standardizing consequences for specific behaviors.

In terms of the above query, we can conceive of the school study as having involved the entire population in question, thereby obviating inferential considerations. However conceived, there was never a question of generalizing these results to all middle schools, or even to other middle schools in the same city. The problems facing the staff (study 1) and the proposed solutions to these problems (study 2) were highly specific to the setting from which the factors emerged. Whether population sampling even makes sense, therefore, depends in large measure upon the question being asked.

A second response to the above query revolves around the concept of representativeness more generally conceived. A well known principle of representative design (Brunswik, 1949; Brown and Ungs, 1970) holds that generalizations depend both on the representativeness of responders (person sampling) and on the representativeness of the situations to which individuals respond (stimulus sampling): the former supports generalizations regarding the conditions under which results are obtained, and the latter regarding the situations to which the results apply. Q methodology approximates situational representativeness in the breadth of its statement samples—while Q sorting, the person's attitude is repeatedly brought into play across the wide variety of possibilities contained in the Q sample; on the other hand, survey studies, virtually without exception, fail to provide for representativeness on the stimulus side. Moreover, behavior tends to be more variable from situation to situation than from person to person, and so it is on the stimulus side that the principle of representativeness is most in need of application.

Third, it should not be overlooked that Q factors are themselves generalizations. For example, the Q sort representing the perspective of factor A (see Table 5) is the way, *in general*, that persons of this type think, and the factor B Q sort is how persons of the B variety think. . .*in general*. Factors represent qualitatively different modes of thought that retain their distinctive features no matter how many persons of each kind are included in a study.

Finally, there can never be a guarantee that the factors discovered in a sample of respondents will necessarily be exhaustive. In a study of administrators, for instance, were we to find factors reflective of Weberian traditionalists, charismatics, and rationalists, it could not be assumed that these three were the only factors: Factors IV, V, and VI might appear in future studies. Nor could it be assumed that these same three factors would all necessarily appear in each and every organization. These are empirical matters. Quite apart from the relative numerical strength of factors, however, what can be done is to compare them as phenomena—e.g., to examine how a traditional-minded administrator differs from a charismatic administrator—and for this we only require a handful of each kind.

2. Does it Matter Whether the Q Sorts are Forced-Normal or Not?

This query has reference to the fact (as illustrated in Table 1) that Q sorts are typically quasinormal in form and are ''forced'' in the sense that all respondents are given instructions for adhering

to it—e.g., that each Q sorter is instructed to assign only two statements the score of +4, only three the score of +3, and so forth. This particular feature of Q technique initially attracted considerable attention, but it is a relatively simple matter to demonstrate that the Q-sort distribution, no matter what shape, has negligible impact on results (e.g., see Brown, 1980, pp. 288–289). Why, then, insist upon a particular shape? The reasons are more theoretical and pragmatic than statistical. Theoretically, the Q sort distribution is an expression of the Law of Error and is an idealized version of the way in which respondents typically respond when left to their own devices (Brown, 1985), i.e., when allowed to distribute the statements ''freely''; the forced distribution is typically platykurtic in the same way that a *t* distribution is a flattened version of a normal distribution.

But the more important reason for employing the forced distribution is pragmatic. To reveal heartbeat, doctors instruct patients to engage in artificial behavior (such as stair-stepping or running in place) so as to induce an operant response, and the same principle applies in Q technique: The individual has preferences (e.g., about alternative policies), and is instructed to engage in the artificiality of Q sorting—artificial in range (e.g., +4 to −4) and in distribution shape (e.g., quasinormal)—so as to force those preferences into the open as an operant response (Stephenson, 1977).

3. *Can Q and R Studies be Linked? For Example, Can the Results of a Q Analysis (Such as the Factor Loadings) be Used in Regression Analyses?*

Such analyses can, of course, be done by straightforward technical extension. Consider the rotated factor loadings in Table 4: These values (suitably transformed logarithmically) could be correlated with any other quantitative variables. Stephenson (1953: pp. 190–218) has also suggested questionnaire construction based on statements which differentiate factors. From Table 5, note the factor scores for the following three statements (scores for factors A to C, respectively):

−1 −1 +3 (statement 9)
0 +4 +1 (statement 15)
+3 −2 −3 (statement 25)

Statement no. 9 obviously distinguishes factor C from the other two factors, and a person presented with the above three items (which are akin to a mini Q sort) and who selected no. 9 as most important could be assumed to be of the factor C type. We would naturally wish to add additional discriminating statements for purposes of reliability, but in this way a questionnaire could be constructed that would serve to identify a person's type membership, and which could then be administered to large samples (see Theiss-Morse et al., 1992, for an example).

But to utilize the Q sort itself for large-sample studies is to rely on a technique that was crafted for other purposes. Large scale correlation and regression studies, for example, are based on overall averages—the regression line is anchored at that point where the means of X and Y intersect—whereas Q is best suited for locating disjunctures and disaggregates (e.g., Rhoads and Sun, 1994), and for the study of single-cases (e.g., Taylor et al., 1994).

4. *Must Q Sorts be Administered in Person, or is it Possible to do Q Studies by Mail?*

Obtaining Q sorts by mail will generally produce the same factors as those produced following conventional procedures (Van Tubergen and Olins, 1979), but the practice is only recommended when there are no alternatives. The reason is obvious: Q methodology is designed to facilitate

a science of subjectivity, and this implies getting more acquainted with a phenomenon rather than achieving distance from it. The best Q studies are generally those in which the Q sorting is immediately followed by an intensive interview during which the person fleshes out the skeletal view contained in the Q sort, and both large samples and mailing—and, more recently, Q sorting via the Internet—mediate against gaining ''a feeling for the organism'' (Brown, 1989). Still, there are obviously occasions when mailing and related means cannot be avoided.

5. Must Analysis of Q-Sort Data be Performed Using Statistical Packages Such as SAS or SPSS, or Are Other Software Packages Available for this Task?

The analysis culminating in Tables 3, 4, and 5 was performed using QMethod (Atkinson, 1992), which is a mainframe freeware program written in Fortran and available in IBM, VAX, and UNIX versions; all are retrievable from the Listserver on the Kent State University mainframe computer. A PC version of QMethod is also accessible from a private site on the World Wide Web. As of this writing, the third version of Stricklin's (1996) PCQ package is being beta-tested: this is a PC-only commercial program that is graphically more sophisticated than the mainframe products. For further details, consult Brown (1996). Both QMethod and PCQ have Q-methodological presuppositions built into them, hence will likely be experienced as user-friendly or not depending on the background of the user. Conventional researchers who depend on large numbers of respondents will undoubtedly be more at home with SPSS or SAS and automatic options (such as Varimax rotation) for the factor analysis of Q-sort data; however, those who have delved into Q as a methodology (as opposed to a mere statistical procedure) will gravitate to these more tailor-made packages containing centroid analysis, judgmental rotation, and other features not to be found in the large commercial programs.

6. Q Methodology Seems to Rely Heavily on the Personal Judgments of the Researcher (e.g., the Purposive Selection of Statements for the Q Sample and of the Persons to Whom Q Sorts are to be Administered, the Judgmental Rotation of Factors, Interpretation of the Results, etc.) Isn't the Outcome Therefore Simply a Function of Decisions Made by the Researcher? And if so, Then is This Really Science?

There is a humorous comment about the British, that in place of thinking they have traditions. In the same way, much of what passes for objectivity in science is little more than custom and routine, and it is this that often creates the impression that personal judgment has been safely reined in. No self-respecting scientist, for example, would ever dream of adjusting the location of the regression line on the basis of some personal whim, and the same can presumably be said for the theoretical rotation of factors in Q methodology. Hence, Varimax or other rotational scheme is conventionally used where choice would otherwise be.

Much of the criticism of investigator whim and bias comes from those who are inexperienced with regard to Q methodology and who therefore imagine the investigator to be freer than is actually the case. Although not demonstrated in the case study presented above, statements assembled into a Q sample are typically selected according to a plan (typically structured as a factorial design) that constrains judgment. To provide a brief illustration: Suppose we were invited by an organization to appraise the skill/role fit of its members, and assume further that interviews with members produced comments such as the following:

1. I prefer goals to be clear, with subgoals and responsibilities well articulated.
2. I can analyze data all day and be just happy as a lark.
3. I value opportunities to be creative.
4. I like opportunities to solve problems. And so forth.

Any of a number of conceptual frameworks might be used to insert order into this chaotic mass of opinion, but let's assume that we have settled on Simon's (1960) schema—that there are two types of decision: (a) programmed and (b) nonprogrammed; and that there are two types of decision making techniques: (c) traditional and (d) modern. We can now return to the statements above and begin, at least hypothetically, to categorize them: hence statement (1) is of the type (ac), statement (2) of the type (ad), and so forth. Once the statements have been placed into the four categories, we might select, say, m = 10 replicates from each, for a Q sample size of N = 40. There is obvious choice involved within each of the cells, but also constraints imposed by the factorial design. Even the selections within cells are constrained by principles, e.g., that the 10 statements of the (ac) type should be as different as possible so as to enhance breadth (i.e., representativeness) in the Q sample. And the procedures adopted for statement selection apply as well to the selection of respondents—e.g., by gender (male/female), level of management (upper/middle), etc. Further details are contained in the standard volumes on Q methodology (Brown, 1980; Stephenson, 1953; McKeown and Thomas, 1988).

Those concerned that investigator bias will exercise undue influence over results fail to appreciate the central role of the Q sorter, whose own subjective understanding is at the center of the enterprise. Statements in a Q sample are not assumed to carry meaning a priori, as in an attitude scale; rather, the Q sorter projects meaning onto the statement, a posteriori. Even in the Simon example above, it is a central principle of Q methodology that the hypothetical meanings that informed the Q sample construction give way to the actual meanings attributed by the person performing the Q sort.

The issue of theoretical rotation of factors is not one that can be satisfactorily addressed in the space available (for a worked example, see Brown, 1980: pp. 224–239). Suffice it to say that it is rooted in the principles of interbehaviorism (Kantor, 1959) and abductory logic (Peirce, 1955), and that unlike other procedures (e.g., Varimax), it is not aimed at finding a preexisting structure but at probing and examining subjective space. Moreover, as anyone who has ever engaged in theoretical rotation already knows full well, the investigator rotating the factors is often educated in the process by the data themselves, and this dialectical interaction (an aspect of Kantor's interbehaviorism) affects subsequent rotational decisions. An assumption behind the habitual use of Varimax (or any of the other automatic rotational algorithm) is that there is a single geometric model that applies universally across all problems and contexts, whereas an assumption of theoretical rotation in Q methodology is that each situation is unique and has its own logic and determinism which the computer program cannot know, but which can be incorporated via the understanding of an investigator familiar with the situation.

As to the factors and their interpretation: Like the Q sorts which comprise them, the factors represent subjective understandings—the Q sorters' understandings—and the investigator trying to grasp those meanings quickly realizes that in order to do so it is necessary to enter into a receptive state of mind that is as devoid as possible of preconceived ideas that might get in the way of listening. Subordination of bias to discipline of this kind cannot be vouchsafed, of course, but a strong constraint on any investigator's predispositions is the array of factor scores (see Table 5) to which any interpretation, biased or otherwise, must conform. As demonstrated in the case study reported above, the interpretation of a factor must come to terms with the entire gestalt of all the statements, i.e., the interpretation must have the kind of coherency that incorporates the entire pattern of elements.

It bears mentioning that Stephenson was not only a renown psychologist (Ph.D. London, 1929) but also a physicist (Ph.D. Durham, 1926), and that he had a lifelong interest in science. Such parading of credentials guarantees nothing, of course, but hopefully gives pause to those who might be otherwise inclined to dismiss his admittedly unusual ideas as being without scientific merit.

We conclude this section by noting additional sources available for information about Q methodology: (1) The journal *Operant Subjectivity*, issued quarterly since 1977; (2) the annual scientific meeting of the International Society for the Scientific Study of Subjectivity, begun in 1985; and (3) the Listserve discussion forum Q-METHOD, established in 1991 at Kent State University. A bibliography of applications now numbering more than 2000 appears on a continuing basis in *Operant Subjectivity*.

IV. APPLICATIONS OF Q METHODOLOGY IN PUBLIC ADMINISTRATION: A BIBLIOGRAPHICAL GUIDE AND RESEARCH AGENDA

As mentioned and illustrated in the above sections, Q methodology is useful for both public administration academicians and practitioners. It is pertinent to the study of many fields and areas of public administration, including general public administration topics, public personnel administration, public management, decision making, and public policy making. For each of these areas, we describe in this section several applications of Q methodology, and then we suggest possible avenues for future research using Q. We should note that this section only suggests Q methodology's potential contribution to the study of public administration, and it is not intended to be perceived as exhaustive in the development of that potential. We believe that the possibilities for application of Q methodology are boundless wherever subjectivity is implicated.

A. The Use of Q Methodology in the Study of General Public Administration

Q methodology is useful in studying situations in which individuals are likely to hold an opinion or viewpoint about a matter or an event occurring around them. Thus Q methodology is well suited to investigate the viewpoints and attitudes of bureaucrats, which has received much attention in public administration. In fact, several public administration scholars have used Q methodology to better understand the views of bureaucrats toward issues such as affirmative action programs, bureaucratic discretion, and administrative ethics (Decourville and Hafer, 1995; Hiruy, 1987; Wood, 1993).

To illustrate how Q can be used to identify different attitudes toward a subject, consider the Q study of Thai administrators by Vajirakachorn and Sylvia (1990). The objective of this research was to investigate the influence of traditional Buddhist culture and modern bureaucracy on Thai administrative attitudes. Ninety-four Thai administrative elites in the Ministry of Interior were asked to sort and rank 54 statements.[7] From this process, four administrative types emerged: the first group endorsed modern and bureaucratic values most strongly; the second group was characterized by mixed attitudes and held values that were midway between modern bureaucratic principles and Buddhist traditional ideas; and the third and fourth groups expressed a slight mix in opinion, but in general they expressed more agreement with values that correspond to traditional bureaucratic practices, such as planning in management, rationalism and scientific reasoning in making decisions, and merit-based civil service. This research illustrates the utility of Q in discerning attitude differences among public administrators.

In another comparative study of public administrators, Gough et al. (1971) compared the managerial perspectives and preferences of American and Italian public managers. The Q sorts of Italian administrators were analyzed and the following administrative types were identified: innovator, mediator, actionist, the moderate, achiever, and the realist. Then, the authors collected Q sorts from 110 American administrators to analyze typological variations among the two

cultures (Gough et al., 1971: p. 256). The addition of American public administrators revealed some stylistic variations in perspectives and practices among the two groups of administrators.

The result of the Q sorts showed that American administrators perceived interpersonal relationships and career opportunities to be more important, whereas Italian administrators were more concerned with job security and structure. In terms of the administrative types identified in the analysis of Italian administrators, American bureaucrats identified more with the mediator role; that is, they perceived themselves as tolerant, modest in demands, and generous in relationships. American administrators were least likely to assume the role of actionist; that is, one who is tough-minded, decisive, and indifferent to the feelings of subordinates. With this research, Gough et al. (1971) used Q methodology to identify alternative self-understandings of managerial responsibilities and preferences.

Other applications of Q methodology in public administration include studies of administrative roles and types of American public administrators (Durning and Osuna, 1994; Johnson, 1973; McDale, 1989; Scheb, 1982; Vroman, 1972), work orientations of middle-level American managers (Shah, 1982), and American city-manager prototypes (Faulhaber, 1969). Also, Yarwood and Nimmo (1975) examined different stakeholders' perceptions of bureaucratic images and how these images differed between academicians and other stakeholders. In another study, Yarwood and Nimmo (1976) explored the ways in which administrators, legislators, and citizens orient themselves to bureaucracy and the accuracy of these groups in estimating the attitudes of each other toward the bureaucracy.

More recently, Sun (1992) used Q methodology to examine public administration scholars' and practitioners' opinions regarding the practical use of scholarly research in Taiwan. Sun (1992) identified four factors, two of which were bipolar. The first factor was bipolar: one group of respondents sensed scholarly research was not utilized because of organizational and political constraints, whereas the other group believed that a gap existed between scholarly research and reality because scholars and practitioners failed to communicate with one another. Individuals loading on the second factor judged that practitioners failed to use scholarly research because the two entities hold different assumptions about the functions of public administration. Like the first factor, the third factor was bipolar, and one group viewed institutional constraints as the primary barrier to practical use of scholarly research. The other group believed that ''public administration is primarily an art and relies on experience and skill rather than book knowledge'' (Sun, 1992: p. 291). Subjects that loaded on the final factor held the opinion that scholars and practitioners operate independently and irrespective of one another.

Each of the factors that emerged from Sun's analysis revealed different viewpoints to explain why practitioners do not apply scholarly research. This information is valuable because it enables researchers to address specific misgivings of practitioners toward scholarly research, as well as provide some direction as to how to bridge the gap between academia and applied practice.

Beyond the use that has already been made of Q methodology to research general public administration topics, it could be employed to study a myriad of other public administration issues. For example, a Q study of bureaucratic responsibility would provide a new approach for examining an enduring and important issue. Most of the empirical research on bureaucratic responsibility has relied primarily on R methodology and has frequently concentrated on a single method of ensuring bureaucratic responsibility. However, with Q methodology different conceptions of bureaucratic responsibility could be considered and evaluated simultaneously. In this regard, a useful framework for such a study would be Gilbert's (1959) two-by-two typology of administrative responsibility. The concourse of statements would represent the four categories that emerge from dividing the horizontal axis into internal and external categories of responsibility and splitting the vertical axis into formal and informal means of control. The researcher

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

could specify different conditions for sorting the statements. For example, subjects could be asked to order the statements to reflect to whom the administrator feels most responsible or, alternatively, to indicate to whom the subject perceives that he or she should be responsible. Moreover, bureaucrats' perceptions of responsibility could be compared to other important stakeholders, such as elected officials and the public, to identify the extent to which views of bureaucratic responsibility converge among these groups.

As suggested by research described above, Q methodology facilitates the study of administrative roles and could be applied to study other frameworks. For example, a Q study could be based on Faerman et al.'s (1990) competing values model of managerial roles, which has been used extensively to study public management (see, for example, Giek and Lees, 1993). So far, the empirical research using this model has relied on extensive closed-end questionnaires that address numerous managerial tasks (Ban, 1995). An alternative approach would be to use Q methodology to allow public managers to operantly define and formally model their attitudes and perspectives towards the managerial roles set forth in the competing values framework. The competing values model is based on a two-dimensional scheme: the horizontal axis is split into two categories of focus, internal versus external, and the vertical axis is divided into two managerial approaches, control versus flexibility. The quadrants that emerge reflect existing models from the classic management literature: (A) human relations model, (B) open systems model, (C) internal process model, and (D) rational goal model. Each quadrant or model includes two specific roles. Hence, A is identified with (a) facilitator and (b) mentor; B with (c) innovator and (d) broker; C with (e) monitor and (f) coordinator; and D with (g) director and (h) producer. Administrators would be asked to weight the alternative roles and to sort the statements to reflect their role orientations.

Another widely cited typology that would be an excellent candidate for a Q study is Down's (1967) typology of administrators. To date, little effort has been made to verify the following five categories of administrators suggested by his typology: climbers, conservers, zealots, advocates, and statesmen (Rainey, 1991). Thus, we could ascertain how many, if any, of the types of administrators suggested by Downs emerge when individuals sort a representative body of statements regarding their administrative motives.

B. Researching Public Personnel Administration with Q Methodology

The use of Q methodology to better understand attitudes and orientations of administrators indicates that it might also be particularly useful to identify the personnel-related concerns of other employees, such as job satisfaction and motivation. In fact, several studies have used Q methodology for this purpose. For example, Sylvia and Sylvia (1986) collected 43 completed Q sorts about job satisfaction and work motivation from a randomly selected sample of mid-level rehabilitation managers. From this study, three factors were discovered and classified. The first, ''the positive concerned supervisor,'' identified managers as being concerned for subordinates, achievement, recognition, and work as a source of satisfaction. The second characterized job satisfaction as stemming from a positive attitude toward advancement, work, and co-workers. Individuals loading on the third factor experienced feelings of job satisfaction for a number of the same reasons as found in the first two factors, as well as from the freedom they were granted to try new ideas and programs.

Job satisfaction is widely studied using R methodology and, as illustrated above, some effort has been made to use Q to study job satisfaction. Despite this research, no coherent framework of factors that determine job satisfaction has surfaced. According, to Rainey (1991: p. 146), this absence of a coherent framework is not surprising ''because it is obviously unrealistic to try to generalize about how much any single factor affects satisfaction.'' Nevertheless,

progress might be made toward identifying the relative importance of different factors for job satisfaction through a Q method study. If such a study did provide insight into the facets of the job—such as supervision, pay, and promotion—that contribute to an individual's job satisfaction, this tool could be used to shape agency practices, training, and development (Erb, 1987).

Another personnel-related area of research that has been studied via Q technique is work motivation. Gaines et al. (1984) explored the relationship between perceptions of promotion and motivation in two Connecticut Police Departments. Specifically, they investigated the need structures of police officers and the extent to which promotion fulfilled those needs.

Applications of Q methodology in the study of work motivation could provide insights into what needs, motives, and values are most important to public sector employees. Such research could draw from a number of existing typologies such as Murray's List of Basic Needs (1938), Maslow's (1954) need hierarchy, Alderfer's (1972) ERG model, and Rokeach's (1973) Value Survey.

In addition, Q could be used to study the types of incentives that induce public sector employees to contribute positively to their agency. The following frameworks are suitable for studying this using Q methodology: Herzberg et al. (1957), Locke (1968), and Lawler (1973). Another potential application of Q would involve identifying methods and techniques that motivate employees. A concourse could be constructed to represent various methods and techniques employed to facilitate high performance in public organizations, such as performance appraisals, merit pay, pay-for-performance, bonuses, job redesign, job rotation, flex time, and quality circles. The completed sorts would show how employees would view and value alternative efforts to improve organizational performance.

Q would also be an appropriate technique to use to develop a self-assessment performance tool. Employees would be asked to sort a group of statements pertaining to their performance and the resulting Q sorts would represent the employees' own constructions of their performance strengths and weaknesses. As part of a performance appraisal system, Q could facilitate and structure feedback discussions and suggest employee skills and knowledge that need further development. Also, as suggested by Chatman (1989), Q could be used to assess the extent to which an individual ''fits'' into a specific public agency setting or job.

C. Researching Public Management with Q Methodology

Public management research is closely linked to general public administration and public personnel scholarship. Scholars have employed Q to describe organizational culture (O'Reilly et al., 1991) and to understand leadership (Dunlap and Dudley, 1965; Thomas and Sigelman, 1984; Wong, 1973). Beyond these uses of Q methodology, Q could be used to explore public-private distinctions. For example, scholars could explore patterns among public and private managers on a wide range of subjects, such as leadership, service motivation, organizational tasks and functions, and organizational characteristics.

Another potential application of Q would be to assess the trust of managers, an increasingly important topic in public management. Most of the existing research on trust is based on attitude surveys and personal interviews of public managers (Carnevale, 1995). Alternatively, Q methodology could be used to explore public managers' worldviews toward trust. A concourse of statements could be constructed that captures dimensions that are fundamental in understanding managerial philosophies (Wong, 1973: p. 35). These include:

1. The degree to which individuals perceive that people are trustworthy or untrustworthy;
2. The degree to which individuals believe that people are altruistic or selfish;

3. The degree to which individuals believe that people are independent and self-reliant or, alternatively, dependent and conformist;
4. The degree to which individuals believe that people are simple versus highly complex.

The two dimensions, trust (1 and 2) and control (3 and 4), suggest something about the kinds of management methods, personnel policies, and work procedures to which an individual would respond or elect to use as a manager (Carnevale, 1995). With a good understanding of one's own perceptions and the perceptions of others who work in the organization, a manager could shape agency practices and culture to maximize performance by facilitating views congruent with the mission and goals of the agency.

D. Researching Decision Making and Public Policy with Q Methodology

Often scholars analyze decision processes according to a contingency-theory perspective. That is, in some situations, administrators are able to adopt rational decision approaches when the following conditions are known: all relevant goals are clearly stated, all values for assessing these goals and levels of attainment of them are known, preferences of goals can be ranked, all alternative means of achieving these goals are examined, and the most efficient alternative is selected (Downs, 1967: p. 80). In many situations, however, the conditions outlined above cannot be met. Administrators encounter uncertainty, competing demands, unclear goals, and limited information and data. As Brown (1980: p. 71) has said, ''the situation is highly complex and multivalued. . . ; subjective judgment reigns; decisions are made and consequences accepted; values and preferences are everywhere involved. The methodological problem is one of modeling this phenomenon in all of its rich complexity, and of holding it steady for examination.''

Q has been used to secure agreement among decision-making participants, to weigh alternatives and their consequences, and to reduce the time required to make decisions (Cooper and Dantico, 1990; Coke and Brown, 1976; Grunig, 1969; Nutt, 1984). As described in the previous section, it can be used to assist with strategic planning (Gargan and Brown, 1993).

To illustrate further how Q methodology can be used to facilitate planning decisions, consider the Q study of land use options carried out by Fairweather et al. (1994). A concourse of 50 alternative land use options for the Mackenzie/Waitaki Basin in New Zealand was developed. Each option was presented visually, economically, and ecologically (integrated on a single card), and stakeholders were asked to sort the options to reflect their preference for future land use in the Mackenzie/Waitaki Basin. Three distinctive preferences for land use emerged:

- plantations: ''most important feature is the role of large plantation for production on the hills and lower slopes, and for conservation on the higher rainfall flats''
- grazing/trees: key features include a ''combination of trees and grazing for production, comprising plantations and grazing on the hills, and shelterbelts on the lower slopes and higher rainfall flats''
- conservation: ''the essential features are small plantations and conservation on hills, larger plantations and conservation on lower slopes, and retention of views on higher rainfall flats'' (Fairweather et al., 1994: p. 107).

The researchers intend to extend the planning process by conducting more detailed economic modeling and evaluation of the social consequences of the three land use preferences.

In the field of public policy, Q methodology has been used to explore the attitude climate among legislators and administrators on issues of policy making and policy implementation (Cunningham and Olshfski, 1986) and to evaluate public programs (Garrard and Hausman, 1985; Oring and Plihal, 1993).

Kim's (1991) study of community development block grants (CDBG) illustrates how Q can be used to judge implementation effectiveness. Kim collected 33 statements that covered policy design, intergovernmental relations, organization of implementing agency, implementing environment, and evaluation of the implementation process model. Kim administered the Q sort to citizens, General Accounting Office (GAO) administrators, and other employees of federal, state, and local agencies. Three factor were discovered, each of which represents distinct implementation concerns. Kim called the first the ''bureaucratic guild's view'' (reflecting bureaucratic interests concerned with CDBG implementation); the second, ''policy critic's view'' (includes critical view of implementation given CDBG's substantive goals); and the third, ''local view'' (represented perspective of local officials—a bottom-up perspective). The analysis also revealed statements which all respondents ranked in a similar way, indicating consensus among participants with respect to certain CDBG implementation problems. As this study illustrates, stakeholders are likely to perceive implementation problems differently depending on their position, role, and interests. Kim (1991: p. 163) concludes:

> Better understanding of the views and perceptions of other participants, however, is vital for creating a workable intergovernmental environment in which the complex undertaking will be handled by way of continuing political bargaining and negotiations among intergovernmental actors involved in the process of implementation of federally mandated programs.

Q methodology is a useful tool for facilitating decision and policy making by clarifying preferences and priorities, improving communication, and expediting the process by illustrating the specific policies about which a group is in fundamental agreement and disagreement. Q could be used, for example, to aid elected officials in budget negotiations or reaching consensus on community planning priorities.

V. CONCLUSION

In this chapter, we have discussed the motivations for using Q methodology and described in some detail how to carry out Q studies. We have suggested that some researchers and practitioners will use Q methodology pragmatically to answer important research and practical public administration questions from a perspective that differs from the usual R method approach. Other researchers and practitioners may turn to Q methodology in reaction to the shortcomings of R methods, which are founded on positivism, an epistemology that is being increasingly challenged by theorists in all social science disciplines.

Whatever the motivation for using Q methodology, public administration researchers and practitioners will find that this method can be valuable for their work. As we have described in the chapter, researchers have used Q method to investigate important issues in general public administration, public personnel administration, public management, and public policy, and this methodology is well suited for exploration of other key issues in these research areas. Q methodology can also be a valuable tool for public managers and policy analysts to identify and understand conflicting values, preferences, and opinions concerning organizational and policy issues. Also it has been used, and should be used further, in policy evaluations. In addition, it can contribute to the democratization of management and policy making by allowing the voices of stakeholders and the interested public to be more fully articulated and understood.

A researcher or practitioner who wishes to conduct a Q methodology study can do so by following the general procedures described in the case study of strategic planning in part III of this chapter. Many questions that might arise in such a study are addressed in the section on ''questions frequently asked about Q methodology.'' We believe that the value of Q methodol-

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ogy, both as a pragmatic tool and as a nonpositivist or postpositivist research method, is much more than sufficient to reward public administration researchers and practitioners fully for their efforts to master it.

NOTES

1. The procedures for carrying out Q method research are discussed in more detail in section III of this chapter. The most useful and complete methodological guides are Brown (1980, 1993) and McKeown and Thomas (1988). After mastering the basics of Q methodology, researchers may want to read Stephenson (1953), *The Study of Behavior*, which laid the foundation for Q methodology as the science of subjectivity (Brown, 1995).
2. See Dryzek (1990), Chapters 8 and 9, for an in-depth comparison of the use of survey research (an R methodology) and Q methodology to investigate public opinion.
3. We should note that most ''good'' positivists have been quite aware of the limits of science. For example, Stephenson, the creator of Q methodology, regarded himself as a positivist and would not have taken the extreme position attributed to positivists by some postmodernists. Stephenson's view of science was not defended on the basis of objectivity, but of detachment, that is trying to establish conditions under which the observer might have reason to believe that what was being observed was independent of the self and its desires. Ironically, it is precisely in terms of detachment that R methodology falters: the measuring device (e.g., an autocratic management scale) carries the observer's undetached meaning along with it; in Q, on the other hand, we permit the respondent's meaning to reign, as detached from whatever meaning we might have given to the same statements. It is Q, not R, that comes closest to achieving detachment (which again, is not a claim of ''objectivity'').
4. The importance of context is one of the principal ideas of the policy sciences movement that traces its lineage back to Harold Lasswell (see Ascher, 1986; 1987).
5. Ascher (1987) argues that both the policy sciences and Q methodology are ''staunchly non-positivist.'' The policy sciences movement dates back to the early 1950s.
6. Many non-positivists and post-positivists believe that the focus on individual unique realities should be balanced with a discussion of the limited variety of these realities. In fact, all is not specificity, there is also communality of perspectives, and Q method is well suited to identify those commonalities among the different realities.
7. Statements about the two main concepts were collected for the following nine issues: source of authority, dominant values in work, decision making patterns, recruitment, placement, transfer and promotion, superior-subordinate relationships, work performance, accountability, and group-orientation.

REFERENCES

Adcock, C.J. (1954). *Factorial Analysis for Non-mathematicians*, Melbourne University Press, Melbourne.

Alderfer, C.P. (1972). *Existence, Relatedness, and Growth: Human Needs in Organizational Settings*, Free Press, New York.

Ascher, W. (1986). ''The Evolution of the Policy Sciences: Understanding the Rise and Avoiding the Fall,'' *Journal of Policy Analysis and Management*, *5*: 365–373.

Ascher, W. (1987). ''Subjectivity and the Policy Sciences,'' *Operant Subjectivity*, *10*: 73–80.

Atkinson, J. (1992). QMethod (computer program), Computer Center, Kent State University, Kent, OH.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Balfour, D.L. and W. Mesaros (1994). ''Connecting the Local Narratives: Public Administration as a Hermeneutic Science,'' *Public Administration Review*, *54*: 559–564.
Ban, C. (1995). *How do Public Managers Manage? Bureaucratic Constraints, Organizational Culture, and the Potential for Reform*, Jossey-Bass Publishers, San Francisco, CA.
Brown, S.R. and T.D. Ungs (1970). ''Representativeness and the Study of Political Behavior,'' *Social Science Quarterly*, *51*: 514–526.
Brown, S.R. (1980). *Political Subjectivity: Applications of Q Methodology in Political Science*, Yale University Press, New Haven, CT.
Brown, S.R. (1985). ''Comments on 'The Search for Structure,'' *Political Methodology*, *11*: 109–117.
Brown, S.R. (1988). *Perspectives on the Commission's Role and Objectives*, Report to the International Commission on Central American Recovery and Development, Stockholm.
Brown, S.R. (1989). ''A Feeling for the Organism: Understanding and Interpreting Political Subjectivity,'' *Operant Subjectivity*, *12*: 81–97.
Brown, S.R. (1992–1993). ''On Validity and Replicability,'' *Operant Subjectivity*, *16*: 45–51.
Brown, S.R. (1993). ''A Primer on Q Methodology,'' *Operant Subjectivity*, *16*: 91–138.
Brown, S.R. (1994). ''Representative Exposure and the Clarification of Values,'' Paper read at a meeting of the Policy Sciences Institute, Yale University School of Law, New Haven, CT.
Brown, S.R. (1994–1995). ''Q Methodology as the Foundation for a Science of Subjectivity,'' *Operant Subjectivity*, *18*: 1–16.
Brown, S. (1996). ''Q Methodology and Qualitative Research,'' *Qualitative Health Research*, *6*: 561–567.
Brunswik, E. (1949). *Systematic and Representative Design of Psychological Experiments*, University of California, Berkeley and Los Angeles, CA.
Burt, C. and W. Stephenson (1939). ''Alternative Views on Correlations between Persons,'' *Psychometrika*, *4*: 269–281.
Burt, C. (1940). *The Factors of the Mind*, University of London Press, London.
Carnevale, D.G. (1995). *Trustworthy Government*, Jossey-Bass Publishers, San Francisco, CA.
Chartier, R. (1995). *Forms and Meanings*, University of Pennsylvania Press, Philadelphia, PA.
Chatman, J.A. (1989). ''Improving Interactional Organizational Research: A Model of Person-Organization Fit,'' *Academy of Management Review*, *14*: 333–349.
Coke, J.G. and S.R. Brown (1976). ''Public Attitudes About Land Use Policy and Their Impact on State Policy-makers,'' *Publius*, *6*: 97–134.
Cooper, D.R. and M.K. Dantico (1990). ''Priority Setting Techniques for the Budgetary Process: An Experimental Comparison to Determine What Works,'' *Experimental Study of Politics*, *9*(1): 94–117.
Cunningham, R. and D. Olshfski (1986). ''Interpreting State Administrator-Legislator Relationships,'' *Western Political Science Quarterly*, *39*: 104–117.
DeCourville, N. and C. Hafer (1995). ''Attitudes Toward Affirmative Action Programs: A Q-Methodological Study,'' *Canadian Psychology*, *36*(2a): 114.
De Lancer, P.D. (1998). ''Principal Component Analysis, Factor Analysis, and Other Clustering Techniques,'' *Handbook of Research Methods in Public Administration*, G.J. Miller and M.L. Whicker, (eds.), Marcel Dekker, New York, NY.
Denhardt, R.B. (1981). ''Toward a Critical Theory of Public Organization,'' *Public Administration Review*, *41*: 628–635.
Downs, A. (1967). *Inside Bureaucracy*, Little Brown, Boston, MA.
Dryzek, J.S. (1982). ''Policy Analysis as a Hermeneutic Activity,'' *Policy Sciences*, *14*: 309–329.
Dryzek, J.S. (1990). *Discursive Democracy*, Cambridge University Press, Cambridge, England.
Dunlap, M.S. and R.L. Dudley (1965). ''Quasi-Q-sort Methodology in Self-evaluation of Conference Leadership Skill,'' *Nursing Research*, *14*: 119–125.
Durning, D. and D. Edwards (1992). ''The Attitudes of Consolidation Elites: An Empirical Assessment of Their Views of City-County Mergers,'' *Southeastern Political Review*, *20*: 355–383.
Durning, D. and W. Osuna (1994). ''Policy Analysts' Role and Value Orientations,'' *Journal of Policy Analysis and Management*, *13*: 629–657.
Ellis, C. and M.G. Flaherty (eds.) (1992). *Investigating Subjectivity: Research on Lived Experience*, Sage Publications, Newbury Park, CA.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Erb, M. (1987). *Identification of Training Needs: A Focus Group Interview/Q-sort Methodology*, Unpublished Dissertation, Western Kentucky University.

Faerman, S.R., R.E. Quinn, M.P. Thompson, and M.R. McGrath (1990). *Supervising New York State: A Framework for Excellence*, Governor's Office of Employee Relations, Albany, NY.

Fairweather, J.R., S. Swaffield, L. Langer, J. Bowring, and N. Ledgard (1994). *Preferences for Land Use Options in the Mackenzie/Waitaki Basin: A Q-method Analysis of Stakeholders' Preferences for Visual Images of Six Land Uses on Four Land Forms*, Agribusiness and Economic Research Unit, Lincoln University, Canterbury, New Zealand.

Farmer, D.J. (1995). *The Language of Public Administration: Bureaucracy, Modernity and Post Modernity*, University of Alabama Press, Tuscaloosa, AL.

Faulhaber, R.L. (1969). *City-Manager Prototypes*, Unpublished M.P.A. thesis, Kent State University.

Finkel, N. (1995). *Commonsense Justice: Jurors' Notions of the Law*, Harvard University Press, Cambridge, MA.

Fisher, F. (1990). *Technocracy and the Politics of Expertise*, Sage Publications, Newbury Park, CA.

Forester, J. (1990). ''No Planning or Administration without Phenomenology?'' *Public Administration Quarterly, 50*: 50–65.

Forester, J. (1993). *Critical Theory, Public Policy and Planning Practice*, State University of New York Press, Albany, NY.

Gaines, L.K., N. Van Tubergen, and M.A. Paiva (1984). ''Police Officer Perceptions of Promotion as a Source of Motivation,'' *Journal of Criminal Justice*, *12*: 265–275.

Gargan, J.J. and S.R. Brown (1993). '''What is to be Done?' Anticipating the Future and Mobilizing Prudence,'' *Policy Sciences*, *26*: 347–359.

Garrard, J. and W. Hausman (1985). ''The Priority Sort: An Empirical Approach to Program Planning and Evaluation,'' *Evaluation Studies Review Annual 11*: 279–286.

Giek, D.G. and P.L. Lees (1993). ''On Massive Change: Using the Competing Values Framework to Organize the Educational Efforts of the Human Resource Function in New York State Government,'' *Human Resource Management*, *32*: 9–28.

Gilbert, C. (1959). ''The Framework of Administrative Responsibility,'' *Journal of Politics*, *21*: 373–407.

Gough, H.G., R. Misiti, and D. Parisi (1971). ''Contrasting Managerial Perspectives of American and Italian Public Administrators,'' *Journal of Vocational Behavior*, *1*: 255–262.

Grunig, J.E. (1969).''Information and Decision Making in Economic Development,'' *Journalism Quarterly*, *46*: 565–576.

Guba, E.G. (1985). ''What Can Happen as a Result of a Policy?'' *Policy Studies Review*, *5*: 11–16.

Harmon, M.M. (1989). '''Decision' and 'Action' as Contrasting Perspectives in Organization Theory,'' *Public Administration Review*, *49*: 144–150.

Herzberg, F., B. Mausner, R.O. Peterson, and D.F. Capwell (1957). *Job Attitudes: Review of Research and Opinion*, Psychological Service of Pittsburgh, Pittsburgh, PA.

Hiruy, M. (1987). *Exploring the Perspectives of Ethics: The Case of Public Administrators in the United States*, Unpublished Ph.D dissertation, Kent State University.

Hummel, R.P. (1987). *The Bureaucratic Experience*, 3rd ed., St. Martin's Press, New York, NY.

Jennings, B. (1987). ''Policy Analysis: Science, Advocacy, or Counsel?'' *Research in Public Policy Analysis and Management*, *4*: 121–134.

Johnson, P.E. (1973). *Functional Typologies of Administrators: A Q-sort Analysis*, Unpublished Ph.D dissertation, University of Iowa.

Kantor, J.R. (1959). *Interbehavioral Psychology*, 2nd ed., Principia Press, Granville, OH.

Kelly, M. and S. Maynard-Moody (1993). ''Policy Analysis in the Post-positivist Era: Engaging Stakeholders in Evaluating the Economic Development District Programs,'' *Public Administration Review*, *53*: 135–142.

Kim, S.E. (1991). *Factors Affecting Implementation Effectiveness: A Study of Community Development Block Grant Implementation*, Unpublished Ph.D. dissertation, Kent State University.

Kirkhart, L. (1971). ''Toward a Theory of Public Administration,'' *Toward a New Public Administration: The Minnowbrook Perspective*, F. Marini (ed.), Chandler Publishing Co., Scranton, PA, pp. 127–164.

Lasswell, H.D. (1951). ''The Policy Orientation,'' *The Policy Sciences*, D. Lerner and H.D. Lasswell (eds.), Stanford University Press, Stanford, CA, pp. 3–15.
Lasswell, H.D. (1963). *The Future of Political Science*, Atherton, New York, NY.
Lawler, E.E. (1973). *Motivation in Work Organizations*, Brooks Cole, Pacific Grove, CA.
Lewis, C.S. (1960). *Studies in Words*, Cambridge University Press, Cambridge, England.
Locke, E.A. (1968). ''Toward a Theory of Task Motivation and Incentives,'' *Organizational Behavior and Human Performance*, *3*: 157–159.
Maslow, A.H. (1954). *Motivations and Personality*, Harper and Row, New York, NY.
Maxwell, J. (1996a). *Problem Identification at Sill Middle School*, Center for Applied Conflict Management, Kent, OH.
Maxwell, J. (1996b). *Possible Solutions to the Discipline Problem at Sill Middle School*, Center for Applied Conflict Management, Kent, OH.
McCool, D. (1989). ''Response to Beryl Radin and Terry L. Cooper,'' *Public Administration Review*, *49*: 170.
McDale, S.E. (1989). *The Identification of Employee Types Through Q-methodology: A Study of Part-time and Seasonal Recreation and Parks Employees*, Unpublished M.A. thesis, Pennsylvania State University.
McGee, F. (1971). ''Comment: Phenomenological Administration—a New Reality,'' *Toward a New Public Administration: The Minnowbrook perspective*, F. Marini (ed.), Chandler Publishing Co., Scranton, PA, pp. 164–171.
McKeown, B. and D. Thomas (1988). *Q Methodology*, Sage Publications, Newbury Park, CA.
McKeown, B. (1990). ''Q Methodology, Communication, and the Behavioral Text,'' *Electronic Journal of Communication*, 1.
McKeown, B. (in press). ''Circles: Preliminaries for a Hermeneutical science,'' *Operant Subjectivity*.
Mesaros, W. and D.L. Balfour (1993). ''Hermeneutics, Scientific Realism, and Social Research: Towards a Unifying Paradigm for Public Administration,'' *Administrative Theory and Praxis*, *15*: 25–37.
Miller, D. and P.H. Friesen (1984). *Organizations: A Quantum View*, Prentice-Hall, Englewood Cliffs, NJ.
Murray, H.A. (1938). *Explorations in Personality*, Oxford University Press, New York, NY.
Nutt, P.C. (1984). ''A Strategic Planning Network for Nonprofit Organizations,'' *Strategic Management Journal*, *5*: 57–75.
O'Reilly, C.A., J.A. Chatman, and D.F. Caldwell (1991). ''People and Organizational Culture: A Profile Comparison Approach to Assessing Person-organization Fit,'' *Academy of Management Journal*, *34*: 487–516.
Oring, K.E. and J. Plihal (1993). ''Using Q-methodology in Program Evaluation: A Case Study of Student Perceptions of Actual and Ideal Dietetics Education,'' *Journal of American Dietetic Association*, *93*: 151–157.
Peat, F.D. (1990). *Einstein's Moon: Bell's Theorem and the Curious Quest for Quantum Reality*, Contemporary Books, Chicago, IL.
Peirce, C.S. (1955). ''The Law of Mind,'' *Philosophical Writings of Peirce*, J. Buchler (ed.), Dover, New York, NY, pp. 339–353.
Radin, B.A. and T.L. Cooper (1989). ''From Public Action to Public Administration: Where Does it Lead?'' *Public Administration Review*, *49*: 167–169.
Rainey, H.G. (1991). *Understanding and Managing Public Organizations*, Jossey-Bass Publishers, San Francisco, CA.
Rhoads, J.C. and T. Sun (1994). ''Studying Authoritarianism: Toward an Alternative Methodology,'' *Southeastern Political Review*, *22*: 159–170.
Roe, E. (1994). *Narrative Policy Analysis*, Duke University Press, Durham, NC.
Rokeach, M. (1973). *The Nature of Human Values*, Free Press, New York, NY.
Rust, J. and S. Golombok (1989). *Modern Psychometrics: The Science of Psychological Assessment*, Routledge, London.
Sasson, T. (1995). *Crime Talk: How Citizens Construct a Social Problem*, Aldine de Gruyter, New York, NY.
Scheb, J.M. (1982). *Merit Selection, Role Orientation and Legal Rationalization: A Q-Technique Study of Florida State District Courts*, Unpublished Ph.D dissertation, University of Florida.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Shah, S. (1982). *Work Orientations of Middle-level Managers from the Public Sector*, Unpublished Ph.D dissertation, Kent State University.
Simon, H.A., D.W. Smithburg, and V.A. Thompson (1950). *Public Administration*, Knopf, New York, NY.
Simon, H.A. (1960). *The New Science of Management Decision*, Harper, New York, NY.
Stephenson, W. (1935). ''Correlating Persons Instead of Tests,'' *Character and Personality*, *4*: 17–24.
Stephenson, W. (1953). *The Study of Behavior: Q-Technique and Its Methodology*, University of Chicago Press, Chicago, IL.
Stephenson, W. (1977). ''Factors as Operant Subjectivity,'' *Operant Subjectivity*, *1*: 3–16.
Stephenson, W. (1978). ''Concourse Theory of Communication,'' *Communication*, *3*: 21–40.
Stephenson, W. (1980). ''Consciring: A General Theory for Subjective Communicability,'' *Communication Yearbook 4*, D. Nimmo (ed.), Transaction, New Brunswick, NJ, pp. 7–36.
Stephenson, W. (1983). ''Quantum Theory and Q-methodology: Fictionalistic and Probabilistic Theories Conjoined,'' *Psychological Record*, *33*: 213–230.
Stephenson, W. (1986). ''Protoconcursus: The Concourse Theory of Communication, *Operant Subjectivity*, *9*: 37–58, 73–96.
Stephenson, W. (1989). ''Quantum Theory of Subjectivity,'' *Integrative Psychiatry*, *6*: 180–195.
Stricklin, M. (1996). PCQ (computer program), Author, Lincoln, NE.
Sun, T. (1992). ''Indigenization of Public Administration Knowledge in Taiwan,'' *Asian Thought and Society*, *17*: 97–112.
Sylvia, R.D. and K.M. Sylvia (1986). ''An Empirical Investigation of the Impacts of Career Plateauing,'' *International Journal of Public Administration 8*: 227–241.
Taylor, P., D.J. Delprato, and J.R. Knapp (1994). ''Q-methodology in the Study of Child Phenomenology,'' *Psychological Record*, *44*: 171–183.
Theiss-Morse, E., A. Fried, J.L. Sullivan, and M. Dietz (1992). ''Mixing Methods: A Multistage Strategy for Studying Patriotism and Citizen Participation,'' *Political Analysis 3*, J. Stimson (ed.), University of Michigan, Ann Arbor, MI, pp. 89–121.
Thomas, D.B. and L. Sigelman (1984). ''Presidential Identification and Policy Leadership: Experimental Evidence on the Reagan Case,'' *Policy Studies Journal*, *12*: 663–675.
Torgerson, D. (1986). ''Between Knowledge and Politics: Three Faces of Policy Analysis,'' *Policy Sciences*, *19*: 33–59.
Vajirakachorn, S. and R.D. Sylvia (1990). ''Administrative Attitudes of Elite Officials in a Buddhist Polity,'' *Operant Subjectivity*, *13*: 163–174.
Van Tubergen, N. and R.A. Olins (1979). ''Mail vs Personal Interview Administration for Q Sorts: A Comparative Study,'' *Operant Subjectivity*, *2*: 51–59.
Vroman, H. (1972). *Types of Administrators and Some of Their Beliefs—a Q Factor Analysis*, Unpublished Ph.D dissertation, University of Iowa.
Wong, H.S. (1973). *Utilizing Group O-technique to Index Leadership Behavior in Task-oriented Small Groups*, Unpublished M.A. thesis, University of Hawaii.
Wood, R.A. (1983). *Perspectives on Prosecutorial Discretion: A Q-methodology Analysis*, Unpublished Ph.D dissertation, University of Missouri-Columbia.
Yarwood, D.L. and D.D. Nimmo (1975). ''Perspectives for Teaching Public Administration,'' *Midwest Review of Public Administration*, *9*: 28–42.
Yarwood, D.L. and D.D. Nimmo (1976). ''Subjective Environments of Bureaucracy: Accuracies and Inaccuracies in Role-taking Among Administrators, Legislators, and Citizens,'' *Western Political Quarterly*, *29*: 337–352.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Appendix 1
Algebra

Sarmistha R. Majumdar
Rutgers University, Newark, New Jersey

Algebra is a useful branch of mathematics mainly concerned with the generalization of arithmetic. It deals not only with numbers but also with letters that may represent a large variety of numbers. Algebra is mainly used in solving equations by using various algebraic techniques. The answers to the equations can be checked by replacing the letters in the equations with the solved values. This makes it possible to balance the equations and help solve any problems that may have been expressed in algebraic terms (Pine, 1992: p. 1-1).

Linear Equations and Graphs

Linear algebra which can only be applied to linear equations enable the expression of a complex system of equations in a simplified form (Dowling, 1992: p. 215). A linear equation is a first degree equation in which each variable is of the first degree, i.e. the variables can be expressed as straight lines on a graph (Hall and Bennett, 1985: p. 84). For example, $2m = 12$ is a linear equation since m has an exponent of 1. But the equation $s = 16$ is not a linear equation as the exponent of s is 2. It is therefore not an equation of the first degree. A linear equation may have more than one variable, for example, $6x + 2 = 5y + 8$ where the two variables are x and y.

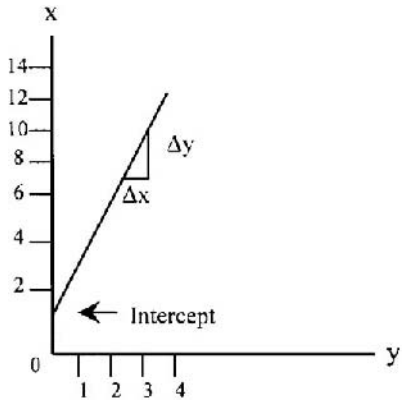
In solving linear equations, like in any other equations, the first step involves simplification of the equation. This requires the elimination of fractions and parentheses. Second, isolation of terms of the specified variable as a single term of the equation. Finally, solving the equation by factoring out the specified variable followed by division of its coefficient (Hall and Bennett, 1985: p. 88).

In a linear equation, the analytical relationship that may exist between the two variables can be expressed in the form of a straight line by using the Cartesian coordinate system (Pine, 1992: pp. 5–18). If there exists two unknown variables x and y, the relationship between the variables in an equation for example, $y = 3x + 1$ can be plotted on the graph by plotting the values of y against x, provided we know or have chosen a numerical value for x.

In the equation $y = 3x + 1$, we choose the values of x as shown below, the corresponding values of y can be easily calculated and plotted on the graph as a straight line.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

x	y	
0	1	(0, 1)
1	4	(1, 4)
2	7	(2, 7)
3	10	(3, 10)
4	13	(4, 13)



In the above graph, the point where the straight line intersects the y axis, is called the y intercept. At this point, the value of $y = 1$ for the line $y = 3x + 1$. The x intercept of the line can be plotted by setting the value of y at zero and solving for x. The steepness of the plotted line is usually called the slope of the line; it tells us of the rate of increase in y with the increase in x. In calculating the slope of a straight line, we usually take into consideration the ratio of change in y as a result of change in x. This ratio of change or the slope of a line passing through the points (x, y) and (x and y) is denoted by Δ (Greek delta) and is calculated as:

$$\Delta = \frac{\text{change in y } (y_2 - y_1)}{\text{change in x } (x_2 - x_1)} = \frac{\Delta y}{\Delta x}$$

Exponentiation and Logarithmic Functions

In algebra, exponentiation and logarithmic functions are often used to express the concept of growth and decay in solving problems in science and social science. The exponent notation which includes both powers and roots is a superscript which tells us how many times to use the base as a multiplier of 1. For example, x^2 means x(x)(1). Since the factor 1 is understood, the most commonly used expression for $x^2 = x(x)$. Thus for any equation $a^x = N$, we can say that x is the exponent, a is the base and N is the number that a to the x equals N (Pine, 1992: p. 7-1).

The exponent in an exponential expression can be any real number. The exponential function deals mainly with a variable rather than a constant. For example, $f(y) = 2^y$ is an exponential function but $f(y) = y^2$ is not an exponential function since the exponent is the constant 2 (Hall and Bennett, 1985: p. 302).

Some of the rules of operation of exponents are as follows:

Ia. $a^2\,(a)^4 = a(a)(a)(a)(a)(a) = a^6$
b. $b^x\,(b^y) = b^{x+y}$
II. $a^x/a^y = a^{x-y}$
$a^4/a^2 = a(a)(a)(a)/a(a) = a^2$
III. $(a^x)^y = a^{xy}$
$(a^3)^2 = a^6$
IV. $a^{-x} = 1/a^x$
$3^{-2} = 1/3^2 = 1/9$
V. $a^0 = 1$
$3^0 = 1$

In graphing exponential functions, all exponential functions $f(x) = b^x$ pass through (0, 1) since $b^0 = 1$. An exponential function where $b > 1$ is an increasing function and it denotes exponential growth function. If $b < 1$, it is a decreasing function and is often used to express exponential decay function (Hall and Bennett, 1985: p. 304).

The logarithmic function is the inverse of exponential function. The inverse of exponential function $f(x) = b^y$ can be written as a logarithmic function of $y = \log_b x$, where $\log_b x$ is the exponent to which b must be raised to get x.

Examples of Inversion:

Exponential Form	Logarithmic Form
$5^2 = 25$	$\log_5 25 = 2$
$6^{-2} = 1/36$	$\log_6(1/36) = -2$
$2^{1/2} = \sqrt{2}$	$\log_2 \sqrt{2} = 1/2$

Usually, the base for a logarithm is any positive number with the exception of 1. The common logarithm of x is $\log_{10} x$ or simply log x. It is the exponent to which 10 must be raised to get x (Dowling, 1980: p. 163). In common logarithm of x, the power to which 10 must be raised to get x is as follows:

$10^1 = 10$	$\log 10 = 1$
$10^2 = 100$	$\log 100 = 2$
$10^3 = 1{,}000$	$\log 1{,}000 = 3$
$10^4 = 10{,}000$	$\log 10{,}000 = 4$
$10^0 = 1$	$\log 1 = 0$
$10^{-1} = 0.1$	$\log 0.1 = -1$
$10^{-2} = 0.01$	$\log 0.01 = -2$

The properties of logarithms help to solve exponential equations and provide a means for simplifying many algebraic expressions. The logarithmic properties for b, x and y positive numbers, n a real number and $b/= 1$ are as follows:

a. $\text{Log}_b xy = \log_b x + \log_b y$

b. $\text{Log}_b x/y = \log_b x - \log_b y$

c. $\text{Log}_b x = n\log_b x$

d. $\text{Log}_b \sqrt{x} = 1/n \log_b x$

Matrix Algebra

A matrix refers to a rectangular array of numbers arranged either in the form of a chart or a table. Thus, matrices provide the means to store information in an orderly and organized way. Capital letters are usually used to denote matrices. In a matrix, the numbers, parameters or variables are referred to as elements. The numbers are usually arranged in horizontal rows and vertical columns. In the matrix the row number always precedes the column number and the number of rows (r) and columns (c) defines the order of the matrix ($r \times c$).

In a square matrix, the number of rows is the same as that of the columns. If a single column dominates a matrix such that its dimensions are $r \times 1$, it is a column vector. Similarly, if a matrix has only a single row with dimensions $1 \times c$, it is a row vector. A transposed matrix is one where the rows of A can be converted to columns and the columns of A to rows. A transposed matrix is denoted by any capital letter, example A′.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Examples:

$$A = \begin{bmatrix} 1_{11} & 2_{12} & 5_{13} \\ 3_{21} & 4_{22} & 0_{23} \end{bmatrix}$$

$r \times c = 2 \times 3$ (general matrix)

$$B = \begin{bmatrix} 4 & 0 & -8 \\ -5 & 2 & 3 \\ 1 & -4 & 7 \end{bmatrix}$$

$r \times c = 3 \times 3$ (square matrix)

$$C = \begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix}$$

$r \times c = 3 \times 1$ (column vector)

$D = [1 \quad 3 \quad 5 \quad 2]$

$r \times c = 1 \times 4$ (row vector)

$C' = [5 \quad 2 \quad 1]$

Transposed matrix of C

In a general matrix A with the dimension of 2×3, the subscript refers to the placement of numbers or elements. The first subscript identifies the row and the second number identifies the column. Thus the subscript in matrix A which is 12 refers to the position of the number which is located in the first row and the second column.

Addition and Subtraction of Matrices

In the addition or subtraction of matrices A and B, the matrices should be of the same order. Each number in addition of matrices (A + B) or in subtraction of matrices (A − B) is obtained by either adding or subtracting the corresponding entries in A and B. If the matrices are not of the same order or dimension, then A + B or A − B cannot be defined.

Examples:

Addition of Matrices

$$A = \begin{bmatrix} 8 & 3 \\ 10 & 4 \end{bmatrix}_{2\times2} + B = \begin{bmatrix} 1 & 2 \\ 5 & 3 \end{bmatrix}_{2\times2} = \begin{bmatrix} 8+1 & 3+2 \\ 10+5 & 4+3 \end{bmatrix}_{2\times2} = \begin{bmatrix} 9 & 5 \\ 15 & 7 \end{bmatrix}_{2\times2}$$

Subtraction of Matrices

$$C = \begin{bmatrix} 5 & 8 \\ 9 & 6 \end{bmatrix}_{2\times2} - D = \begin{bmatrix} 4 & 1 \\ 6 & 5 \end{bmatrix}_{2\times2} = \begin{bmatrix} 5-4 & 8-1 \\ 9-6 & 6-5 \end{bmatrix}_{2\times2} = \begin{bmatrix} 1 & 7 \\ 3 & 1 \end{bmatrix}_{2\times2}$$

Multiplication of Matrices

In matrix algebra, scalar multiplication involves the multiplication of every number in the matrix with any simple number (3, −2, 0.01) or a scalar. The product of scalar multiplication k(A) when $k = 2$:

$$A = \begin{bmatrix} 1 & 3 \\ 9 & 7 \end{bmatrix}_{2\times2} \quad kA = \begin{bmatrix} 2(1) & 2(3) \\ 2(9) & 2(7) \end{bmatrix}_{2\times2} = \begin{bmatrix} 2 & 6 \\ 18 & 14 \end{bmatrix}_{2\times2}$$

Vector multiplication in matrix algebra involves the multiplication of a row vector A by column vector B. One of the prerequisites is that each vector should have the same number of elements so that each element in one vector can be multiplied with the corresponding element in the other vector.

Example:

$$A = \underset{1\times 3}{[1 \quad 2 \quad 5]} \qquad B = \underset{3\times 1}{\begin{bmatrix} 3 \\ 6 \\ 7 \end{bmatrix}} \qquad AB = 1(3) + 2(6) + 5(7) = 3 + 12 + 35 = 50$$

Inverse Matrices

The inverse of a matrix is used to solve equations and to find a curve that best fits the data (Hall and Bennett, 1985: p. 428). Inverse matrix is a square and a nonsingular matrix that satisfies the relationship:

$$AA^{-1} = A^{-1}A$$

An inversed matrix can be obtained by applying the formula:

$$A^{-1} = \frac{1}{|A|} \text{Adj } A$$

Thus in linear algebra, inverse matrix performs the same function as the reciprocal in ordinary algebra.

In the inversion of the given matrix A, the following steps are important.

$$A = \begin{bmatrix} 5_{a_{11}} & 3_{a_{12}} & 6_{a_{13}} \\ 4_{a_{21}} & 2_{a_{22}} & 7_{a_{23}} \\ 8_{a_{31}} & 5_{a_{32}} & 3_{a_{33}} \end{bmatrix}$$

$$A^{-1} = 1/|A| \quad \text{Adj } A$$

(a) Evaluation of the determinant by taking the first element of the first row, i.e. a_{11}, and mentally deleting the row and column in which it appears. Then multiplying a_{11} by the determinant of the remaining elements (Dowling, 1992: p. 244). For example:

$$|A| = a_{11}\begin{bmatrix} 2_{a_{32}} & 7_{a_{23}} \\ 5_{a_{32}} & 3_{a_{33}} \end{bmatrix} + a_{12}{}^{(-1)}\begin{bmatrix} 4_{a_{21}} & 7_{a_{23}} \\ 8_{a_{31}} & 3_{a_{33}} \end{bmatrix} + a_{13}\begin{bmatrix} 4_{a_{21}} & 2_{a_{22}} \\ 8_{a_{31}} & 5_{a_{32}} \end{bmatrix}$$

$$= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

$$= 5(6 - 35) - 3(12 - 56) + 6(20 - 16)$$

$$= -145 + 132 + 24 = 11$$

(b) Finding the cofactor matrix by replacing every element a_{ij} in matrix A with its cofactor $|C_{ij}|$. Then finding the adjoint matrix which is a transpose of the cofactor matrix. Thus the cofactor matrix is

$$C = \begin{bmatrix} \begin{vmatrix} 2 & 7 \\ 5 & 3 \end{vmatrix} & -\begin{vmatrix} 4 & 7 \\ 8 & 3 \end{vmatrix} & \begin{vmatrix} 4 & 2 \\ 8 & 5 \end{vmatrix} \\ -\begin{vmatrix} 3 & 6 \\ 5 & 3 \end{vmatrix} & \begin{vmatrix} 5 & 6 \\ 8 & 3 \end{vmatrix} & -\begin{vmatrix} 5 & 3 \\ 8 & 5 \end{vmatrix} \\ \begin{vmatrix} 3 & 6 \\ 2 & 7 \end{vmatrix} & -\begin{vmatrix} 5 & 6 \\ 4 & 7 \end{vmatrix} & \begin{vmatrix} 5 & 3 \\ 4 & 2 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} -29 & 44 & 4 \\ 21 & -33 & -1 \\ 9 & -11 & -2 \end{bmatrix}$$

$$\text{and Adj A} = C' = \begin{bmatrix} -29 & 21 & 9 \\ 44 & -33 & -11 \\ 4 & -1 & -2 \end{bmatrix}$$

$$\text{Thus } A^{-1} = \frac{1}{11}\begin{bmatrix} -29 & 21 & 9 \\ 44 & -33 & -11 \\ -4 & -1 & -2 \end{bmatrix} = \begin{bmatrix} -29/11 & 21/11 & 9/11 \\ 44/11 & -33/11 & -11/11 \\ -4/11 & -1/11 & -2/11 \end{bmatrix}$$

Linear Equations in Matrix Algebra

Matrix algebra enables the expression of linear equations in simple and concise forms. Example of Linear Equation:

$3x_1 + 2x_2 = 10$

$5x_1 + 4x_2 = 27$

Matrix form = AX = B

where $A = \begin{bmatrix} 3 & 2 \\ 5 & 4 \end{bmatrix}$ (coefficient matrix) $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ (solution vector) $= B = \begin{bmatrix} 10 \\ 27 \end{bmatrix}$ (vector of constant term)

Matrix Inverstion in Linear Equations

$3x_1 + 2x_2 = 14$

$4x_1 + 5x_2 = 28$

$X = A^{-1}B$

$$\begin{bmatrix} 3 & 2 \\ 5 & 4 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 36 \end{bmatrix}$$

Inverse of A where $|A| = 3(5) - 2(4) = 7$. The cofactor matrix of A is:

$$C\begin{bmatrix} 5 & -4 \\ -2 & 3 \end{bmatrix}$$

$$\text{and Adj A} = C' = \begin{bmatrix} 5 & -2 \\ -4 & 3 \end{bmatrix}$$

$$\text{Thus } A^{-1} = 1/7 \quad \begin{bmatrix} 5 & -2 \\ -4 & 3 \end{bmatrix} = \begin{bmatrix} 5/27 & -2/7 \\ -4/7 & 3/7 \end{bmatrix}$$

Then substituting in $X = A^{-1}B$ and simply multiplying matrices

$$X = \begin{bmatrix} 5/27 & -2/7 \\ -4/7 & 3/7 \end{bmatrix}\begin{bmatrix} 14 \\ 28 \end{bmatrix} = \begin{bmatrix} 10 & -8 \\ -8 & 12 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

Thus $x_1 = 2$ and $x_2 = 4$.

BIBLIOGRAPHY

Dowling, E.T. (1992). *Introduction to Mathematical Economics*, New York: McGraw Hill Inc.
Hall, J.W., and R.D. Bennett. (1985). *College Algebra with Application*, Boston: Prindle, Weber and Schmidt.
Pine, C. (1992). *The Algebra Project*, Newark, New Jersey: Rutgers.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Appendix 2
Distribution of t

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Appendix

T-TABLE Distribution of t

	Level of Significance for One-Tailed Test					
	.10	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test					
df	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.984
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291
	.80	.90	.95	.98	.99	.999

Source: Table III of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, Addison Wesley Longman Ltd., London (1974), 6th edition (previously published by Oliver & Boyd Ltd., Edinburgh). By permission of the authors and publishers.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Index

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016