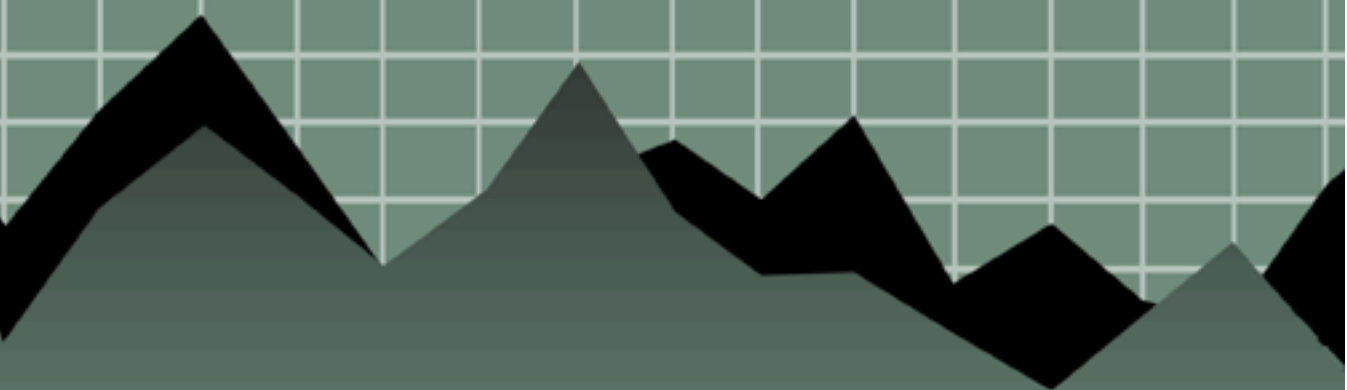
Evaluation in Practice

A Methodological Approach

edited by

Richard D. Bingham
Cleveland State University

Claire L. Felbinger
American University



SECOND EDITION

Evaluation in Practice

A Methodological Approach

SECOND EDITION

Richard D. Bingham
Cleveland State University

Claire L. Felbinger
American University

CHATHAM HOUSE PUBLISHERS
SEVEN BRIDGES PRESS, LLC
NEW YORK · LONDON

Seven Bridges Press
135 Fifth Avenue
New York, NY 10010-7101

Publisher: Ted Bolen
Managing Editor: Katharine Miller
Composition: Rachel Hegarty
Cover design: Rachel Hegarty
Printing and Binding: Bang Printing

Library of Congress Cataloging-in-Publication Data

Bingham, Richard D.
Evaluation in practice : a methodological approach / Richard D. Bingham, Claire L. Felbinger. — 2nd. ed.
p. cm.
Includes bibliographical references.
ISBN 1-889119-57-1
1. Evaluation research (Social action programs)—United States. 2. Social sciences—Research—Methodology. I. Felbinger, Claire L. II. Title.
H62.5 U5 B56 2002
300'.7'2—dc21 2001006312
CIP

Manufactured in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

We would like to thank the following individuals, organizations, and publishers who have kindly granted permission to include the following materials in this book:

"Improving Cognitive Ability in Chronically Deprived Children," by Harrison McKay, Leonardo Sinisterra, Arline McKay, Hernando Gomez, Pascuala Lloreda, reprinted with permission from *Science* 200, no. 4339 (21 April 1978): 270–78. Copyright 1978 by American Association for the Advancement of Science.

"Evaluation of a Multimethod Undergraduate Management Skills Development Program," by Marian M. Extejt et al., in *Journal of Education for Business* (March/April 1996): 223–31. Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 1319 18th St. N.W., Washington, D.C. 20036-1802. Copyright 1996.

"Community Posthospital Follow-up Services," by Ann Solberg, in *Evaluation Review: A Journal of Applied Social Research* 7, no. 1 (February 1983): 96–109, copyright © 1983 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"Conservation Program Evaluations: The Control of Self-Selection Bias," by Tim M. Newcomb, in *Evaluation Review: A Journal of Applied Social Research* 8, no. 3 (June 1984): 425–40, copyright © 1984 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"Regulatory Strategies for Workplace Injury Reduction: A Program Evaluation," by Garrett E. Moran, in *Evaluation Review: A Journal of Applied Social Research* 9, no. 1 (February 1985): 21–33, copyright © 1985 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"The Effects of Early Education on Children's Competence in Elementary School," by Martha B. Bronson, Donald E. Pierson, and Terrence Tivnan, in *Evaluation Review: A Journal of Applied Social Research* 8, no. 5 (October 1984): 615–29, copyright © 1984 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"Nutrition Behavior Change: Outcomes of an Educational Approach," by Patricia K. Edwards, Alan C. Acock, and Robert L. Johnson, in *Evaluation Review: A Journal of Applied Social Research* 9, no. 4 (August 1985): 441–59, copyright © 1985 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"A Little Pregnant: The Impact of Rent Control in San Francisco," by Edward G. Goetz, in *Urban Affairs* 30, no. 4 (March 1995): 604–

12, copyright © 1995 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"The Costs and Benefits of Title XX and Title XIX Family Planning Services in Texas," by David Malitz, in *Evaluation Review: A Journal of Applied Social Research* 8, no. 4 (August 1984): 519–36, copyright © 1984 by Sage Publications, Inc. Reprinted by permission of Sage Publications, Inc.

"A Cost-Effectiveness Analysis of Three Staffing Models for the Delivery of Low-Risk Prenatal Care," by Elaine A. Graveley and John H. Littlefield, in *American Journal of Public Health* 82, no. 2 (February 1992): 180–84. Reprinted with permission from the American Public Health Association.

"Attitude Change and Mental Hospital Experience," by Jack M. Hicks and Fred E. Spaner, in *Journal of Abnormal and Social Psychology* 65, no. 2 (1962): 112–20.

"How Effective Is Drug Abuse Resistance Education? A Meta-Analysis of Project DARE Outcome Evaluations," by Susan T. Ennett, Nancy S. Tobler, Christopher L. Ringwalt, and Robert Flewelling, in *American Journal of Public Health* 84, no. 9 (September 1994): 1394–1401. Reprinted with permission from the American Public Health Association.

"Are Training Subsidies for Firms Effective? The Michigan Experience," by Harry J. Holzer, Richard N. Block, Marcus Cheatham, and Jack H. Knott, in *Industrial and Labor Relations Review* 46, no. 4 (July 1993): 625–36. Reprinted with permission from Cornell University.

"Changing the Geography of Opportunity by Expanding Residential Choice: Lessons from the Gautreaux Program," by James E. Rosenbaum, in *Housing Policy Debate* 6, no. 1 (1995): 231–69. Copyright 1995 by Fannie Mae Foundation. Reprinted with permission from the Fannie Mae Foundation.

"Changes in Alcohol Consumption Resulting from the Elimination of Retail Wine Monopolies," by Alexander C. Wagenaar and Harold D. Holder, reprinted with permission from *Journal of Studies on Alcohol* 56, no. 5 (September 1995): 566–72. Copyright by Alcohol Research Documentation, Inc., Rutgers Center of Alcohol Studies, Piscataway, NJ 08854.

"Fifth-Year Report: Milwaukee Parental Choice Program," by John F. Witte, Troy D. Sterr, and Christopher A. Thorn, typescript (December 1995). Reprinted with permission from John F. Witte.

"The Effectiveness of School Choice in Milwaukee," by Jay P. Greene, Paul E. Peterson, and Jiangtao Du, in *Learning from School Choice*, edited by Paul E. Peterson and Bryan C. Hassel (Brookings Institution Press, 1998). Reprinted with permission from the Brookings Institution.

Preface

In teaching program and policy evaluation courses in various graduate public administration programs, we have found that students have serious difficulty in critiquing evaluations done by others. Many students have problems distinguishing one type of design from another. If the design used by a researcher is not a clone of one of the designs found in the textbook, many students are not able to classify or categorize it.

In addition, students often believe that merely because an article was published in a scholarly or professional journal, it is perfect. They fail to see how an author did or did not control for various methodological problems.

We conceive of an evaluation course as a methods course in which students learn the techniques of program and policy evaluation and the methodological problems encountered when conducting evaluation research. Part of the learning process involves the ability to assess the evaluations of others—not solely to find fault with the work of others but to develop a keen understanding of the difficulties of conducting evaluations. Peter Rossi and Katharine Lyall make this point clearly:

> It is easy enough to be a critic: All pieces of empirical research are more or less flawed. There are simply too many points at which mistakes can be made or unforseen events intrude to permit perfection in either design or execution. We believe that critics have a responsibility to make two kinds of judgments about criticisms they offer: First it is necessary to make distinctions, if possible, between those mistakes that constitute serious flaws and those that are less serious defects. Admittedly, this is a matter of judgment, yet we do believe that some sort of weighting ought to be suggested by responsible critics as a guide to those who have not digested the enormous volume of memoranda, working papers, analyses, and final reports produced by the experiment. Second, it is only fair to distinguish between defects that arise from incorrect planning and other errors of judgment and those that arise out of events or processes that could not have been anticipated in advance. This is essentially a distinction between "bad judgment" and "bad luck," the former being a legitimate criticism and the latter calling for sympathetic commiseration (1978, 412–13).

We would like to take Rossi and Lyall's comments one step further. According to the dictionary, a critic is "one who expresses a rea-

soned opinion on any matter as a work of art or a course of conduct, involving a judgment of its value, truth, or righteousness, an appreciation of its beauty or technique, or an interpretation" (*Webster's* 1950, 627).

Thus, to be a critic requires positive judgments as well as negative. We consider it important to point to cases in which researchers have developed interesting approaches to overcome problems of difficult evaluation design. Chapter 8 of this book is illustrative. Tim Newcomb devised a unique comparison group in his evaluation of an energy conservation program in Seattle. As "critics," we are delighted when we run across this kind of creative thinking.

It should be clear by now that the purpose of *Evaluation in Practice* is to illustrate the techniques of different research designs and the major design problems (termed *problems of internal validity*) encountered in real-world evaluation research. Each chapter, 5 through 16, presents a brief introduction to an evaluation design and uses an evaluation as an example of that design. The article is then followed by an explanation and critique of the work.

One caveat: The book does not cover all forms of evaluation monitoring, assessment, process evaluations, and the like (with the exception of brief discussions in chapter 1). We are concerned with impact evaluations. This does not suggest that these functions are not important, only that the purpose of this book is much more modest.

Evaluation in Practice is written for students in graduate professional degree programs and for general graduate courses in the social sciences. Although we both teach in graduate public administration programs, this book is not that narrowly construed. The articles in the book are taken from a variety of disciplines illustrating the commonality of evaluation. The book provides examples of both policy and program evaluations from a multitude of disciplines. What is important is the methodology and not the substantive fields represented by the articles. The textual material in the chapters is somewhat limited as our concern is with practical applications of the various designs. A list of supplementary readings follows the introductions to many of the chapters for those interested in more detailed explanations of the design being discussed. The book has equal applicability in public policy courses, public administration, planning, urban studies, sociology, political science, criminal justice, education, social welfare, and the health professions, to name a few.

The book makes no assumptions about the reader's background beyond an undergraduate degree. A number of the articles use statistical techniques, but it is not necessary that the student be well versed in statistics. Unless we specify otherwise, the student can assume that the statistics are appropriately applied. It is the design methodology that is important in *Evaluation in Practice*, not statistics.

Organization of the Book

The book is composed of eight parts: introduction, experimental designs, quasi-experimental designs, reflexive designs, cost-benefit and cost-effectiveness analyses, other designs, on your own, and dilemmas of evaluation. Part I consists of four chapters. The introductory chapter addresses process evaluations, output evaluations, and outcome evaluations. The chapter is based on the premise that evaluation is a continuous process in which different types of evaluation of a particular program are appropriate at different times and for different audiences. The second chapter introduces four basic evaluation designs and discusses threats to validity. The third chapter presents the basics of measurement

and discusses the reliability and validity of measurement. The last chapter in the section is concerned with two particular types of measurement in vogue today—performance measurement and benchmarking.

Part II consists of three chapters discussing and illustrating three experimental evaluation designs. The feature common to each is the random assignment of subjects into experimental and control groups. The chapters cover the pretest-posttest control group design, the Solomon Four-Group Design, and the posttest-only control group design.

Part III covers quasi-experimental evaluation designs and is composed of four chapters. They discuss the pretest-posttest comparison (not control) group design, regression-discontinuity design, interrupted time-series comparison group design, and the posttest-only comparison group design.

Part IV also concerns quasi-experimental designs but is differentiated from the designs in part III in that the part IV groups are compared only with each other. The two types of evaluation design illustrated in this section are the one-group pretest-posttest design and the simple time-series design.

Part V covers two variations, or expansions, of other designs, which are included for their uniqueness. These are cost-benefit and cost-effectiveness analyses.

Part VI also covers unique variations on basic designs. They are patched designs and meta-evaluation designs.

Part VII requires students to work on their own. The three chapters in this section present three interesting policy and program evaluations with no explanation and critique. It is up to the students to evaluate the articles on their own and to apply what they have learned thus far.

The book concludes with three articles in Part VIII. This section illustrates the dilemmas of evaluation as it presents two articles that use the same data set to evaluate a parental choice program in education but come to differing conclusions. Our colleague Chieh-Chen Bowen provides the critique.

Finally, we would like to thank a number of people for their assistance on this book. First, we thank the Urban Center at Cleveland State University for institutional support. But we especially appreciate the comments and guidance we received from a number of our colleagues—Joe Wholey of the University of Southern California, Roger Durand of the University of Houston at Clear Lake, Patricia Shields of Southwest Texas State University, and especially Paul Culhane of Northern Illinois University. Their comments on our proposal for this second edition were absolutely critical to the book. Also thanks to American University students Jane Chan and Robin Gluck for assistance with manuscript production. Finally, we appreciate the assistance of the editorial and publishing services of David Estrin and the help of Bob Gormley, Katharine Miller, and Ted Bolen of Chatham House Publishers. Of course, we retain the ultimate responsibility.

We have high hopes for this second edition. We are confident that if students thoroughly understand the readings and discussion presented here, they will be capable of rationally reviewing most of the program or policy evaluations or evaluation designs that they are likely to encounter during their professional careers.

References

Rossi, Peter H., and Katharine C. Lyall. 1978. "An Overview Evaluation of the NIT Experiment." In *Evaluation Studies Review Annual*, vol. 3. Beverly Hills, Calif.: Sage Publications, 1950.

Webster's New International Dictionary. 2d ed. Springfield, Mass.: Merriam.

PART I

Introduction

CHAPTER 1

The Process of Evaluation

The evaluation of agency programs or legislative policy is the use of scientific methods to estimate the successful implementation and resultant outcomes of programs or policies for decision-making purposes. Implicit in this definition are the many levels on which a program or policy can be evaluated and the many potential audiences that may be interested in utilizing the evaluation. A single approach or method is not common to all evaluations. We hope that during the course of this book, students of evaluation will become acquainted with the multiplicity of approaches to evaluation and develop the ability to choose appropriate designs to maximize internal validity and meet consumers' evaluation needs.

Good evaluations use scientific methods. These methods involve the systematic process of gathering *empirical* data to test *hypotheses* indicated by program's or policy's *intent. Empirical data* are observable, measurable units of information. In evaluation, one does not just "feel" that a program is operating effectively; the data demonstrate that this is or is not the case. *Hypotheses* are assertions about program impacts. In other words, hypotheses state changes that should occur to program recipients as a direct result of the program. For example, children in a nutrition program should be healthier after the program than they were before. These hypotheses should be linked with the policy's or program's intent. Typically, *intent* refers to policymakers' hopes regarding the program's outcomes. Unfortunately, identifying intent is not always a simple process. Bureaucratic agencies attempt to translate legislative intent into procedures and regulations. When the intent is unclear or contradictory, the translation can be manipulated into a program that does not resemble anything the legislators had in mind. Even with clear intent, poor program design can obscure original intent. In the best of all worlds, evaluators look for broad-based goals and for clearly stated measurable objectives by which these goals can be reached. Nevertheless, it is the job of the evaluator to try to break through these often unclear guidelines and to provide usable findings in a timely manner in the hopes of informing the decision-making process.

Types of Evaluation

Several general types of evaluation correspond to what some evaluators consider to be successive hierarchical levels of abstraction.

One type of evaluation is not "better" than another; each is appropriate to a different set of research questions. Timothy Bartik and Richard Bingham refer to this as a continuum of evaluations (1997, 247). This continuum is illustrated in figure 1.1. Evaluation is a science and an art. The artful part is the successful matching of the level of the evaluation and the appropriate scientific design within the resource and time constraints of the consumer of the evaluation.

Process Evaluations

The first sentence of this chapter mentions measuring the *implementation* of programs. *Implementation* is the process by which a program or policy is designed and operated. *Process evaluations* focus on the means by which a program or policy is delivered to clients. Karen Sue Trisko and V.C. League identify five levels, or approaches, to evaluation (which Bartik and Bingham adapt to their continuum), two of which are process or, as they also refer to them, formative evaluations. Although Trisko and League refer to these as levels, the term *approach* seems to be more appropriate in the evaluation context. The first process approach is monitoring daily tasks. In this approach to evaluation, fundamental questions of program operation are the focus of inquiry. Indeed, questions such as "Are contractual obligations being met?" or "Are staff adequately trained for their jobs?" are pursued at this level. Basically, these evaluations look to uncover management problems or assure that none are occurring. They deal with the behavior and practices of the program's staff. Work analysis, resource expenditure studies, management audits, procedural overhauls, and financial audits are indicative of evaluations at this point in the continuum.

These evaluations often involve an inspection of the fundamental goals and objectives of the program. Indeed, evaluations cannot occur in the absence of direction concerning what the program is supposed to do. It is shocking how many programs operate in the absence of written goals and objectives! Sometimes policymakers cannot satisfy all relevant political players unless they are not specific about goals and objectives. Even when such directives exist, evaluators often find that the actual operation does not seem to fit the intent of the written guidelines. This may be because the goals are outdated, at which point staff need to reevaluate the goals and objectives in light of the current environment. Sometimes programs develop a life of their own, and evaluators can point out ways in which the program can get back on track.

Even if the purpose of the evaluation is not to assess process, evaluators find it necessary to gather an inventory of goals and

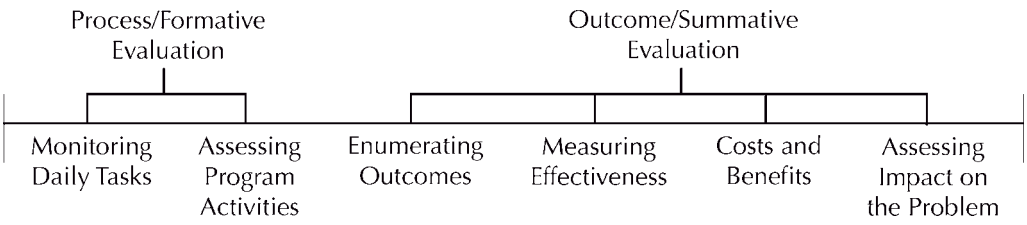


FIGURE 1.1
Continuum of Evaluation

Source: Bartik and Bingham 1997, 248.

objectives to determine the predicted impact of the program. They then must reconcile the written material with what they observe before a full-blown evaluation can occur. Joseph Wholey and his colleagues at the Urban Institute (Horst et al. 1974) termed the process by which this is done "evaluability assessment." *Evaluability assessments* are process evaluations that are performed so that the evaluator and the evaluation client can come to agreement on which aspects of the program will be part of the final product and what resources are necessary to produce the desired document. The benefits of process evaluations should not be underestimated.

The second approach to process evaluation concerns assessing program activities and client satisfaction with services. This approach is concerned with what is happening to program participants. Among the questions considered at this level are "What is done to whom and what activities are actually taking place?" or "How could it be done more efficiently?" or "Are the clients satisfied with the service or image of the service?"

Both the first and second aspects of process evaluations involve subjective measures at times and also require staff and client involvement to complete. Some researchers, such as Donald Campbell and Julian Stanley (1963, 6–12), refer to process evaluations as "pre-experiments." But, once again, the value of process evaluations should not be understated. It makes little sense to attempt to assess the impact of a program if it is run incorrectly or if the consumer being served is not known. The results of these evaluations can be as basic as pointing out to an agency that it does not have evaluable goals or objectives or informing it that its accounting procedures are shoddy. If one ever wonders why it seems so difficult to evaluate the activities of federal bureaucrats, one need only look to the enabling legislation of many of our major national programs. Putting together Congressional majorities in controversial legislation often leads to murky legislative intent.

Process evaluations frequently focus on the way a program is implemented. Program managers may be interested professionally in organizing and running a fine-tuned bureau and in periodically assessing the efficiency of operations. During the course of these evaluations, better ways of doing business may be discovered in addition to blatant inefficiencies. For instance, staff may suggest alternate ways of organizing the service production, or they may point out innovative techniques or tools discovered in their own professional development. Those who fund the activities are also concerned about efficiency (minimizing waste)—although sometimes to the detriment of effectiveness (obtaining the desired effect). During times of fiscal austerity, evaluations of this type can be beneficial. Unfortunately, evaluations are often the first activities to be cut during budget slashing. In order to understand how important process evaluations are as precursors to impact evaluations, see case 1.1; it is one of our favorite evaluation stories as told by Michael Patton.

Impact Evaluations

The next two approaches are referred to as *impact, outcome,* or *summative evaluations.* Impact evaluations focus on the end results of programs. The first impact evaluation, enumerating outcomes, looks at whether the program's or policy's objectives have been met. Here we are interested in quantifying what is happening to program participants. Questions may be "What is the result of the activities conducted by the program?" or "What happened to the target population because of those activities [was it the expected outcome]?" or "Should different activities be

CASE STUDY

Case 1.1 Preimplementation Evaluation Program

A state legislature established a demonstration program to teach welfare recipients the basic rudiments of parenting and household management. The state welfare department was charged with the responsibility for conducting workshops, distributing brochures, showing films, and training case workers on how low-income people could better manage their meager resources and how they could become better parents. A single major city was selected for pilot testing the program, and a highly respected independent research institute was contracted to evaluate the program. Both the state legislature and the state welfare department were publicly committed to using the evaluation findings for decision making.

The evaluators selected a sample of welfare recipients to interview before the program began. They collected considerable data about parenting, household management, and budgetary practices. Eighteen months later, the same welfare recipients were interviewed a second time. The results showed no measurable change in parenting or household management behavior. In brief, the program was found to be ineffective. These results were reported to the state legislators, some of whom found it appropriate to make sure the public learned about their accountability efforts through the newspapers. As a result of this adverse publicity, the legislature terminated funding for the program—a clear instance of utilization of evaluation findings for decision making.

Now, suppose we wanted to know why the program was ineffective. That question could not be answered by the evaluation as conducted because it focused entirely upon measuring the attainment of the program outcomes, that is, the extent to which the program was effective in changing the parenting and household management behavior of welfare recipients. As it turned out, there is a very good reason why the program was ineffective. When the funds were initially allocated from the state to the city, the program became immediately embroiled in the politics of urban welfare. Welfare organizations questioned the right of government to tell poor people how to spend their money or rear their children: "You have no right to tell us to manage our households according to white, middle-class values. And who is this Frenchman named Piaget who's going to tell us how to raise our American kids?"

As a result of these and other political battles, the program was delayed and further delayed. Procrastination being the better part of valor, the first parenting brochure was never printed, no household management films were ever shown, no workshops were held, and no case workers were ever trained. In short, *the program was never implemented—but it was evaluated!* It was then found to be ineffective and was killed.

The lesson: Be sure the program exists before you evaluate it!

Source: Patton 1978, 149–50.

substituted?" In other words, are the objectives of the program being met?

Traditionally, when people think of program evaluation, impact evaluations are what they usually have in mind. Impact evaluations are easy to conceptualize because they revolve around directly assessing outputs. How many output units were delivered? How many free lunches were served, applications processed, client hours logged, workers trained? Because these evaluations are easily conceptualized, designing them tends to be straightforward. In addition, because "impacts" are easily understood, these types of evaluations are more easily justified and fundable than are process evaluations.

The second type of impact evaluation involves measuring effectiveness, which concerns either "Was the program cost effective?" or "What would have happened to the target population in the absence of the program?" In contrast to the previous approach, one looks at whether and to what extent the goals of the program or policy are being met. This can be either a substantive analysis of effectiveness or can involve factoring in the costs of the program relative to its benefits. These evaluations focus on measuring outcomes. Impact evaluations tend to be more objective because it is not necessary to rely solely on clients or staff to gather data (although their assistance is often helpful). The data can be extracted from records (if process evaluations are favorable) or from observing or testing or measuring phenomena.

The evaluation methods covered in this book are generally of the impact variety. We say "generally" because both of the authors believe that scientific methodology and rigor can be used, though in varying degrees, in any type of evaluation. However, impact evaluations lend themselves quite easily to empirical investigation. The methods described here assume this empirical dimension.

Policy Evaluations

The final kind of evaluation considers the long-term consequences of a program or policy—assessing the impact on the problem. The kinds of questions asked here include "What changes are evident in the problem?" or "Has the problem (e.g., poverty, illiteracy) been reduced as a result of the program or policy?" These evaluations are difficult to assess empirically. Presumably there is a national policy concerning the eradication of poverty and hunger. How does one assess the consequence of programs that are delivered simultaneously and sometimes at cross-purposes to a target population? How does one measure "poverty"? In the true sense of the word, this type of evaluation can be called *policy evaluation.* David Nachmias and Claire Felbinger (1982, 300–308) describe this type of evaluation as one that occurs when a political body considers changing or continuing a major social or political program—when prevailing policy is challenged in the agenda-setting forum. In this situation the role of the evaluator is rather vague. So many variables can have an impact on the long-term process concerning such broad social issues that it is difficult to assess impact. Many decisions regarding the relative "worth" of differential policies are best evaluated in a political arena.

This does not mean that "policies" as such cannot be empirically evaluated. The example in chapter 19 evaluates a public policy issue—alcohol consumption. At issue is whether consumption patterns differ under conditions of state-owned versus privatized liquor stores.

Meta-Evaluations

Another evaluation approach attempts to make sense out of accumulated evaluation findings. *Meta-evaluations* are syntheses of evaluation research findings. They look for

commonalities among results, measures, and trends in the literature. They reuse the extant research findings.

Meta-evaluations are quite similar to literature reviews. In science, one is interested in the cumulative nature of research. Literature reviews attempt to make sense out of the research that precedes the current effort. Literature reviews are considered qualitative exercises. Equally equipped researchers can disagree on the interpretation of research results. Persons involved in meta-evaluations try to quantify the review, presumably making a more objective statement. Harris Cooper (1982) argues that meta-analytic procedures can be systematic and scientific. A number of quantitative methods have been developed, most notably by Gene Glass (1976; 1978) and Richard Light (1979; see also Light and Smith 1971). Often evaluations focus on whether the average, or mean, measure of the evaluation criterion variable for the group that participated in the program is any different from a similar group that did not participate. Glass and his associates found conceptually similar criterion, or dependent, variables across studies and developed what they call the "effect size" between the means of experimental and control groups, standardizing on the basis of standard deviation. They estimate the size of the total effect of these programs by aggregating across studies. Light, in contrast, uses the original data, as opposed to group means; pools the findings in different sites; and reanalyzes the data. He includes only those measures that are identical across studies. This technique reduces the number of studies he can aggregate. However, the findings are not subject to measurement differences. An example of a meta-evaluation is in chapter 16. The idea of systematically aggregating evaluation results has received considerable research attention. William Bowen and Chieh-Chen Bowen (1998, 72) identify seven steps in meta-evaluations:

- Conceptualize the relationship under consideration;
- Gather a set of studies that have tested the specified relationship;
- Design a coding sheet to record characteristics of the condition under which each study was conducted;
- Examine each study and, using the coding sheet, record the conditions under which it was conducted;
- Compute the "effect size" for each study;
- Statistically analyze the characteristics and effect sizes for all the studies;
- Write a research report.

Utilization of Evaluation Findings

Regardless of what type of evaluation is performed, it is not useful unless it is completed in time for decision makers to use the findings as input for their decision-making process. It makes little sense to perform an elegantly designed evaluation the results of which arrive a day after the decision on the program's fate is cast. This is one difference between evaluation research and academic research. Academic researchers will usually risk a timely report for the elegance of a properly constructed, controlled design. Evaluation researchers hope to construct the "best" design, but they err on the side of expedience when a decision deadline is near. That does not mean that evaluation research is necessarily slipshod. Rather, evaluation research is not worth much if it is not utilized. Evaluators must recognize the trade-offs made in the interest of providing material in a timely manner.

How can one determine whether the results of an evaluation are utilized? In the 1970s, evaluation researchers spent a great

deal of time trying to trace whether specific evaluations were utilized by decision makers. The first problem they encountered was that they could not agree on the definition of utilization. When restrictive definitions were proposed, the record of utilization was bleak. Less restrictive definitions led to more pleasing results. Evaluators, however, were frustrated with the degree to which they felt their efforts were being used to affect policy.

Nachmias and Felbinger (1982) recast the utilization question and put utilization into a broader perspective. For them, utilization need not be a discreet action. Policymakers do not operate in an information vacuum; they tend to bring to the decision-making process the variety of their life experiences. Indeed, when the results of an evaluation suggest some action that, to them, is counterintuitive or politically inexpedient, utilization is not immediately assured. As the volume of evidence grows, however, the information gleaned at an earlier time may be brought to bear in the context of a future decision. This type of utilization is virtually impossible to document.

Nachmias and Felbinger suggest that the process by which information is utilized can be conceptualized by viewing the decision process as part of a cycle of life events that shape the decision makers' perspectives. The diagram they use to explain this process is shown in figure 1.2.

Figure 1.2 is instructive in this context because it displays the various points at which decisions (hence, utilization) are made and the various approaches and clients for evaluation information of a single program. The feedback loop at point 1 stands for ongoing, short-term evaluation of the implementation of a program—process evaluations. This ongoing process is intended to inform program managers of the success of basic program operations. The constant feedback allows them to fine-tune the existing program. Generally, then, evaluations at this point are short-term in nature, process oriented, and aimed at the needs of program managers (the consumers and utilizers of the evaluation). Clearly, costly experimental designs are inappropriate for these kinds of evaluations.

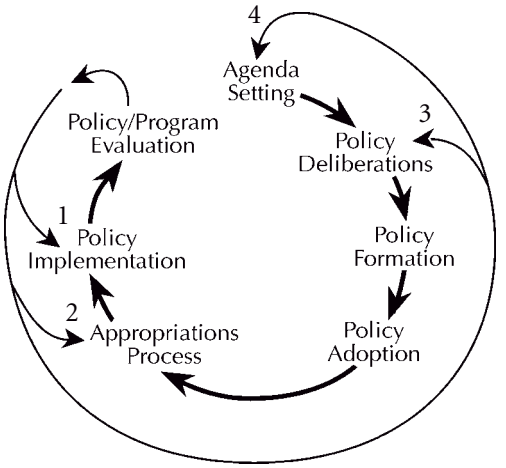


FIGURE 1.2
Evaluation Utilization Subcycles in the Legislative-Budgetary Cycle

Source: Nachmias and Felbinger 1982, 305.

Evaluations at point 2 are of the impact variety. They usually allow for a longer-term evaluation. The designs described in this book are certainly appropriate for this kind of evaluation. The clients for these evaluations are decision makers who are interested in program oversight. In the Nachmias and Felbinger example, these consumers are members of Congress engaged in annual appropriation hearings. However, these consumers can be any interested publics making decisions on the utility of programs. These decision makers are not constrained by the information presented in the specific impact evaluations; they are free to use the portions they think are relevant and politically feasible.

Point 3 evaluations are concerned with the outcomes of programs—policy evaluations. When programs are highly controversial or when new programs are competing for scarce dollars, policy evaluations affect whether the general policy should continue. When prevailing policies are being questioned, fundamental questions such as "Should the government really be involved in this?" or "Does this really matter?" come to fore. The evaluations of the 1970s income maintenance experiments are of this sort, as instituting the proposals would have constituted a major shift in national welfare policy. These evaluations are typically of a long-term variety and command the most costly evaluations. Again, decision makers at this point may utilize the information from the evaluation—in the case of income maintenance they did. They cannot help, however, allowing their feelings to be shaped by ideological concerns and information from other sources.

The feedback loop at point 4 is illustrative of how evaluations of the past affect decision making in the present and can be considered a personal meta-evaluation exercise. Nachmias and Felbinger suggest that utilization research cast at this point would require longitudinal case studies. The loop is descriptive; the research suggested seems unmanageable. Nachmias and Felbinger hoped to end attempts to evaluate the utilization of specific results and to measure the behavior of decision makers. Evaluators were still concerned, however, about the value decision makers place on research results. Michael Patton and his colleagues (1978) suggest that evaluators should concern themselves with something they can control—the conduct of evaluations. They suggest that evaluators should be concerned with designing utilization-focused evaluations. William Tash and Gerald Stahler (1982, 180–89) followed a step-by-step method to enhance the utilization of their research results:

1. Identify specific user audiences and tailor recommendations to these groups.
2. Formulate recommendations in cooperation with the user audience.
3. Direct specific recommendations toward a broader policy and program model.
4. Assess the impact of recommendations over time.
5. Present the recommendations in an empathetic way.

Tash and Stahler caution against gearing evaluations toward one general "imaginary" user audience. They argue that this type of casting tends to develop recommendations that are not specific enough to provide managerial or policymaking direction. The recommendations often do not take into account the mid- and lower-level bureaucrats' needs—especially when the recommendations call for them to alter actions. Therefore, once the users are specified, it may be necessary to produce multiple versions of the same evaluation report to address the concerns of different groups.

Tash and Stahler also suggest that evaluations, and especially the recommendation formulation process, be viewed as a joint venture that would make the role of staff more participatory. As such, the process would enhance utilization because those who participate have a stake in the process. In addition, staff involvement can lead to constructive and doable changes; otherwise recommendations may not make sense in the given context for reasons unknown to the evaluator. Regular meetings regarding interim results are also suggested.

Tash and Stahler suggest that by casting the recommendations more broadly, the utilization of the research findings can be more generalizable to other contexts. Although this may seem to contradict step 2, they suggest that broad implications can be the focus of an additional section of the report.

When possible, both strategically and financially, plans should be made to conduct a follow-up assessment sometime down the road to determine the extent to which the recommendations have been followed. Unfortunately, unless the users are committed to this reassessment, evaluators seldom know the extent to which their proposals have been followed. Reassessment may be more likely, however, if the evaluating unit is part of the evaluated organization.

In step 5 Tash and Stahler are referring to the "style" of the recommendation presentations. Style demonstrates the respect the evaluator has for the evaluated and their environment. Respect detracts from the belief that all evaluators are interlopers who are quick to find fault with staff. Tash and Stahler suggest that rather than merely providing a laundry list of recommendations, an evaluator may present recommendations in the form of case studies that hypothetically explain the intent of the recommendations. Regardless of the presentation, though, it makes sense that utilization would be enhanced by the common-sense method of presenting the results as you would like to receive them. This does not mean that the evaluator is co-opted by the organization, but neither is she or he aloof toward it.

Harry Hatry, Kathryn Newcomer, and Joe Wholey (1994, 594) suggest that to get the most out of evaluations, those at higher levels in the organization should "create incentives for—and remove disincentives to—performance-oriented management and management-oriented evaluations." Some of these incentives are the following:

- Regardless of who sponsors the evaluation, evaluators should seek input from program managers and staff on evaluation objectives and criteria. Where appropriate, evaluators should include the program manager and key program staff on the evaluation team, or as reviewers of the evaluation design and draft reports. Program managers should be kept aware of the progress of evaluations and be given the opportunity to review evaluation findings before they are made public.
- The legislative body, chief executive, or agency head might mandate periodic program evaluations, or at least regular monitoring and reporting of program outcomes.
- The legislative body, chief executive, or agency head could ask program managers to set target levels of performance in terms of key service quality and outcome indicators at the beginning of the year—and to report progress in achieving those targets quarterly and at the end of the year.
- To the extent feasible, the chief executive or agency head should take steps to build achievement of program results into performance appraisal systems for managers and supervisors.
- The chief executive or agency head could develop performance contracts with program managers and with contractors to deliver specific results—both outputs and outcomes.
- To encourage managers and staff to identify opportunities to improve efficiency and effectiveness, legislators or executives could permit agencies to retain a share (say 50 percent) of any savings they achieve from improved service delivery that results in lower costs. The agency should be given considerable flexibility in the use of such funds. The savings probably should be provided only for a limited amount of time, such as one or two years, even though the improvement lasts for many years.
- Legislators or chief executives could give agencies and programs the option of

> either continuing to be under strict personnel controls and restrictions on line-item/object class transfers or being held accountable for program results with substantial added flexibility regarding personnel and financial management (594–96).

Internal versus External Validity

Another difference between academic and evaluation research is the different emphasis each puts on the value of internal versus external validity. A design is internally valid to the extent that the impact is attributable to the treatment and no other factors. External validity refers to the generalizability of the research findings to other sites and situations. Evaluators and laboratory scientists prefer to maximize the former, policy researchers the latter (although both would say they wished to maximize both simultaneously). Why the difference? Program managers, program funding agents, and interested publics are most interested in finding out whether *their* particular program works—whether the impacts can be traced to their program. In doing so, they tend to control for so many factors that the study becomes "generalizable" only to itself. In other words, the program has the observed impacts only under the conditions of the specific situation. In this way evaluators are like scientists who seek to isolate the cause of a disease by ruling out all other factors with laboratory controls. In contrast, policy researchers seek to build theory that is based on the generalizability of their findings. They wish to apply the results more broadly. For example, they may wish to explain expenditures in American cities: What general patterns seem to emerge? Their predictions are to American cities in general, not necessarily the city in which you live. One approach is not more valid than another—only appropriate to different situations. By maximizing specificity, however, generalizability is diminished.

The trade-offs between internal versus external validity need not result in the lack of appreciation of the findings of either practitioners or academics. When evaluators do what Tash and Stahler referred to in step 3, they are increasing the external validity of their models by addressing the concerns of wider audiences. Careful academic research, likewise, can be used by evaluators to provide a grounding in theory for the relationships they hypothesize and a critique of the prevailing methodology. In other words, both types of research are complimentary. Academics need the experience of the pragmatically grounded evaluators and vice versa.

Because evaluations often occur after a program is in place and because they are often conducted under severe time constraints, evaluators are sometimes tentative about the quality of their findings. When they are overly cautious about accepting a treatment's impact when that impact truly exists, the researcher commits a Type II error. When that occurs, a policy or program can be terminated when it, in fact, is beneficial. If one imputes significance to an impact mistakenly, a Type I error is committed. It seems that because the "best" design and "best" measures of variables are unavailable, evaluation findings tend to be attenuated and Type II errors arise. Students are cautioned not to be so critical of their findings that they cancel a program that appears to be effective when the evaluation findings are "not statistically significant."

Ethics and Evaluations

Unlike most academic research in the social sciences, evaluation research has an ethical quality which in a very real sense touches people's lives and may cost jobs. These ethical

considerations and trade-offs are all a part of the art and duty of any evaluation. A number of ethical considerations are involved in program evaluation. First, the evaluator should be aware of the purposes for which an evaluation is commissioned. Evaluations for covert purposes should be avoided. Those commissioned to whitewash or to make a program look good are examples of unethical evaluations because the clients dictate the results. Results should emerge from carefully constructed research; they should not be dictated from above. Evaluations commissioned to kill programs are also unethical for the same reason. Another covert purpose to be avoided is when it is clear that the consumer of the evaluation wants to target the replacement of current employees. Although some evaluations may show that the program is good or bad or that someone should be replaced, one should not design evaluations with a preconceived notion of what the results will be.

Unscrupulous evaluators who are in their profession just for the money and who will tailor their results are often referred to as "Beltway Bandits." This phrase refers to the unscrupulous consulting firms that sprang up on the Washington, D.C., Beltway, a highway that encircles the District of Columbia, during the heyday of federal funding of social program evaluations. Not all these firms were unethical. Evidence suggests, however, that almost anyone at the time could get away with calling him- or herself an evaluator regardless of training. Over time, the reputation of these "bandits" became known, and they no longer were granted contracts.

Another ethical concern is confidentiality. The evaluator often has access to information of a confidential nature over the course of a study. Some of this information could be damaging to individuals who supply the information or to their superiors or subordinates. Also, some of the information (such as an individual's income) need not be divulged, although access to that information is necessary to arrive at aggregate statistics. An evaluator has to be conscious of confidentiality concerns whether or not confidentiality was assured.

As mentioned earlier, the results of an evaluation can have a real, personal impact on employees, clients, and agencies. Care should be taken to perform evaluations professionally and empathetically. These are ethical considerations encountered in the business of evaluation.

References

Bartik, Timothy J., and Richard D. Bingham. 1997. "Can Economic Development Programs Be Evaluated?" In *Dilemmas of Urban Economic Development: Issues in Theory and Practice*, ed. Richard D. Bingham and Robert Mier. Thousand Oaks, Calif.: Sage Publications, 246–77.

Bowen, William M., and Chieh-Chen Bowen. 1998. "Typologies, Indexing, Content Analysis, and Meta-Analysis." In *Handbook of Research Methods in Public Administration,* ed. Gerald J. Miller and Marcia L. Whicker. New York: Marcel-Dekker, 51–86.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Cooper, Harris M. 1982."Scientific Guidelines for Conducting Integrative Research." *Review of Educational Research* 52, no. 2 (Summer): 291–302.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis," *Educational Researcher* 5: 3–8.

——. 1978."Integrating Findings: The Meta-Analysis of Research." In *Review of Research in Education,* ed. L.S. Schulman. Itasca, Ill.: Peacock, 351–79.

Hatry, Harry P., Kathryn E. Newcomer, and Joseph S. Wholey. 1994."Conclusion: Improving Evaluation Activities and Results." In *Handbook of Practical Program Evaluation,* ed. Wholey, Hatry, and Newcomer. San Francisco: Jossey-Bass, 590–602.

Horst, Pamela, Joe. N. Nie, John W. Scanlon, Joseph S. Wholey. 1974."Program Management and the Federal Evaluation." *Public Administration Review* 34, no. 4 (July/August): 300–308.

Light, Richard J. ed. 1983. *Evaluation Studies Review Annual*, vol. 8, entire issue.

Light, Richard J., and Paul V. Smith. 1971. "Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies." *Harvard Educational Review* 41 (November, 1971): 429–71.

Light, Richard J. 1979."Capitalizing on Variation: How Conflicting Research Findings Can Be Helpful for Policy." *Educational Researcher* 8 (October): 3–11.

Nachmias, David, and Claire L. Felbinger. 1982. "Utilization in the Policy Cycles: Toward a Conceptual Framework." *Policy Studies Review* 2, no. 2 (Fall): 300–308.

Patton, Michael Q. 1978. *Utilization-Focused Evaluation.* Beverly Hills, Calif.: Sage.

Strosberg, M. A., and Joseph S. Wholey. 1983. "Evaluability Assessment: From Theory to Practice in the Department of Health and Human Services." *Public Administration Review* 43, no. 1: 66–71.

Tash, William R., and Gerald J. Stahler. 1982. "Enhancing the Utilization of Evaluation Findings." *Community Mental Health Journal* 18, no. 3 (Fall): 180–89.

Trisko, Karen Sue, and V. C. League. 1978. *Developing Successful Programs.* Oakland, Calif.: Awareness House, chap 7.

CHAPTER 2

Evaluation Designs

In the preface and chapter 1, we explicitly noted that this book does not cover all types of evaluation, in particular process evaluations, although they are important, even critically important. But the focus here is on outcome evaluations (or impact evaluations, terms we will use synonymously) because they are really the only way to determine the success of programs in the public and nonprofit sectors. In the private sector there is always the bottom line—the profit and loss statement. Did the company show a profit? How much? The outcome evaluation is the profit and loss statement of the public sector. Did the program work? How effective is it compared to other alternatives?

Very often, the trouble with new public programs is that they are highly touted as being responsible for the improvement in some service delivery because politicians want to take credit for some improvement in that condition, regardless of cause and effect. President Bill Clinton wanted to take credit for a balanced budget occurring two years ahead of the planned balanced budget; however, it was a robust economy that deserved the credit. Moreover, President Clinton wanted to take credit for the robust economy, when he actually had very little to do with it.

The same behavior occurs in state and local governments and in nonprofit agencies. Mayor Rudolph Giuliani of New York City was proud of his community policing initiatives and quick to claim that these initiatives had led to reduced crime in New York City. Of course it may be something else, such as demographics (fewer teenage males), that really "caused" the reduction in crime. Yet community policing has been given credit for reducing crime, and mayors all over the country have jumped on the bandwagon and are initiating community policing activities in their communities merely based on these two events occurring at the same time. New York's crime rate declined at the same time that community policing efforts were initiated and the number of teenage males was reduced.

Now, we do not want to be seen as throwing cold water on a potentially good program. Community policing may be everything it is believed to be. We are simply urging caution. A new program should be evaluated before it is widely cloned.

The fact remains, however, that the public sector has the propensity to copy and clone programs before they are known to work or not work. Examples are numerous, including boot camps, midnight basketball,

neighborhood watches, and the DARE (Drug Abuse Resistance Education) program (Butterfield 1997). DARE has been adopted by 70 percent of the school systems in the United States, yet evaluations of the program (one is included in this book) call the program's efficacy into question. So much pressure is placed on school officials to "do something about the drug problem" that programs like DARE are adopted more for their political purposes than for what they do—because they *may* do something. But if they are doing nothing, they are using up portions of school days that could be better spent on traditional learning.

So this is why we are emphasizing outcome evaluations. Before the public sector, or the nonprofits, run around helter skelter adopting the latest fad just to seem progressive, it is first important to see if programs have their desired outcomes, or at least are free of undesirable outcomes.

The Problem of Outcome Evaluations

The questions "Does the program work?" and "Is it the program that caused the change in the treatment group or is it something else?" are at the heart of the problem of doing outcome evaluations. In other words, the evaluator wants to compare what actually happened with what "would have happened if the world had been exactly the same as it was except that the program had not been implemented" (Hatry, Winnie, and Fisk 1981, 25). Of course this task is not possible because the program does exist, and the world and its people have changed ever so slightly because of it. Thus, the problem for evaluators is to simulate what the world would be like if the program had not existed. Would those children have ever achieved the eleventh-grade reading level without that reading program? Would all of those people have been able to quit smoking without the American Cancer Society's smoking cessation program? How many of those attending the Betty Ford Clinic would have licked alcoholism anyway if the clinic had not existed? The answers to those and similar questions are impossible to determine exactly, but evaluation procedures can give us an approximate idea what would have happened.

The Process of Conducting Outcome Evaluations

How does one go about conducting an outcome evaluation? A pretty simple step-by-step process makes it fairly clear. That process consists of the following:

1. Identify Goals and Objectives
2. Construct an Impact Model
3. Develop a Research Design
4. Develop a Way to Measure the Impact
5. Collect the Data
6. Complete the Analysis and Interpretation

Identify Goals and Objectives

It is difficult to meaningfully evaluate program outcomes without having a clear statement of an organization's goals and objectives. Most evaluators want to work with clearly stated goals and objectives so that they can measure the extent to which those goals and objectives are achieved. For our purposes we will define a program's goal as the broad purpose toward which an endeavor is directed. The objectives, however, have an existence and reality. They are empirically verifiable and measurable. The objectives, then, measure progress towards the program's goals.

This seems simple enough, but in reality many programs have either confusing or nonexistent goals and objectives. Or, since many goals are written by legislative bodies, they are written in such a way as to defy interpretation by mere mortals. Take the goals set for the Community Action Program by Congress il-

lustrating the near-impossibility of developing decent program objectives. A Community Action Program was described in Title II of the Economic Opportunity Act as one

> (1) which mobilizes and utilizes resources, public or private, of any urban or rural, or combined urban and rural, geographical area (referred to in this part as a "community"), including but not limited to multicity unit in an attack on poverty;
>
> (2) which provides services, assistance, and other activities of sufficient scope and size to give promise of progress toward elimination of poverty or cause or causes of poverty through developing employment opportunities, improving human performance, motivation, and productivity, or bettering the condition under which people live, learn, and work;
>
> (3) which is developed, conducted, and administered with the maximum feasible participation of residents of the areas and members of the groups served; and
>
> (4) which is conducted, administered, or coordinated by public or private nonprofit agency (other than a political party), or a combination thereof. (Nachmias 1979, 13)

This is indeed Congress at its best (and only one sentence). Imagine the poor bureaucrat given the task of developing objectives from this goal.

Construct an Impact Model

Constructing an impact model does not imply constructing a mathematical model (although in some cases this might be done). Typically, the impact model is simply a description of how the program operates, what services are given to whom, when, how. In other words, it is a detailed description of what happens to the program participant from beginning to end and what the expected outcomes are. It is also, in essence, an evaluability assessment.

Develop a Research Design

The evaluator now needs to decide how she or he is going to determine if the program works. This will usually be accomplished by selecting one of the basic models of outcome evaluations described below or one or more of their variants (elaborated on in parts II through V of this book).

Develop a Way to Measure the Impact

Now comes the difficult part. The evaluator must develop ways to measure change in the objectives' criteria. What is to be measured and how is it to be measured? In the nonprofit sector, for example, if the goal of our program is to improve peoples' ability to deal with the stresses of everyday living, we must identify specific measures which will tell us if we have been successful in this regard. The next chapter will deal with these difficulties.

Collect the Data

Again, this is not as easy as it appears. Are we going to need both preprogram and postprogram measures? How many? When should the measurement be taken? Are we measuring the criteria for the entire population or for a sample? What size sample will be used?

Complete the Analysis and Interpretation

This seems to be straightforward enough. Analyze the data and write up the results. In many cases this is simple, but in some cases it is not. Chapters 20 and 21 of this book illustrates a case in point. In these chapters, two different research teams use the same data set to evaluate an educational voucher program in Milwaukee. John Witte and his colleagues at the University of Wisconsin look at the data and conclude that the program was ineffective. At Harvard, Jay Greene and his colleagues, use the same data but a different

methodology and conclude that the program does lead to improved educational achievement for participants. So even analysis and interpretation has its pitfalls.

The Four Basic Models of Outcome Evaluations

Reflexive Designs

With reflexive designs, the target group of the program is used as its own comparison group.[1] Such designs have been termed reflexive controls by Peter Rossi and Howard Freeman (1985, 297). Donald Campbell and Julian Stanley (1963) considered them to be another type of pre-experimental designs. The two types of reflexive designs covered here are simple before-and-after studies that are known as the *one-group pretest-posttest design and the simple time-series design.* The essential justification of the use of reflexive controls is that in the circumstances of the experiment, it is reasonable to believe that targets remain identical in relevant ways before and after the program.

One-Group Pretest-Posttest Design

The design shown in figure 2.1a, the one-group pretest-posttest design, is probably the most commonly used evaluation design; it is also one of the least powerful designs. Here, the group receiving the program is measured on the relevant evaluation criteria before the initiation of the program (A1) and again after the program is completed (A2). The difference scores are then examined (A2–A1), and any improvement is usually attributed to the program. The major drawback of this design is that changes in those receiving the program may be caused by other events and not by the program, but the evaluator has no way of knowing the cause.

Simple Time-Series Design

Figure 2.1b shows the second of the reflexive designs, the simple time-series design. Time-series designs are useful when preprogram measures are available on a number of occasions before the implementation of a program. The design thus compares actual postprogram data with projections drawn from the preprogram period. Here, data on the evaluation criteria are obtained at a number of intervals prior to the program (A1, A2, A3, . . . A*n*). Regression or some other statistical technique is used to make a projection of the criteria to the end of the time period covered by the evaluation. Then the projected estimate is compared with actual post-program data to estimate the amount of change resulting from the program (Ar – Ap). But, as with the case of the one-group pretest-posttest design, the drawback to this design is the fact that the evaluator still has no way of knowing if the changes in the program participants are caused by the program or some other events.

Quasi-Experimental Design

With the quasi-experimental design, the change in the performance of the target group is measured against the performance of a comparison group. With these designs, the researcher strives to create a comparison group that is as close to the experimental group in all relevant respects. Here, we will discuss the *pretest-posttest comparison group design*, although there are others, as we will see in part III of this book.

Pretest-Posttest Comparison Group Design

The design in figure 2.1c, the pretest-posttest comparison group design, is a substantial improvement over the reflexive designs because the evaluator attempts to create a comparison

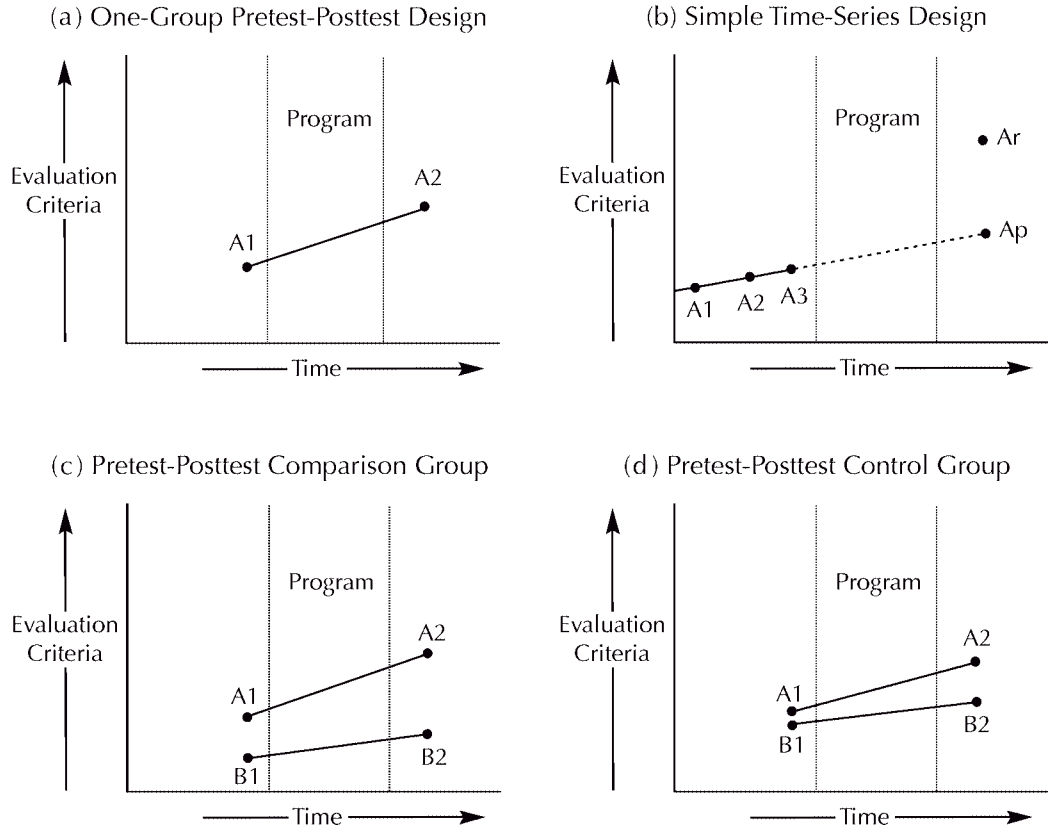


FIGURE 2.1
Evaluation Designs

Source: Adapted from Richard D. Bingham and Robert Mier, eds. 1997. *Dilemmas of Urban Economic Development, Urban Affairs Annual Reviews* 47. Sage, 253.

group that is as similar as possible to the program recipients. Both groups are measured before and after the program, and their differences are compared [(A2 – A1) – (B2 – B1)]. The use of a comparison group can alleviate many of the threats to validity (other explanations which may make it appear that the program works when it does not), as both groups are subject to the same external events. The validity of the design depends on how closely the comparison group resembles the program recipient group.

Experimental Design

A distinction must be made between experimental designs and the quasi-experimental design discussed above. The concern here is with the most powerful and "truly scientific" evaluation design—the controlled randomized experiment. As Rossi and Freeman have noted, randomized experiments (those in which participants in the experiment are selected for participation strictly by chance) are the "flagships" in the field of program evaluation because they allow program per-

sonnel to reach conclusions about program impacts (or lack of impacts) with a high degree of certainty. The experimental design discussed here is the *pretest-posttest control group design.*

Pretest-Posttest Control Group Design

The pretest-posttest control group design shown in figure 2.1d is the most powerful evaluation design presented in this chapter. The only significant difference between this design and the comparison group design is that participants are assigned to the program and control groups randomly. The participants in the program group receive program assistance, and those in the control group do not. The key is random assignment. If the number of subjects is sufficiently large, random assignment assures that the characteristics of subjects in both groups are likely to be virtually the same prior to the initiation of the program. This initial similarity, and the fact that both groups will experience the same historical events, mature at the same rate, and so on, reasonably ensures that any difference between the two groups on the postprogram measure will be the result of the program.

When to Use the Experimental Design

Experimental designs are frequently used for a variety of treatment programs, such as pharmaceuticals, health, drug and alcohol abuse, and rehabilitation. Hatry, Winnie, and Fisk (1981, 42) have defined the conditions under which controlled, randomized experiments are most likely to be appropriate. The experimental design should be considered when there is likely to be a high degree of ambiguity as to whether the outcome was caused by the program or something else. It can be used when some citizens can be given different services from others without significant harm or danger. Similarly, it can be used when the confidentiality and privacy of the participants can be maintained. It can be used when some citizens can be given different services from others without violating moral or ethical standards. It should be used when the findings are likely to be generalizable to a substantial portion of the population. And finally, the experimental design should be used when the program is an expensive one and there are substantial doubts about its effectiveness. It would also be helpful to apply this design when a decision to implement the program can be postponed until the evaluation is completed.

A Few General Rules

Hatry, Winnie, and Fisk also pass on a few recommendations about when the different types of designs should be used or avoided (1981, 53–54). They suggest that whenever practical, use the experimental design. It is costly, but one well-executed conclusive evaluation is much less costly than several poorly executed inconclusive evaluations.

They recommend that the one-group pretest-posttest design be used sparingly. It is not a strong evaluation tool and should be used only as a last resort.

If the experimental design cannot be used, Hatry, Winnie, and Fisk suggest that the simple time-series design and one or more quasi-experimental designs be used in combination. The findings from more than one design will add to the reliance which may be placed on the conclusion.

Finally, they urge that whatever design is initially chosen, it should be altered or changed if subsequent events provide a good reason to do so. In other words, remain flexible.

Threats to Validity

Two issues of validity arise concerning evaluations. They are issues with regard to the design

of the research and with the generalizability of the results. Campbell and Stanley have termed the problems of design as the problem of *internal validity* (1963, 3). Internal validity refers to the question of whether the independent variables did, in fact, cause the dependent variable.

But evaluation is concerned not only with the design of the research but with its results—with the effects of research in a natural setting and on a larger population. This concern is with the *external validity* of the research.

Internal Validity

As was discussed earlier, one reason experimental designs are preferred over other designs is their ability to eliminate problems associated with internal validity. Internal validity is the degree to which a research design allows an investigator to rule out alternative explanations concerning the potential impact of the program on the target group. Or, to put it another way, "Did the experimental treatment make a difference?"

In presenting practical program evaluation designs for state and local governments, Hatry, Winnie, and Fisk constantly alert their readers to look for external causes that might "really" explain why a program works or does not work. For example, in discussing the one-group pretest-posttest design, they warn:

> Look for other plausible explanations for the changes. If there are any, estimate their effect on the data or at least identify them when presenting findings. (1981, 27)

For the simple time-series design, they caution:

> Look for plausible explanations for changes in the data other than the program itself. If there are any, estimate their effects on the data or at least identify them when presenting the findings. (31)

Concerning the pretest-posttest comparison group design, they say:

> Look for plausible explanations for changes in the values other than the program. If there are any, estimate their effect on the data or at least identify them when presenting the findings. (35)

And even for the pretest-posttest control group design, they advise:

> Look for plausible explanations for the differences in performance between the two groups due to factors other than the program. (40)

For many decision makers, issues surrounding the internal validity of an evaluation are more important than those of external validity (the ability to generalize the research results to other settings). This is because they wish to determine specifically the effects of the program *they* fund, *they* administer, or in which *they* participate. These goals are quite different from those of academic research, which tends to maximize the external validity of findings. The difference often makes for heated debate between academics and practitioners, which is regrettable. Such conflict need not occur at all if the researcher and consumers can reach agreement on design and execution of a project to meet the needs of both. With such understanding, the odds that the evaluation results will be used is enhanced.

Campbell and Stanley (1963) identify eight factors that, if not controlled, can produce effects that might be confused with the effects of the experimental (or programmatic) treatment. These factors are Hatry, Winnie, and Fisk's "other plausible explanations." These threats to internal validity are history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality, and selection-maturation interaction. Here we discuss only seven of these

threats. Selection-maturation interaction is not covered because it is such an obscure concept that real world evaluations ignore it. Campbell and Stanley state that an interaction between such factors as selection with maturation, history, or testing is unlikely but mostly found in multiple-group, quasi-experimental designs (5, 48). The problem with the interaction on the basis of selecting the control and/or experimental groups with the other factors is that the interaction may be mistaken for the effect of the treatment.

History

Campbell and Stanley define history simply as "the specific events occurring between the first and second measurement in addition to the experimental variable"(1963, 5). Historical factors are events that occur during the time of the program that provide rival explanations for changes in the target or experimental roup. Although most researchers use the Campbell and Stanley nomenclature as a standard, Peter Rossi and Howard Freeman chose different terms for the classical threats to validity. They distinguish between long-term historical trends, referred to as "secular drift," and short-term events, called "interfering events"(1986, 192–93). But, regardless of the terminology, the external events are the "history" of the program.

Time is always a problem in evaluating the effects of a policy or program. Obviously, the longer the time lapse between the beginning and end of the program, the more likely that historical events will intervene and provide a different explanation. For some treatments, however, a researcher would not predict an instantaneous change in the subject, or the researcher may be testing the long-term impact of the treatment on the subject. In these cases, researchers must be cognizant of intervening historical events besides the treatment itself that may also produce changes in the subjects. Not only should researchers identify these effects; they should also estimate or control for these effects in their models. For example, chapter 8 presents the results of an evaluation of a home weatherization program for low-income families in Seattle. The goal of the program was to reduce the consumption of electricity by the low-income families. During the time the weatherization of homes was being implemented, Seattle City Light increased its residential rates approximately 65 percent, causing customers to decrease their consumption of electricity. The evaluator was thus faced with the task of figuring how much of the reduced consumption of electricity by the low-income families was due to weatherization and how much to an effect of history—the rate increase.

Maturation

Another potential cause of invalidity is maturation or changes produced in the subject simply as a function of the passage of time. Maturation is different from the effects of history in that history may cause changes in the measurement of *outcomes* attributed to the passage of time whereas maturation may cause changes in the *subjects* as a result of the passage of time.

Programs that are directed toward any age-determined population have to cope with the fact that, over time, maturational processes produce changes in individuals that may mask program effects. For example, it has been widely reported that the nation's violent crime rate fell faster during 1995 than any other time since at least the early 1970s. This change occurred shortly after the passage of the Crime Act of 1994 and at the same time that many communities were implementing community policing programs. President Clinton credits his Crime Act with the reduction in violent crime and the mayors of the community policing cities credit community policing. But it all may be just an artifact. It is

well documented that teens commit the most crimes. Government figures show that the most common age for murderers is eighteen, as it is for rapists. More armed robbers are seventeen than any other age. And right now the number of people in their late teens is at the lowest point since the beginning of the Baby Boom. Thus, some feel that the most likely cause of the recent reduction of the crime rate is the maturation of the population ("Drop in Violent Crime" 1997).

One other short comment: The problem of maturation does not pertain only to people. A large body of research conclusively shows the effects of maturation on both private corporations and public organizations (for example, Aldrich 1979; Kaufman 1985; Scott 1981).

Testing

The effects of testing are among the most interesting internal validity problems. Testing is simply the effect of taking a test (pretest) on the score of a second testing (posttest). A difference between preprogram and postprogram scores might thus be attributed to the fact that individuals remember items or questions on the pretest and discuss them with others before the posttest. Or the pretest may simply sensitize the individual to a subject area—for example, knowledge of political events. The person may then see and absorb items in the newspaper or on television relating to the event that the individual would have ignored in the absence of the pretest. The Solomon Four-Group Design, to be discussed in chapter 6, was developed specifically to measure the effects of testing.

The best-known example of the effect of testing is known as the "Hawthorne Effect" (named after the facility where the experiment was conducted). The Hawthorne experiment was an attempt to determine the effects of varying light intensity on the performance of individuals assembling components of small electric motors (Roethlisberger and Dickson 1939). When the researchers increased the intensity of the lighting, productivity increased. They reduced the illumination, and productivity still increased. The researchers concluded that the continuous observation of workgroup members by the experimenters led workers to believe that they had been singled out by management and that the firm was interested in their personal welfare. As a result, worker morale increased and so did productivity.

Another phenomenon similar to the Hawthorne effect is the placebo effect. Subjects may be as much affected by the knowledge that they are receiving treatment as by the treatment itself. Thus, medical research usually involves a placebo control. One group of patients is given neutral medication (i.e., a placebo, or sugar pill), and another group is given the drug under study.

Instrumentation

In instrumentation, internal validity is threatened by a change in the calibration of a measuring instrument or a change in the observers or scorers used in obtaining the measurements. The problem of instrumentation is sometimes referred to as *instrument decay*. For example, a battery-operated clock used to measure a phenomenon begins to lose time—a clear measure of instrument decay. But what about a psychologist who is evaluating a program by making judgments about children before and after a program? Any change in the psychologist's standards of judgment biases the findings. Or take the professor grading a pile of essay exams. Do not all of the answers soon start to look alike?

Sometimes the physical characteristics of cities are measured by teams of trained observers. For example, observers are trained to judge the cleanliness of streets as clean, lightly littered, moderately littered, or heavily littered. After the observers have been in the field for a while, they begin to see the streets as

average, all the same (lightly littered or moderately littered). The solution to this problem of instrument decay is to retrain the observers.

Regression Artifact

Variously called "statistical regression," "regression to the mean," or "endogenous change," a regression artifact is suspected when cases are chosen for inclusion in a treatment based on their extreme scores on a variable. For example, the most malnourished children in a school are included in a child nutrition program. Or students scoring highest in a test are enrolled in a "gifted" program, whereas those with the lowest scores are sent to a remedial program. So what is the problem? Are not all three of these programs designed precisely for these classes of children?

The problem for the evaluation researcher is to determine whether the results of the program is genuinely caused by the program or by the propensity for a group over time to score more consistently with the group's average than with extreme scores. Campbell and Stanley view this as a measurement problem wherein deviant scores tend to have large error terms:

> Thus, in a sense, the typical extremely high scorer has had unusually good "luck" (large positive error) and the extremely low scorer bad luck (large negative error). Luck is capricious, however, so on a posttest we expect the high scorer to decline somewhat on the average, the low scorers to improve their relative standing. (The same logic holds if one begins with the posttest scores and works back to the pretest.) (1963, 11)

In terms of our gifted and remedial programs, an evaluator would have to determine if declining test scores for students in the gifted program are due to the poor functioning of the program or to the natural tendency for extreme scorers to regress to the mean. Likewise, any improvement in scores from the remedial program may be due to the program itself or to a regression artifact. Campbell and Stanley argue that researchers rarely *account* for regression artifacts when subjects are selected for their extremity, although they may acknowledge such a factor may exist.

In chapter 9, Garrett Moran attempts to weed out the impact of regression artifact from the program impact of a U.S. government mine safety program aimed at mines with extremely high accident rates.

Selection Bias

The internal validity problem involving selection is that of uncontrolled selection. Uncontrolled selection means that some individuals (or cities or organizations) are more likely than others to participate in the program under evaluation. Uncontrolled selection means that the evaluator cannot control who will or will not participate in the program.

The most common problem of uncontrolled selection is target self-selection. The target volunteers for the program and other volunteers are likely to be different from those who volunteer. Tim Newcomb was faced with this problem in trying to determine the impact of the home weatherization program aimed at volunteer low-income people (see chapter 8). Those receiving weatherization did, in fact, cut their utility usage. But was that because of the program or the fact that these volunteers were already concerned with their utility bills? Newcomb controlled for selection bias by comparing volunteers for the weatherization program with a similar group of volunteers at a later time.

Experimental Mortality

The problem of experimental mortality is in many ways similar to the problem of self-selection, except in the opposite direction. With experimental mortality, the concern is

why subjects drop out of a program rather than why they participate in it. It is seldom the case that participation in a program is carried through to the end by all those who begin the program. Dropout rates vary from project to project, but unfortunately, the number of dropouts is almost always significant. Subjects who leave a program may differ in important ways from those who complete it. Thus, postprogram measurement may show an inflated result because it measures the progress of only the principal beneficiaries.

External Validity

Two issues are involved with external validity—*representativeness of the sample* and *reactive arrangements*. While randomization contributes to the internal validity of a study, it does not necessarily mean that the outcome of the program will be the same for all groups. Representativeness of the sample is concerned with the generalizability of results. One very well-known, and expensive, evaluation conducted a number of years ago was of the New Jersey Income Maintenance Experiment (Rees 1974). That experiment was designed to see what impact providing varying amounts of money to the poor would have on their willingness to work. The study was heavily criticized by evaluators, in part because of questions about the generalizability of the results. Are poor people in Texas, or California, or Montana, or Louisiana likely to act the same as poor people in New Jersey? How representative to the rest of the nation are the people of New Jersey? Probably not very representative. Then are the results of the New Jersey Income Maintenance Experiment generalizable to the rest of the nation? They are not, according to some critics.

Sometimes evaluations are conducted in settings which do not mirror reality. These are questions of reactive arrangements. Psychologists, for example, sometimes conduct gambling experiments with individuals, using play money. Would the same individuals behave in the same way if the gambling situation were real and they were using their own money? With reactive arrangements, the results might well be specific to the artificial arrangement alone (Nachmias and Nachmias 1981, 92–93).

Handling Threats to Validity

The reason that experimental designs are recommended so strongly for evaluations is that the random assignment associated with experimental evaluations cancels out the effect of any systematic error due to extrinsic variables which may be related to the outcome. The use of a control group accounts for numerous factors simultaneously without the evaluator's even considering them. With random assignment, the experimental and control groups are, by definition, equivalent in all relevant ways. Since they will have the exact same characteristics and are also under identical conditions during the study except for their differential exposure to the program or treatment, other influences will not be confounded with the effect of the program.

How are the threats to validity controlled by randomization? First, with history, the control and experimental groups are both exposed to the same events occurring during the program. Similarly, maturation is neutralized because the two groups undergo the same changes. Given the fact that both groups have the same characteristics, we might also expect that mortality would effect both groups equally; but this might not necessarily be the case. Using a control group is also an answer to testing. The reactive effect of measurement, if present, is reflected in both groups. Random assignment also ensures that both groups have equal representation of extreme cases

so that regression to the mean is not the case. And, finally, the selection effect is negated (Nachmias and Nachmias 1981, 91–92).

Other Basic Designs

The brief description of several evaluation designs covered earlier in this chapter will be amplified and expanded upon in later chapters, and variations on the basic designs will be discussed. For example, a design not yet covered, the Solomon Four-Group Design, a form of experimental design, will be discussed in chapter 6. Two variations of designs covered in the first edition of *Evaluation in Practice* are not covered here, however—the *factorial design* and the *regression-discontinuity design*. We do not have a chapter on either design because we were not able to identify *good* examples of the procedures in the literature. Instead, they will be discussed here.

A *factorial design*, a type of experimental design, is used to evaluate a program that has more than one treatment or component. It is also usually the case that each level of each component can be administered with various levels of other components. In a factorial design, each possible combination of program variables is compared—including no program. Thus, a factorial evaluation involving three program components is similar to three separate experiments, each investigating a different component.

One of the major advantages of the factorial design is that it allows the researchers to identify any interaction effect between variables. David Nachmias stated:

> In factorial experiments the effect of one of the program variables at a single level of another of the variables is referred to as the simple effect. The overall effect of a program variable averaged across all the levels of the remaining program variables is referred to as its main effect. Interaction describes the manner in which the simple effects of a variable may differ from level to level of other variables. (1979, 33)

Let us illustrate with an overly simplistic solution. Figure 2.2 illustrates a program with two components, A and B. In an evaluation of the operation of a special program to teach calculus to fifth graders, component A is a daily schedule of 30 minutes of teaching with a_1 indicating participation in the program and a_2 indicating nonparticipation. Component B is a daily schedule of 30 minutes of computerized instruction, with b_1 indicating participation and b_2 indicating nonparticipation. The possible conditions are the following:

a_1b_1	teaching and computer
a_1b_2	teaching only
a_2b_1	computer only
a_2b_2	neither teaching nor computer

Also suppose that a_1b_2 improves the students' knowledge of calculus by one unit, that a_2b_1 also improves the students' knowledge of calculus by one unit, but that a_1b_1 improves the students' knowledge of calculus by three units. This is an example of an interaction effect. The combination of receiving both components of the program produces greater results than the sum of each of the individual components.

Of course in the real world, factorial evaluations are seldom that simple. In an

		Treatment *A*	
		a_1	a_2
		(yes)	(no)
Treatment *B*	b_1 (yes)	a_1b_1	a_2b_1
	b_2 (no)	a_1b_2	a_2b_2

FIGURE 2.2
The 2^2 Factorial Design

evaluation of a negative income tax experiment conducted in Seattle and Denver, there were two major components—(1) cash transfer treatments (NIT) and (2) a counseling/training component. But the experiment was quite complicated because more than one version of both NIT and counseling/training were tested. With NIT, three guarantee levels and four tax rates combined in such a way as to produce eleven negative income tax plans (one control group receiving no treatment made twelve NIT groups in all). For counseling/training, there were three treatment variants plus a control group receiving no counseling/training, making a total of four. Thus, there were forty-eight possible combinations of the twelve NIT plans and the four counseling/training options. On top of this, the families in the program were also randomly assigned to a three-year or five-year treatment duration. Thus, instead of forty-eight possible treatment/control groups, there were ninety-five, if all possible combinations were considered given the two treatment durations. Such a massive experiment would have cost a small fortune, so in the interests of economy and practicality, the experiment did not test every possible combination of the treatments.

The *regression-discontinuity design* is a quasi-experimental design with limited applicability. In those cases for which the design is appropriate, however, it is a strong method used to determine program performance. Two criteria must be met for the use of the regression-discontinuity design:

1. The distinction between treated and untreated groups must be clear and based on a quantifiable eligibility criterion.
2. Both the eligibility criterion and the performance or impact measure must be interval level variables (Langbein 1980, 95).

These criteria are often met in programs that use measures such as family income or a combination of income/family size to determine program eligibility. Examples of these include income maintenance, rent subsidy, nutritional and maternal health programs, or even in some cases more aggregate programs such as the (former) Urban Development Action Grants. The U.S. Department of Housing and Urban Development's Urban Development Action Grants' (UDAG) eligibility criteria consisted of a summated scale of indexes of urban distress. This scale approximated an interval scale and determined a city's ranking for eligibility. For a regression-discontinuity analysis, one would array the units of analysis (cities) by their UDAG scores on the X axis and cut the distribution at the point where cities were funded (impact), and then plot a development score of some kind on the Y axis. See the following example for further explanation. Donald Campbell and Julian Stanley caution that these eligibility cutting points need to be clearly and cleanly applied for maximum effectiveness of the design (1963, 62). Campbell (1984) refers to the unclear cut points as "fuzzy cut points."

In the regression-discontinuity design, the performance of eligible program participants is compared to that of untreated ineligibles. The example in figure 2.3 from Laura Langbein's book depicts a regression-discontinuity design assessing the impact of an interest credit program of the Farmer's Home Administration on the market value of homes by income level (1980, 95–96). Figure 2.3 shows that the income eligibility cut-off for the interest credit program is $10,000. Because income is an interval variable and it is assumed that the eligibility criterion is consistently upheld, the program's impact is measured as the difference in home market value at the cut-off point. For the program to be judged successful, estimated home values for those earning $9,999 should be statistically significantly higher than for those with incomes of $10,001.

To perform the analysis, two regression lines must be estimated—one for the treated eligibles and the other for the untreated ineligibles. For each, the market values of the home (Y) is regressed on income (X). Although it is intuitively pleasing in this example that both lines are upward sloping (i.e., that market value increases with income), that measure of program impact is at the intercepts (*a*) of the regression lines. If the difference between a_t ("*a*" for treated eligibles) and a_{ut} ("*a*" for untreated ineligibles) is statistically significant, then the program is successful. If not, the program is ineffective (the *a*s were the same) or dysfunctional (a_{ut} was statistically significantly different from a_t and higher than a_t).

Langbein describes the rationale of the design as follows:

> It assumes that subjects just above and just below the eligibility cut point are equivalent on all potentially spurious and confounding variables and differ only in respect to the treatment. According to this logic, if the treated just-eligibles have homes whose value exceeds those of the untreated just-ineligibles, then the treatment should be considered effective. The linear nature of the regression makes it possible to extrapolate these results to the rest of the treated and untreated groups (1980, 95–96).

Thus, the design controls for possible threats to internal validity by assuming that people with incomes of $9,999 are similar in many respects to those who earn $10,001.

Several cautions are advanced to those considering using this design. First, if the eligibility cut-off for the program in question is the same as that for other programs, multiple treatment effects should be considered so as not to make the results ambiguous. Second, the effects of self-selection or volunteerism among eligibles should be estimated. A nested experiment within the design can eliminate this threat to internal validity. Third, when people with the same eligibility are on both sides of the cut point line, a fuzzy cut point has been used. If the number of such misclassifications is small, the people in the fuzzy gap can be eliminated. One should use this technique cautiously, however. The further apart on the eligibility criterion the units are, the less one is confident that the treated and untreated are comparable. A variation of a fuzzy cut point occurs when ineligibles "fudge" their scores on the eligibility criterion. This non-random measurement error is a threat to the validity of the design.These effects should be minimized (by careful screening) and at least estimated. Fourth, because regression is a linear statistic, a visual examination of the bivariate distributions ensures that nonlinear trends are not being masked. An associated caution is that one should remember that the program's effect is being measured at only a relatively small band of eligibility/ineligibility; consequently, a visual inspection of overall trends seems reasonable for one's claiming programmatic impact.

The Need for an Evaluability Assessment

As mentioned in chapter 1, evaluability assessments are made to determine the extent to which a credible evaluation can be done on a program before committing full resources to the evaluation. The process was developed by Joseph Wholey (1979) in the 1970s but has been widely adopted by others (in particular Rutman 1980). An evaluability assessment addresses two major concerns: program structure, and the technical feasibility of implementing the evaluation methodology.

The concern with program structure is essentially overcome by conducting a mini-process evaluation. The first step in conducting an evaluability assessment is to prepare a document model of the program. Here, the evaluator prepares a description of the pro-

gram as it is supposed to work according to available documents. These documents include legislation, funding proposals, published brochures, annual reports, minutes of meetings, and administrative manuals. This is the first step for the evaluator in attempting to identify program goals and objectives and to construct the impact model (description of how the program operates).

The second step in conducting the evaluability assessment is to interview program managers and develop a manager's model. Here the evaluator presents the managers and key staff members with the document model and then conducts interviews with them to determine how their understanding of the program coincides with the document mode. The evaluator then reconciles the differences between the documents model and the manager's model.

The evaluator then goes into the field to find out what is really happening. The documents model and manager's model are not usually enough to yield a real understanding of the program's complex activities and the types of impacts it may be producing. Through the fieldwork the evaluator tries to understand in detail the manner in which the program is being implemented. She or he compares the actual operations with the models she or he has developed. With this work complete, the evaluator is now in a position to identify which program components and which objectives can seriously be considered for inclusion in the impact study.

Finally, the evaluator considers the feasibility of implementing the proposed research design and measurements. Is it possible to select a relevant comparison group? Can the specified data collection procedures be implemented? Would the source of the data and the means of collecting it yield reliable and valid information (Rutman 1980)?

Evaluability assessments will increase the probability of achieving useful and credible evaluations. This was ever so clearly illustrated by one of our favorite evaluation stories by Michael Patton (case 1.1 in chapter 1).

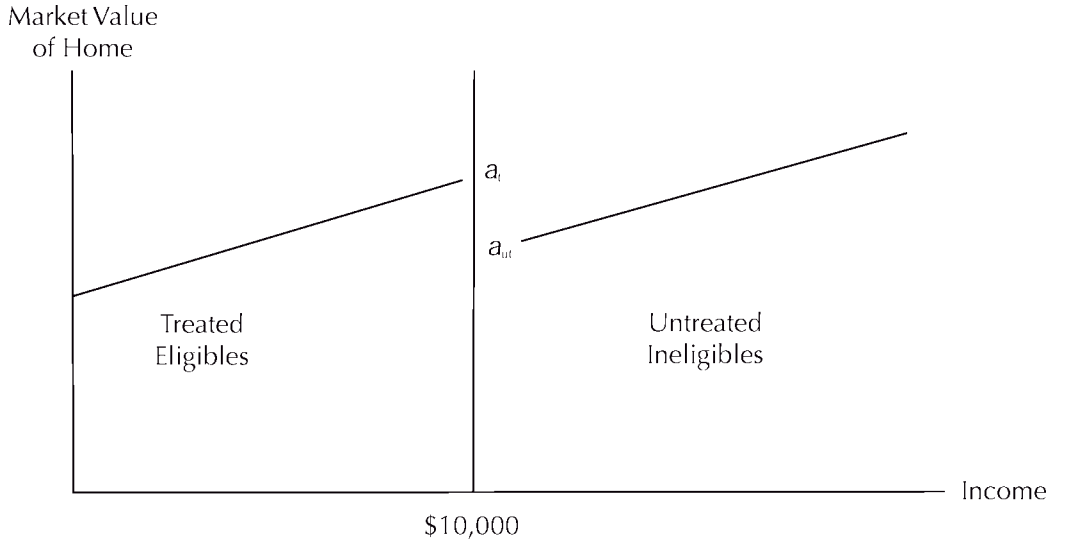


FIGURE 2.3
Regression-Discontinuity Design: The Impact of Subsidized Interest Credit on Home Values

Source: Langbein, 1980, 95.

Note

1. Although a number of authors in the evaluation field use the terms *control group* and *comparison group* interchangeably, they are not equivalent. Control groups are formed by the process of random assignment. Comparison groups are groups that are matched to be comparable in important respects to the experimental group. In this book, the distinction between control groups and comparison groups is strictly maintained.

References

Aldrich, Howard E. 1979. *Organizations and Environments.* Englewood Cliffs, N.J.: Prentice Hall.

Bartik, Timothy J., and Richard D. Bingham. 1997. "Can Economic Development Programs Be Evaluated?" In *Dilemmas of Urban Economic Development: Issues in Theory and Practice*, ed. Richard D. Bingham and Robert Mier. Thousand Oaks, Calif.: Sage, 246–77.

Butterfield, Fox. 1997. "Study Suggests Shortcomings in Crime-Fighting Programs." [Cleveland] *Plain Dealer* (April 17), 3–A.

Campbell, Donald T. 1984."Foreword," in *Research Design for Program Evaluation: The Regression-Discontinuity Approach,* by William M. Trochim. Beverly Hills, Calif.: Sage, 29–37.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Boston: Houghton Mifflin.

_____. 1963. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching,* ed. N.L. Gage. Chicago: Rand McNally, 171–247.

"Drop in Violent Crime Gives Rise to Political Debate." 1997. [Cleveland] *Plain Dealer* (April 14), 5–A.

Hatry, Harry P., Richard E. Winnie, and Donald M. Fisk. 1981. *Practical Program Evaluation for State and Local Governments.* 2d ed. Washington, D.C.: Urban Institute.

Horst, Pamela, Joe Nay, John W. Scanlon, and Joseph S. Wholey. 1974. "Program Management and the Federal Evaluator." *Public Administration Review* 34, no. 4: 300–308.

Kaufman, Herbert. 1985. *Time, Chance, and Organizations: Natural Selection in a Perilous Environment.* Chatham, N.J.: Chatham House.

Langbein, Laura Irvin. 1980. *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation.* Glenview, Ill.: Scott, Foresman.

Nachmias, David. 1979. *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's.

Nachmias, David, and Chava Nachmias. 1981. *Research Methods in the Social Sciences.* 2d ed. New York: St. Martin's.

Patton, Michael Q. 1978. *Utilization-Focused Evaluation.* Beverly Hills, Calif.: Sage.

Posavac, Emil J., and Raymond G. Carey. 1997. *Program Evaluation: Methods and Case Studies.* 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Rees, Albert. 1974. "An Overview of the Labor-Supply Results." *Journal of Human Resources* 9 (Spring), 158–80.

Roethlisberger, Fred J., and W. Dickson. 1939. *Management and the Worker.* Cambridge, Mass.: Harvard University Press.

Rossi, Peter H., and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach.* 3d. ed. Beverly Hills, Calif.: Sage.

Rutman, Leonard. 1980. *Planning Useful Evaluations: Evaluability Assessment.* Beverly Hills, Calif.: Sage.

Scott, W. Richard. 1981. *Organizations: Rational, Natural, and Open Systems.* Englewood Cliffs, N.J.: Prentice Hall.

Skidmore, Felicity. 1983. *Overview of the Seattle-Denver Income Maintenance Experiment Final Report.* Washington, D.C.: U.S. Department of Health and Human Services, May.

Wholey, Joseph S. 1979. *Evaluation: Promise and Performance.* Washington, D.C.: Urban Institute.

Supplementary Readings

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* Cambridge, England: Cambridge University Press, chaps. 3 and 5.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation,* Chicago: Dorsey, chaps. 4 and 6.

CHAPTER 3

Measurement

Program evaluation is an empirical enterprise. The logic of empirical inquiry concerns the investigation of measurable, observable phenomena. Empirical knowledge is objective knowledge. While the choice of what one studies may be subjective, the method by which one does it is not. That is why empirical research is straightforward. Once the subject of the investigation is identified, there are rules by which the investigation proceeds. Measurement is a key component of empirical inquiry. In its broadest sense, "measurement is the assignment of numerals to objects or events according to rules" (Stevens 1951, 1).

Conceptualization and the Process of Operationalization

Concepts are the abstract or underlying properties of variables we wish to measure. The more abstract the concept, the more difficult it is to measure. In evaluation research, some concepts are very easy to measure. For our purpose right now, let us assume that "measure" means how we see the concept in reality. For example, if we think that gender has an impact on the outcome of a treatment, gender is the concept and we would measure it by separating people by how gender is manifest—females and males. We do not "see" gender; rather, we see men and women.

In the physical sciences, concepts are easily measured since they are physically "seen"—mass can be weighed, length can be measured, white blood cells can be counted. Some would say that it is unfortunate that in evaluation and the social sciences the gap between many concepts and how the concept is observed in the world is sometimes quite wide. We rather would like to agree with E.L. Thorndike, who said, "If something exists, it exists in some amount. And if it exists in some amount, it can be measured" (Isaac and Michael 1981, 101). That is what makes evaluation an art, a science, and a theoretically stimulating exercise.

Once you have identified a concept of interest, the first step in measurement is to define it conceptually. Let us take a concept to which most people can relate—political participation. An appropriate *conceptual definition* might be, "any effort by citizens to affect the outcome of political decisions." Now, let us assume that our study of political participation is confined to industrialized democracies in North America. That being the case,

we may wish to narrow our definition a bit to "any *legal* effort by citizens to affect the outcome of political decisions." This *stipulative definition* makes sense in this context. It specifically excludes illegal means of participating, such as political assassinations, bombings, airplane hijackings, or kidnappings. This would be appropriate in the North American context. In some countries, however, one could understand that to exclude these illegal forms would be to miss an important part of the concept of political participation as it is practiced in those countries.

That brings us to the *operational definition*. When we operationalize a concept, we define how it is seen in reality. Given our stipulative definition, how do we know political participation when we see it? Sidney Verba, Norman Nie, and Jae-On Kim (1979) identify several variables that operationalize political participation in their classic studies: voting, contacting elected or appointed officials, contributing money for a candidate, running for elective office, trying to persuade someone of your political views, and campaigning for a candidate. When you operationalize a concept, you name the variable that you subsequently will measure.[1]

Measurement refers to the rules for assigning numerals to values of a variable. According to Fred Kerlinger,

> measurement is the game we play with objects and numerals. Games have rules. It is, of course, important for other reasons that the rules be "good" rules, but whether the rules are "good" or "bad," the procedure is still measurement. (1973, 427)

In order to specify the rules, we need to identify the *unit of analysis* of the phenomenon we are measuring. The *unit of analysis* is the general level of social phenomena that is the object of the observation measured at an identified level of aggregation (e.g., individual persons, neighborhood, cities, nations). The variable must be both conceptually and operationally measured at the same level of aggregation or unit of analysis.

For example, let us go back to political participation. If we are interested in explaining the political participation level of people, the unit of analysis would be the individual and it could be measured in a number of ways. One measure of participation measured at the individual level is to count how many times an individual voted in the past five years. The variable is "voting frequency." Another way is to create an index by adding up the different ways a person has reported that they have participated using the Verba, Nie, and Kim items. The variable is a "participation index." A third way is to ask someone if they voted in the last presidential election. The variable is "voted in last presidential election."

Measurement refers to the rules for assigning numbers to values on a variable. For voting frequency, consider the following rule: For each individual in the study, look at the last five years of election records and count the number of times each individual voted. Each voting incident increases the voting frequency by 1. An alternate rule could be as follows: Count the number of times each individual voted in the past five years and divide it by the number of elections for which they were eligible to vote. This alternative measurement controls for the time a person moves into or out of a voting district and also controls for eligibility to vote by virtue of age.

A variety of ways exist for constructing the participation index. One way is to tally the number of Verba, Nie, and Kim items each individual had done in the past year. The result would be an index which would vary from 0 (no participation) to 6 (participated in all forms of political participation). Another, what Verba, Nie, and Kim have argued as a "better" rule/measure, would be to do a factor analysis of the degrees of each item (number

of contacts, amount of money contributed, number of friends you tried to persuade, etc.) and produce a factor score for each individual.

A rule for the final measure of individual participation could be to ask a person whether or not they voted in the last presidential election and assigning a "1" if they reported to have voted and "0" if they reported to have not voted. Alternately, one could go back to the election records and verify actual voting behavior, making the same numerical assignments.

How about if we are interested in participation in urban areas? The question is "How can we measure urban political participation?" The unit of analysis is cities. Remember, the variable measured must be consistent with the unit of analysis. Therefore, we cannot say, "How many times did the city vote in the past five years?" Cities do not vote, individuals do. We can estimate voter turnout, however, by calculating the percentage of eligible voters who voted in the last election.

Let us take a few more easy examples. If we wish to include gender in our study and the unit of analysis is the individual, we can assign a "1" if the person is a male and a "2" if the person is a female. If the unit of analysis is states, then gender can be operationalized as "percent female." If we wish to measure income and the unit of analysis is the individual, we might ask them their pretax income, but a better method (explained below) would be to give them a card upon which income ranges are printed and ask them to give you the letter of the range which includes their income. If your unit of analysis is census tract, you might use "median income" for the tract or "percentage of the tract's residents with incomes below the poverty line."

In program evaluation, we are typically interested in determining whether a program had its desired effect. Measurement, in this case, can be more difficult than these previous examples. First, there must be some agreement on the desired effect. Presumably, a process evaluation would verify whether the program were set up in a manner consistent with program objectives—for example, is a job training program actually delivering job placement and success skills? Does a Head Start Program have an educational component?

How does one measure the success of a job training program? At the individual level, is it getting a job interview, getting a job, or still being in the job six months after placement? What is a "good" placement rate for the agency? Are there any benchmarks (see chapter 4) that can be used in this regard? For a nutrition education program, one might expect to encounter the most "knowledge" (measured by scores on a test) at the end of the course with knowledge leveling off to a plateau above the original baseline. That is why it is so important to go through the steps of conceptually defining the variable, with stipulations if necessary, with operationalizations and units of analysis consistent with the concept of interest.

Measurement deals with rules, and there are rules concerning how good measures are constructed. First, all measures must have categories which are *mutually exclusive* and *collectively exhaustive*. The mutual exclusivity criterion means that each unit (individual, city, nation, etc.) must fit into only one category. In the social sciences, gender is easy. One can be either male or female, but not both. Race is a bit more difficult. Either all permutations of combinations must be listed or you wind up with a lot of people checking "Other" and the researcher's not knowing what that means substantively. Measures are collectively exhaustive if every unit can fit into a category. In other words, every aspect of the variable has a category and value so that each unit can be assigned to a category that reflects its manifestation of the variable.

A final aspect of measurement concern involves committing what is referred to as the

ecological fallacy. The ecological fallacy involves imputing individual level behavior from aggregate-level measures and statistics. To continue with the political participation example, one cannot infer the voting behavior of an individual based on variables measured at the city level. You can infer the other way around, though. By knowing the voting behavior of all individuals in a city, we can produce an aggregate turnout rate for the city. By carefully matching your concept and measure with the appropriate unit of analysis, you can avoid the ecological fallacy.

Levels of Measurement

All statistical choices are made on the basis of the directionality of the hypothesis, the number of variables involved, and the level of measurement of those variables. Consequently, there is seldom a "choice" of statistics but, rather, an appropriate statistic. The higher the level of measurement, the greater the information about the nature of the differences among units. The higher the level of measurement, the more creative you can be in combining information from multiple variables. The purpose of this section is to review the levels of measurement and their constraints.

Nominal Level

At the nominal level of measurement, names or numerals are assigned to classes of categories in a purely arbitrary manner. In nominal level variables, there is no rank ordering, no greater than or less than implied. The numerals themselves are meaningless except to "code" or organize data. In the gender example above, we assigned "1" to males and "2" to females. The assignment could have been reversed. Scoring a "2" on gender does not mean females have more gender than males. In fact, we could have assigned *any* number, say "1282" for males and "400" for females. They just connote labels for categories. However, it is customary to begin coding with "1" and follow up with consecutive numerals until the number of categories is exhausted. This reduces coding or keying errors which would have to be "cleaned" once the data set is built.

Ordinal Level

When variables are measured at the ordinal level, their values are arranged in a sequence from lesser amounts of the concept to greater amounts of that attribute. It implies a ranking, either individual discrete ranks or grouped ranks, and suggests "better or worse," "more or less" on the variable; however, it does not tell us how much more. An example of discrete ranks is a ranking of the states on an attribute, for example, median income, from highest to lowest with no ties. Grouped ranks can be rankings on which there can be ties, such as the ranking of academic programs by *U.S. News & World Report*. However, survey-type questions that ask respondents whether they "strongly agree," "agree," "neither agree nor disagree (neutral)," "disagree," or "disagree strongly" with a statement are also considered grouped ranks. In this latter case, we know that someone who "strongly agrees" has a greater feeling of agreement than someone who only "agrees"—however, we do not know by how much.

Interval Level

Variables measured at the interval level have a constant unit length between adjacent categories. For example, if we are measuring units in inches, the distance between one inch and two inches is the same as the distance between the 50th and 51st inch—one inch. The unit length between $1 and $2 is the

same as between $1,000 and $1,001. Therefore, interval-level variables have all of the characteristics of ordinal variables with the addition of the constant unit length. This constant unit length allows one to perform mathematical functions—to add, to subtract, to multiply, to divide, to exponentiate, and so forth.[2]

Since you can do more operations with interval-level variables than with ordinal or nominal, is interval-level measurement always the *best* measure? It *usually* is because, if it is appropriate, you can "collapse" categories of adjacent interval-level variables and make them ordinal-level variables. You lose some of the information, but it may be aesthetically pleasing in a cross-tabluation display. You can always make a higher-level variable into a lower-level variable, but not vice versa. However, there are situations in which gathering data at the ordinal level may be more appropriate.

Take the case of income. The *best* measure from the level of measurement perspective is number of dollars of income during the previous year. However, income is considered a personal statistic—not one which many people are willing to share, or, if they are willing, there may be questions about accurate recall. It is better to gather income data as it is reported in the Census—by multiple (i.e., twenty-five) ordinal categories. People are more willing to give income information as long as they are in reasonable ranges. It is better to have less missing data and an ordinal measure than to have lots of missing interval-level data. In fact, the more ordinal rankings, the more the variable approximates an interval-level variable and can be used as such.[3]

By the way, in cases where the unit of analysis is individuals from the general population, it is reasonable to directly use the Census categories as a base for comparability across studies. It would not be advisable in specialty populations, though. In studies of the poor, there are not enough gradations of the lower income levels for reasonable variance. Likewise, a survey of medical doctors would not yield enough variation at the higher limits of income for meaningful analysis.

Special Case of Dummy Variables

Dummy variables designate the presence or absence of a characteristic. They are often constructed to be used as interval-level variables in statistics such as multiple regression since the coefficients are interpretable. Dummy variables are always coded with a "1," designating the presence of an attribute and a "0," as the absence of an attribute. Simply, this is how it works. The regression formula is

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

where Y is the dependent variable, a is the intercept where $Y = 0$, b is the regression coefficient, and X is the independent variable.

Let us say that we think gender has an impact on some dependent variable and we would like to use gender in our regression equation. In fact, we think that being female results in more of the dependent variable. Gender is a nominal-level variable, however, and it is incorrect to use nominal-level variables in statistics that assume interval-level measurement. We can transform the nominal-level variable (gender) to a dummy variable (femaleness) by coding females "1" and males "0" (the absence of femaleness). Look at what that does to the equation. If the value for X_1 is the femaleness dummy value, the value of the b_1X_1 portion of the equation would be 0 for men (because you would be multiplying by 0) and the value for women would be some non-zero value. That is why the only coding of dummy variables that will work is the "0" and "1" coding scheme.

What if the nominal variable of interest has more than two categories? For this

example, let us say the unit of analysis is cities and the nominal variable of interest is region of the country operationalized as North, South, East, and West. You would construct a series of dummy variables totaling the number of nominal categories minus 1. In this case you would construct three (4–1) dummy variables of this form:

If region equals "North,"
  northernness equals 1
If region does not equal "North,"
  northernness equals 0
If region equals "South,"
  southernness equals 1
If region does not equal "South,"
  southernness equals 0
If region equals "East,"
  easternness equals 1
If region does not equal "East,"
  easternness equals 0

We do not include a westernness dummy variable because it would overdetermine the model, statistically speaking. In other words, if we have information on whether or not a city is northern, southern, or eastern, we know which ones are western (those which scored 0 on all the preceding dummies). No additional information is provided by that last dummy variable; in fact, it provides redundant information. Our rule of thumb is that when deciding which dummy to "leave out," in the absence of a compelling theoretical choice (like femaleness was above), leave out the potential dummy variable with the smallest number of cases.

Validity and Reliability of Measures

In program evaluation, as well as in the social sciences generally, the issues of validity and reliability of measures is important, as we tend to measure concepts that are not easily observed in reality. In measurement, we impose rules; in assessing the validity and reliability of our measures, we are evaluating the "goodness" of the rules. Psychologists tend to take tests of validity and reliability very seriously. They even have academic journals in which the articles are solely validity and reliability tests. Good program evaluators will routinely evaluate validity and reliability of measures as well.

Validity

When assessing validity we are concerned with whether we are measuring what we think we are measuring. In other words, a measure is *valid* to the extent it captures or measures the concept or thing it is intended to measure. The farther removed our measures are from the concepts they are intended to measure, the more concerned we are with their validity. Measurement validity also requires that the correct numerals are assigned to specific units. The difference between the correct assignment and the actual assignment is called *measurement error*.

William Bowen and Chieh-Chen Bowen (1998) offer several steps to mitigate against measurement error:

- taking repeated measurements of the same empirical entity;
- when the measurements require instrumentation, using mechanical devices to fix the reference point of observation;
- making electronic observations that print automatically whenever possible;
- taking the average of repeated measurements of the same empirical entity (58).

Another method, training observers, is specifically discussed later in this chapter.

The most common form of validity is also the least rigorous, *face validity*. It answers the question, "On the face of it, does this ap-

pear to be a valid measure of the concept?" Face validity is better than no validity at all. For example, does "voting" appear to be a valid measure of political participation? Most reasonable people would agree that it is. Face validity is enhanced if the measure can be linked to previous published literature. In this case, all scholars use voting as a measure of political participation and have to give a good explanation if they do not use it.

Content validity addresses the following question: "How well does the content of the test [i.e., measure] sample the kinds of things about which conclusions are to be drawn?" (Isaac and Michael 1981, 119) This type of validity is typically important in measures like achievement tests where one is concerned about whether the test items are representative of the universe of items that could measure the concept. In assessing content validity, experts use logically deduced judgments to make an assessments of the validity of the instrument. For example, in assessing the items on a school achievement test for second graders, experts would judge whether the items test knowledge or skills one would expect a second grader to have and whether there is an adequate sample of items to tap all aspects of that achievement.

Criterion Validity

Criterion validity addresses this question: "Does the test compare well with external variables considered to be direct measures of the characteristic or behavior in question?" (Isaac and Michael 1981, 119) Basically, one compares the results of a test with another external test or variable assumed to be measuring the same concept.

In program evaluation, criterion validity often takes the form of correlating multiple measures of a concept that are gathered at the same time. This form of criterion validity is referred to as concurrent validity, because the measures are taken concurrently. High correlations between items that are assumed to be measures of the same phenomenon have high concurrent validity.

Another form of criterion validity is predictive validity. In predictive validity one is less concerned with what the test measures as long as it predicts an outcome. Predictive validity is considered to be the stronger of the two forms of criterion validity (Bowen and Bowen 1998, 60). An example of this form is GRE scores. The question is whether GRE scores are predictive of success in graduate school. If they are, then the GRE is a good admissions tool. In most admission decisions, however, the GRE is seldom the sole predictor of success. Multiple measures are used—undergraduate grade point average, essays, letters of recommendation. To the extent that these multiple measures predict success, they have high predictive validity. To the extent that these measures are highly correlated, they exhibit concurrent validity.

Construct Validity

Construct validity relates to this question: "To what extent do certain explanatory concepts or qualities account for performance on a test?" (Isaac and Michael 1981, 119). L. Cronbach (1970, 143) says that there are three parts to construct validation: suggesting what constructs possibly account for test performance, deriving hypotheses from the theory involving the construct, and testing the hypotheses empirically. What separates construct validity from other forms is its reliance on theory and theoretical constructs, along with empirical inquiry into hypothesized relationships. It is more rigorous in that it hypothesizes not only what should be correlated with the measure, but what should not be correlated on the basis of theory.

In program evaluation there are two techniques commonly used to address construct validation. One is to conduct a "known group" test. For example, if you have con-

structed an index that is supposed to measure liberalism, you could administer the items to a group which is known to be liberal (e.g., elected Democrats in Minnesota, members of Common Cause), and they should receive liberal scores. Moreover, if you administered the same instrument to a group known to be conservative (e.g., the Christian Coalition, Family Research Council), they should score low on the liberalism scale. The two groups would then be expected to score similarly on a scale that is unrelated to liberalism, like the concept of being introverted.

A second technique to establish construct validity is the theoretical and empirical use of factor analysis. Fred Kerlinger (1973) argues that factor analysis is the most powerful method of construct validation. Factor analysis not only shows which measures "hang together" and their relative contribution to the index, but it also empirically demonstrates the divergence of the index from other indexes. Conceptually, this is similar to the divergence of the liberalism index and the introversion index above.

Reliability

Kerlinger says a measure is reliable "if we measure the same set of objects again and again with the same or comparable measuring instrument [and] we get the same or similar results" (1973, 443). A measure is reliable to the extent that essentially the same results are produced in the same situation, and that these results can be reproduced repeatedly as long as the situation does not change. Reliable measures are dependable, stable, consistent, and predictable.

The following example demonstrates the distinction between valid measures and reliable measures. Suppose you went to a doctor's office and your weight was logged in at 135 pounds. When you returned home you wanted to check your bathroom scale for accuracy (validity). When you got on the scale it registered 135 pounds. You then assumed that your scale is a valid measure of your weight. It measures what you had hoped it would measure—an accurate reading of your weight (criterion validity).

Suppose, however, when you came home your weight registered 120 pounds. You could not believe it so you got off the scale and then got back on. Once again it registered 120 pounds. So you got off and on again and it registered 120 pounds. Your bathroom scale is reliable, since it took the same situation (you at 135 pounds) and repeatedly gave the same result (registered the weight as 120 pounds). Since you purchased your scale at a garage sale, however, you are confident that the doctor's scale gives a better representation of your weight than your scale—the doctor's scale is a valid measuring instrument. If you had gotten on and off your doctor's scale and it repeatedly registered 135 pounds, the doctor's scale would be judged both valid and reliable.

The basic characteristics of a reliable measure, then, are stability, equivalence, and internal consistency. Stability is demonstrated by generating the same results under similar conditions. Equivalence means that two raters would give the same rating to a single condition using your measurement rules. Internal consistency means that all the items in a measure or index are related to the same phenomenon. The types of reliability tests used on indexes are covered in the next sections.

Split-Half Reliability

This type of reliability test can be used when there are enough items on an instrument such that you can randomly assign items to two subsets, or halves, and correlate the results of the halves. The higher the correlation, the higher the internal consistency of the index items.

Equivalent Forms

Another method to test the reliability of an index is by using equivalent forms. In this method, the same person responds to two instruments that are hypothesized to be equivalent. The results from the two sets of responses are correlated. Once again, the higher the correlation, the greater the evidence of a stable, reliable index.

Test-Retest

The test-retest method is another measure of equivalence of a measure or index. In this situation, a person responds to an instrument (test) and responds to the same instrument (retest) over a period of time. The results of the two sets of responses are correlated and used as evidence of reliability.

There are some easy methods to increase the reliability of measures—by using multiple measures of a concept, by adding more items to an index, by ensuring that there are no ambiguous items or measures, and by providing clear and unambiguous instructions for raters, coders, and survey respondents. The use of multiple measures will ensure that if you find that one of your measures is unreliable, then you have a back-up measure to use.[4] Other statistical tests of reliability (such as Cronbach's alpha) can be found in all statistical packages built for social scientists. If you construct an index and find that it is not reliable enough (the rule of thumb on Cronbach's alpha is 0.8), adding additional items to the index will increase reliability by a predictable amount. Ambiguous items or measures, particularly in mail surveys, yield unreliable results, as respondents may interpret the items in a manner inconsistent with the surveyor's intent. The same is true for ambiguous instructions.

In program evaluation, when you use trained observers to gather data, there is the added concern over *interrater reliability*. Let us say that you are going to evaluate the effectiveness of a neighborhood beautification program. You take four pictures of street conditions ranging from 1, which is very neat and tidy, to 4, which is a mess. You hire students who will rate the blocks you have randomly sampled from each of the participating neighborhoods. Before you send the students out to collect the data, you have a one-hour training session in which you point out what discriminates a 1 from a 2, a 2 from a 3, and a 3 from a 4. These aspects include the amount of visible litter, condition of trash containers, presence of abandoned automobiles, condition of tree and lawn grooming, visibility, and condition of signage. You then take all the students to a nonparticipating neighborhood and have each of them rate the same block (pretesting the instrument) and compare their ratings. If the ratings differ, point out which aspect to take into account to correctly rate the block. Continue the process until all students rate the blocks similarly (or fire the ones who consistently rate incorrectly).

Once you let the students out on their own, you need to check and make sure they are still rating correctly—measuring reliably. A number of ways exist for ensuring this. First, you can spot check while students are rating. Secondly, you can send another student to rate the same block right after it has been rated and assess the interrater reliability. Thirdly, you can send the students in two-person teams, instructing each to make an individual rating and then compare the ratings, reconciling any discrepancies on the spot.

Whenever you are using observers to gather data, they must be trained, they must practice the technique, and they must be periodically checked to ensure reliability of their data gathering techniques.

Measurement in Perspective

Now that we have covered the basics of measurement, we will look at the measurement of an actual variable—"psychological type" measured by the Myers-Briggs Type Indicator (MBTI). The MBTI is a commonly used psychological indicator that sorts people into 16 psychological types. It is derived from Carl Jung's (1875–1961) theory of psychological type based on normal, as opposed to abnormal, psychology. He hypothesized that people act in predictable ways depending on their innate type. The MBTI itself was developed by Katharine Cook Briggs (1875–1968) and her daughter, Isabel Briggs Myers (1897–1980). MBTI is used in a variety of instances including career development, relationship counseling, team building, and diversity training.

The conceptual definition of psychological type is the innate pattern of preferences of an individual that causes him or her to behave in predictable ways. According to Jungian theory, the preferences are based on how one takes in information, how one organizes and processes information, and where one's attention is most easily focused. Operationally, the indicator is a combination of preference sorting on four different, dichotomous dimensions, or scales as they are referred to in the Myers-Briggs literature. Those dimensions are extraversion[5] versus introversion, sensing versus intuition, thinking versus feeling, and judging versus perceiving. Therefore, psychological type is what we call a multidimensional concept—it has four separate dimensions.

Because Jungian theory is based on normal psychology, neither of the dichotomous anchors of a dimension is better than the other—someone preferring extraversion is not better than one preferring introversion; someone preferring sensing is not more normal than someone preferring intuition.[6] The idea is to sort people on each dimension. Therefore, each dimension is a dichotomous nominal variable; preferences are labels. The combination of preferences is a nominal variable as well. The maximum number of permutations of the combinations is sixteen; therefore, there are sixteen personality types.

The MBTI instrument is a self-report questionnaire consisting of ninety-five items. The items are selected and developed based on their ability to distinguish between the individual dichotomies. Each item contributes to only one of the four dichotomies. Some items contribute more to distinguishing between the two ends of the dichotomy. These items are "weighted" on that basis. Consequently, each item receives a score of 0, 1, or 2 based on how it contributes to the particular classification. Examples of some items which contribute to each of the dichotomies are listed in table 3.1.

Researchers or counselors who use the MBTI must be trained professionally in the conduct, scoring, and interpretation of the MBTI. This requirement addresses one component of the reliability of the MBTI. This training ensures that the directions will be given unbiasedly, that the instrument will be scored accurately, and that the ethical considerations in their interpretation and use are assured. Other factors affect the reliability of the MBTI related to the respondents. First, a respondent might feel pressured to respond in a way that is either socially desirable or in a manner the respondent feels is appropriate in light of the environment in which the MBTI is administered. For example, if the MBTI is administered in a work environment, the respondent may answer in a manner to please superiors. Secondly, if a person is lacking in self-esteem, there may be a tendency to answer in terms of what they perceive is their idealized self rather than with their real preferences. Thirdly, if people have developed skills for their job or home environment (such as a natural introvert "learning" to be-

TABLE 3.1
Items from MBTI by Scale (with Weight)

Extravert/Introvert Scale

1. In a large group, do you more often
   (A) introduce others—Extravert (2)
   (B) get introduced?—Introvert (2)
2. Which word in the pair appeals to you more?
   (A) reserved—Introvert (1)
   (B) talkative—Extravert (2)
3. Which word in the pair appeals to you more?
   (A) party—Extravert (1)
   (B) theatre—Introvert (0)
4. Are you
   (A) easy to get to know—Extravert (1)
   (B) hard to get to know?—Introvert (2)

Sensing/Intuitive Scale

1. Would you rather have as a friend
   (A) someone who is always coming up with new ideas—Intuitive (1)
   (B) or someone who has both feet on the ground—Sensing (2)
2. Which word in the pair appeals to you more?
   (A) facts—Sensing (2)
   (B) ideas—Intuitive (1)
3. Which word in the pair appeals to you more?
   (A) foundation—Sensing (0)
   (B) spire—Intuitive (2)
4. Do you think it is more important to be able
   (A) to see the possibilities in a situation—Sensing (1)
   (B) or to adjust to the facts as they are—Intuitive (0)

Thinking/Feeling Scale*

1. Which word in the pair appeals to you more?
   (A) convincing—Thinking (2)
   (B) touching—Feeling (2)
2. Which word in the pair appeals to you more?
   (A) who—Feeling (0)
   (B) what—Thinking (1)
3. Do you more often let
   (A) your heart rule your head—Feeling (2)
   (B) your head rule your heart?—Thinking (1)
4. Would you rather work under someone who is
   (A) always kind—Feeling (1)
   (B) or always fair?—Thinking (0)

Judging/Perceiving Scale

1. Does following a schedule
   (A) appeal to you—Judging (2)
   (B) or cramp you?—Perceiving (2)
2. Which word in the pair appeals to you more?
   (A) systematic—Judging (2)
   (B) casual—Perceiving (2)
3. Which word in the pair appeals to you more?
   (A) quick—Judging (0)
   (B) careful—Perceiving (1)
4. In getting a job done, do you depend on
   (A) starting early, so as to finish with time to spare—Judging (0)
   (B) or the extra speed you develop at the last minute—Perceiving (1)

* The only scale that demonstrates differences by gender is Thinking/Feeling. Only items common to both genders are included in this listing.

come comfortable with public speaking), they may respond with the learned response rather than their natural preferred response. A fourth threat to reliability is when the respondent sees no clear choice, thus answering inconsistently. Misreading of items or leaving too many items unanswered also has an impact on the reliability of the MBTI. Since those who administer the MBTI have no control over these respondent-based threats to reliability, they must use their training to provide an environment in which the respondent is less likely to be susceptible to those threats.

The types of reliability tests used with the MBTI include split-half and test-retest. The MBTI does not have an equivalent form, as all forms contain the same core of scored questions. The results of the split-half reliability tests using data stored in the national MBTI data bank indicated that the reliability is high among adults regardless of age and gender and lowest among "underachieving" high school students (Myers and McCaulley 1985, 164–69) and students in rural areas. The Cronbach's alpha coefficients exceeded 0.8 in all but the Thinking/Feeling Scale (a=0.76).[7] Examination of the available test-retest data shows remarkable consistency over time. When respondents changed preferences, it was usually only on one scale and typically if original strength of the preference was low on that scale (170–74).

In terms of validity, the items in table 3.1 can be examined on the extent of face validity. What do you think? It is generally held that the items that contribute to each dichotomous scale have face validity. The most important validity for MBTI is construct validity, as the indicator is grounded in theory. To estimate validity of the MBTI, researchers correlated the results with a number of other personality indicators, particularly those constructed by other Jungian theorists. The correlations generally for each scale ranged from 0.4 to 0.75 (175–223).

Conclusion

Measurement is a very important aspect of program evaluation. It involves defining concepts of interest, organizing rules for the assignment of numbers to the ways the concept manifests itself in the empirical world, and ensuring the validity and reliability of the measures. When you read the examples of program evaluations in the remainder of the book, be cognizant of the some of the ingenious way the authors measure outcomes and carefully report evidence concerning the validity and reliability of their measures. Measurement is an art and a science. In program evaluation, measurement is crucial when the future of the program is in the evaluator's hands.

Notes

1. Recall that variables indicate the various ways the concept appears in nature. Variables must take on at least two forms (e.g., male and female) or else they are considered *constants.* Constants cannot be used in statistical analyses as statistics are based on variance.
2. Statistical purists distinguish between interval-level variables and ratio-level variables. Ratio variables have an absolute zero and, thus, one can say, for example, that someone who has $50 has twice as much money as someone who has $25. However, this distinction is seldom followed in practice, and determination of appropriate statistics is not changed by virtue of this distinction. Therefore, we have chosen to consider ratio a special form of interval variable.
3. We also know there is a classical debate between statistical purists and methodological pragmatists on when an ordinal-level variable "approximates" an interval-level variable and, consequently, does not seem to violate the assumptions of interval-level statistics. Suffice it to say that a five-category ordinal variable is

probably not used appropriately as an interval-level variable, whereas the twenty-five-category Census variable on income is used more appropriately.

4. Remember that using multiple measures also allows one to assess the validity of the measures. To the extent that these measures are highly correlated (assuming they are both reliable), then validity is enhanced.
5. This spelling is the one Jung preferred and the one used by professionals trained to administer the MBTI.
6. We will not go into a lengthy discussion about the meaning of each of these dichotomous pairs. The purpose here is to provide a measurement example using the MBTI. Those interested in learning more about MBTI should consult Myers and McCaulley 1985 or McCaulley 1981. There are other related books in the popular press, such as Elliott and Lassen 1998 and Pearman and Albritton 1997.
7. The MBTI has a method to convert the raw scores on the scales to interval-level scores for use in interval-level statistics.

References

Bowen, William M., and Chieh-Chen Bowen. 1998. "Typologies, Indexing, Content Analysis, and Meta-Analysis." In *Handbook of Research Methods in Public Administration,* edited by Gerald J. Miller and Marcia L. Whicker. New York: Marcel-Dekker, 51–86.

Cronbach, L. 1970. *Essentials of Psychological Testing.* 3d ed. New York: Harper and Row.

Elliott, Charles H., and Maureen Kirby Lassen. 1998. *Why Can't I Get What I Want?* Palo Alto, Calif.: Davies-Black Publishing.

Isaac, Stephen, and William B. Michael. 1981. *Handbook in Research and Evaluation.* 2d ed. San Diego: EdITS Publishers.

Kerlinger, Fred N. 1973. *Foundations of Behavioral Research.* 2d ed. New York: Holt, Rinehart and Winston.

Kroeger, Otto, and Janet M. Thueson. 1988. *Type Talk.* New York: Delacorte Press.

McCaulley, Mary H. 1981. "Jung's Theory of Psychological Types and the Myers-Briggs Type Indicator." In *Advances in Psychological Assessment* 5, edited by P. McReynolds. San Francisco: Jossey-Bass, 294–352.

Myers, Isabel Briggs, and Mary H. McCaulley. 1985. Manual: *A Guide to the Development and Use of the Myers-Briggs Type Indicator.* Palo Alto, Calif.: Consulting Psychologists Press.

Pearman, Roger R., and Sarah C. Albritton. 1997. *I'm Not Crazy, I'm Just Not You.* Palo Alto, Calif.: Davies-Black Publishing.

Stevens, S. 1951. "Mathematics, Measurement, and Psychophysics." In *Handbook of Experimental Psychology,* edited by S. Stevens. New York: Wiley.

Verba, Sidney, Norman Nie, and Jae-On Kim.1979. *Participation and Political Equality.* New York: Harper and Row.

CHAPTER 4

Performance Measurement and Benchmarking

Performance measurement and benchmarking are hot topics at all levels of government and in nonprofit organizations. Inspired by David Osborne and Ted Gaebler's *Reinventing Government* and launched onto the national scene by then Vice President Al Gore's *National Performance Review* and the Government Performance Results Act of 1993 (Results Act), performance measurement is a label typically given to efforts undertaken by governments and nonprofit agencies to meet the demand for documenting results (Wholey and Hatry 1992).*

The Results Act came about during a time when there was declining public confidence in government, continuing budget pressures, and increasing demands for effective public services. Its purposes are to shift the focus of management to results (as opposed to inputs and outputs); to improve service delivery, program efficiency, and program effectiveness; to improve public accountability; to support resource allocation and other policy decision making; and, ultimately, to improve public confidence in government. The Results Act requires that agencies undertake strategic planning to identify missions, long-term goals, and strategies for achieving goals. These plans must also explain the key external factors that may affect the achievement of those goals.

The process of performance measurement is defined explicitly in the Results Act; surprisingly, it gives an outline that any organization may use to implement performance measurement and performance management. Each agency prepares an annual performance plan that shall

1. establish performance goals to define the level of performance to be achieved by a program activity;
2. express such goals in an objective, quantifiable, and measurable form;
3. briefly describe the operational processes, skills and technology, and the human, capital, information, or other resources required to meet the performance goals;
4. establish performance indicators to be used in measuring or assessing relevant

* The group referred to as the National Performance Review, or NPR, is now called the National Partnership for Reinvention.

outputs, service levels, and outcomes of each program activity;
5. provide a basis for comparing actual program results with the established program goals; and
6. describe the means to be used to verify and validate measured values. (P.L. 103–62, sec. 4b.)

State and local governments have been measuring the effectiveness of their services for quite some time; however, the focus on performance as an accountability measure has never been as prominent as it is today. The purpose of this chapter is to introduce you to the differences between performance measurement and program evaluation, to cover the components of a performance-based management system, to discuss the components of a good performance measure, and to expose you to the concept of benchmarking as it pertains to performance measurement.

Performance Measurement versus Program Evaluation

The Government Performance Results Act of 1993 has been dubbed "the full employment act for program evaluators" (Newcomer 1997, 9) because the demand for quantitative measures is so pervasive. This does not mean that program evaluation and performance measurement are synonymous. Program evaluation is the more inclusive of the terms. Evaluation efforts are systematic efforts to determine whether a program did what it intended and can focus on inputs, operations, and results (Newcomer 1997, 9). They look at "why" a program did or did not function as intended or "how" it functioned. Performance measures are quantitative indicators that tell us "what" is going on in terms of outputs or outcomes, but do not explain why or how this occurred. They deal with high-performing operations.

Harry Hatry (1997) makes the following comparisons. Program evaluations tend to be performed by people from outside the agency—such as university researchers or outside contractors. This is because program managers have a vested interest in demonstrating that their programs do, in fact, work. Moreover, they have an incentive to want to keep their jobs. So, program evaluations are performed by objective, disinterested outside groups or evaluation units within very large organizations (such as the Department of Housing and Urban Development). Performance measurement, however, is conducted internally with data gathered and analyzed by agency officials. It provides these officials with feedback to perform their jobs more effectively. Performance measurement is conducted annually; program evaluations are not. Full-fledged program evaluations are expensive. Annual evaluations really do not make sense, as one would not expect a fundamental change in the outcomes of a program from year to year. Performance measurement, however, implies delivering existing services more effectively, more efficiently, and with greater satisfaction for citizens. This annual feedback on operations tackles a different set of issues than the causality implied in program evaluations.

But this does not mean that the data generated from annual performance management reviews cannot inform program evaluations. In fact, with performance data already gathered, a program evaluator can target more resources to performing the evaluation since time is saved at the data gathering stage. Much of the data gathered for performance is ultimately usable in the evaluation. Kathy Newcomer's quote at the beginning of this section is appropriate, though; because measurement itself is such an important feature in evaluation, evaluators are already trained to deal with quantitative measurement issues.

Performance-Based Management

Performance measurement for its own sake cannot be worthwhile unless it is integrated into a performance-based management system. Figure 4.1 depicts key steps and critical practices in the implementation of performance-based management as implemented in the Results Act. The first step in performance-based management is to get agreement on the program's mission, goals, and the strategies for achieving these goals. From the beginning, it is important to include members from every level of the organization—board members, citizens, customers, regulators, and any other relevant stakeholders. This ensures that expectations about service delivery will be clear from the outset and that the participants have accepted performance measurement and the use of those measures to improve service delivery. As with any strategic planning activity, this step allows participants to assess the program's environment—the strengths, weaknesses, opportunities, and threats—and use that assessment to inform their deliberations.

The second step is to actually measure the performance according to the plan developed in step 1. Good, appropriate performance measures are ones which empirically demonstrate the results of the program's activities. There is neither time nor resources to measure *all* indicators of performance for a particular goal. Consequently, the number of measures should be limited to the best and most vital ones, with measures linked to each goal articulated in step 1. Understandably, measures that can respond to multiple aspects of the program are more efficient than those addressing a single priority. Measures should also be sufficiently complete and accurate. Although that sounds like a given, "sufficiently complete" means that an attempt is made to measure enough aspects or dimensions per goal to feel good about the validity of the measure on that goal. Accuracy, as well, poses both validity and reliability concerns. However, unlike full-scale evaluations, in performance measurement you have to make trade-offs on costs versus validity and reliability, as this is an annual activity.

As with program evaluations, the results of performance measurement cannot be utilized unless they are in the hands of decision makers *before* they make their decisions. It is impossible to gather annual performance data the day the decision makers meet. Therefore, there should be a *plan* and a *schedule* for the routine (at least quarterly) collection of data. Moreover, responsibility for the collection of the data should be assigned specifically to agency personnel who then are held accountable for gathering the data and are evaluated and rewarded for this as a duty associated with their jobs. Table 4.1 lists some specifications of performance measures to keep in mind when establishing a performance-based process.

Performance measurement is implemented to enhance operations. This should be the goal of the process—as opposed to being a method of punishing employees. At step 1, this should be made clear so that at step 3 the results can be used productively. The results can be used to identify performance gaps, and the program can then be modified and improved during the next reporting period. The data can also be used to generate annual reports and be used by decision makers in their executive roles.

Step 3 involves *performance monitoring* or using the data to improve operations. There are a number of modes the monitoring may take. First, current performance can be compared with prior performance. Obviously, this requires baseline data. Therefore, this type of comparison may not be made in the first year of performance management. The second mode is comparing current performance with performance targets. This is the model utilized in the Results Act and in the state of Oregon's (1997) highly touted strategic plan. These per-

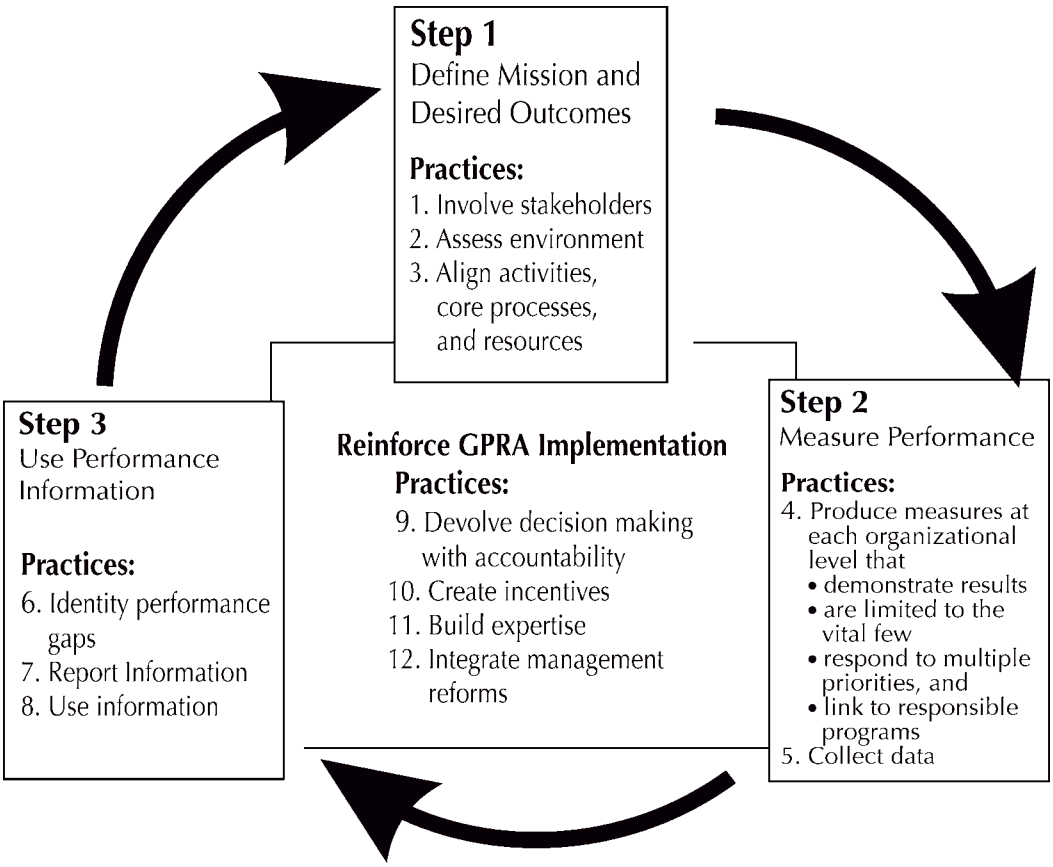


FIGURE 4.1
Performance-Based Management: Key Steps and Critical Practices

Source: U.S. Government Accounting Office, Executive Guide: *Effectively Implementing the Government Results Performance Act*. Washington, D.C.: Government Printing Office, 1996, 10.

formance targets are discussed and agreed upon in step 1. In step 3, the targets are evaluated and modifications in them are suggested. The third method compares the program's performance with the performance of similar programs. This comparison could be between the same programs implemented in different geographical areas. For example, the Environmental Protection Agency could compare performance in divisions with similar functions across its ten regions (or among a subset of comparable regions). A more difficult comparison is between similar programs that are not part of a larger organization. The difficulty comes about in two ways. First, the two programs may not measure performance in the same way, making the comparisons tenuous. Secondly, programs (read "competition") may not want to share information or may be required by confidentiality or proprietary con-

cerns not to release the information. The final comparison mode is through benchmarking. This is discussed in the following section.

In order to be useful in performance monitoring, the performance data should be disaggregated by relevant characteristics. Harry Hatry (1997) identifies this as a major obstacle to the effective use of outcome/performance information.

> Outcome data that are provided are usually too highly aggregated to be meaningful for lower-level personnel. City-wide or state-wide data provide a summary of the outcomes but hide much information as to how the program is working for different categories of customers and for different key service characteristics. Program evaluation personnel routinely break out their data on outcomes by such characteristics. This, however, is not nearly as common in performance measurement in the public sector. For many programs it is clearly desirable to distinguish outcomes for various age groups, persons in different geographical parts of the agencies' jurisdiction, gender, ethnic or racial groups, income category, household size and composition, whether families own or rent their homes, educational level, and so on. In addition outcome data becomes (sic) more useful if segmented by the particular office or facility, such as an individual fire station, specific park, specific library, sanitation district, or social service office. (40)

Once the data are in a usable form, the next step in monitoring performance is for staff to provide explanatory information on the key factors which affected current performance and what can be done to improve performance in the future. Questions should be asked such as, "Why did we not meet the target?" or "Why are the outcome scores so low/high?" Sometimes there is a reasonable explanation beyond staff's control. For example, welfare-to-work programs may be more successful at the beginning of a program when they place those who are most easily placed, a process called "creaming." A performance management response to this situation might be to change the nature of the training portion of the program to better meet the needs of people who are less well equipped to enter the work force. Natural disasters, changes in regulations, reductions in force (RIF), and economic recessions are other examples of uncontrollable factors. You can then make

TABLE 4.1
Quantitative Data Collection

a. Specification of a performance indicator
   - Type information to be collected
   - Data source
     - I. Agency or program records
     - II. Observation
     - III. Survey (mail, telephone, in person)
     - IV. Other (e.g., technology)
   - Data collection instrument
   - Data collection procedures
   - Sampling design (random, purposeful)
   - Sample size

b. Performance indicators should be
   - Relevant
   - Reliable (e.g., based on sufficiently large samples)
   - Valid (e.g., based on high response rates from representative samples)
   - Sensitive
   - Not too costly (e.g., based on small samples)

c. Performance indicators can often be developed from existing informal systems for assessing program performance:
   - Get agreement on standards
   - Use representative samples

statistical adjustments to control for these and any other factors likely to affect performance. The feedback mechanism should not only identify the key factors but provide for a dialogue on how the operation will be modified to enhance performance in the modified environment.

Figure 4.1 also lists some performance-based management practices that can reinforce implementation of the performance monitoring system. By increasing participation and by establishing agreement on the goals, objectives, and measurement system, a manager can delegate greater authority and allow greater flexibility to achieve results. Accountability is enhanced because all actors have accepted their role in the process. The presence of clear, agreed-upon targets makes it easier for managers to create incentives for performance by staff, grantees, or contractors. From their point of view, this takes away some of the uncertainty that occurs when people do not know the basis upon which they will be evaluated—there is little room for managerial subjectiveness to enter the process. Integrating management reforms allows for the redirection of program activities to achieve improved results.

Once you have gone through one round of annual performance measurement, it is reasonable to audit or evaluate the performance measurement system. As a result of the feedback system, you might want to gather contextual information to determine the extent of agreement on goals and strategies to implement them. If the process went smoothly and there were no surprises, there probably was agreement. If there appears to be question of agreement, however, you would want to revisit the mission, goals, and strategic process, making sure all the relevant stakeholders are included in the discussions.

A second aspect of evaluation is an assessment of whether the performance measurement system that is in place is truly the appropriate one. Did you use appropriate measures? Would some other system or method make more sense given the nature of the program? What improvements can be made so that the system produces results? This is a reflection on the process of the implementation of the performance-based management system.

A final utilization question is whether the performance information generated was useful in management as well as in policy decision making. This can be assessed by debriefing the program's management team and noting which, if any, information contributed to their ability to shape the program for improvement. When a particular measure is considered unimportant, is that because it did not measure what it attempted to measure? Should that measure be changed or discarded? For policymakers, was the information available to them before policy decisions were made? Did any policymakers make reference to the measures in their deliberations? This full cycle of evaluation of the performance system results in a dynamic system poised to deal effectively with the program's environment.

Benchmarking

The underlying principle in performance monitoring is improvement of service delivery. It is not enough to simply gather performance data; the information should be used as feedback to continuous improvement in an agency. Greg Meszaros and James Owen make the distinction between performance monitoring and benchmarking, with benchmarking being the "process of comparing internal performance measures and results against those of others in the industry or sector" (1997, 12). They define benchmarking "as a structured series of steps and activities to compare inside products, services, and practices to like activities employed by outside leaders in the industry" (12). So bench-

marking implies comparison and steps to take in the comparison process.

Harry Hatry and Joseph Wholey identify a number of candidates to which an agency can compare itself. The first, and perhaps easiest, comparison is the agency's own previous performance. In other words, the benchmark or baseline from which future improvement is gauged is the previous year's performance. The notion here is that incremental improvement is a reasonable expectation—or that some improvement is better than no improvement at all. A second benchmark can be a comparison with the performance of similar units within the organization. For example, one might compare citizen satisfaction with a city's recreational services with citizen satisfaction with public works services. This requires that data (and services) be comparable enough to merit comparison. Many cities routinely collect citizen satisfaction data from multiservice surveys. These data would be reasonable benchmarks for comparison across departments. Another of the more internal agency benchmarks can come from the comparison of performance outcomes across different client groups. In cities, this comparison can be made, for example, across wards, age groups, income status, or other determinants of "need."

The benchmarks used in the Results Act are preset performance targets agreed upon by the respective agency and monitored by the General Accounting Office. These pre-set targets can be derived in any number of ways (certainly in all the methods discussed in this section). Unlike the benchmarks mentioned earlier, however, these are less routine comparisons upon which success or failure to achieve the targets is more tangible. Consequently, rewards are more easily justified and barriers to performance are quantified.

External benchmarks are increasingly popular. The most basic is a benchmark based on the performance of other jurisdictions or agencies. This requires that the jurisdictions be similar in size, extent of service coverage, income, and other relevant variables. It also requires that comparable performance data are collected across the jurisdictions and that the data are willingly shared. This, of course, assumes a level of trust across the jurisdictions; however, some of these data can be collected from routine reports available from governmental sources and professional associations like the International City/County Management Association.

Another external benchmark can be to state and federal standards in areas where these exist. Examples of standards emanating from the Clean Water Acts or the Environmental Protection Agency easily come to mind. Professional associations also are a source for benchmarking standards. For example, the American Library Association has established standards dealing with per capita holdings and circulation. While these professional association established standards are reasonable targets, one must be aware that sometimes these standards can be inflated in the local context. We would expect the American Library Association to suggest "more" holdings rather than "fewer" holdings. In other words, these standards need to be evaluated in the context of the local environment.

The current trend in benchmarking is to make a comparison to the "best in its class." Richard Fischer (1994, S-4) describes this process as "comparing several competitors on the same benchmarks to see who is best, finding out why that competitor is best, and then using the best practices as a means of achieving better performance in your own program or service." This so-called "modern day benchmarking" began with the Xerox Corporation in 1979 (Omohundro and Albino 1993). Xerox was losing market share to the Japanese in an industry they once dominated. They then compared their machine parts and manufacturing processes with those of the Japanese firms and found they had higher material costs and defection rates than their competitors. They altered their relations with their suppli-

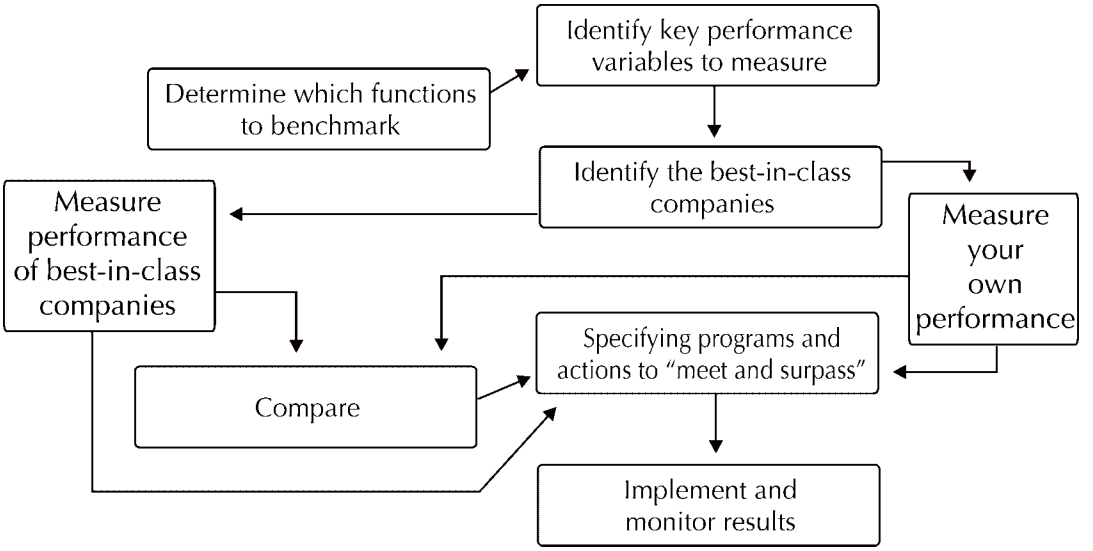


FIGURE 4.2
Seven-Step Process for Benchmarking

Source: Kenneth A. Bruder and Edward M. Gray, "Public Sector Benchmarking: A Practical Approach," *Public Management* 76, no. 9 (September, 1994): S-9–S-14. Reprinted with permission from Kaiser Associates, 2000.

ers and improved the defect rate substantially (Meszaros and Owen 1997, 11), regaining market share. Xerox then compared their warehousing and distribution system to that of L.L. Bean, a firm which is not a competitor but which had a reputation of being best in its class on these functions. This broadens the term "benchmarking" from comparisons with competitors to comparisons with those whose similar functions are considered best in their class.

Kenneth Bruder and Edward Gray (1994) identify seven steps in the public sector's benchmarking process building on the best-in-class notion. This process is displayed in figure 4.2. The first step is to determine which functional areas within your organization will benefit most from benchmarking and which will have the greatest impact. This impact can be in terms of cost, service outcome, room for improvement, or other relevant variable. The second step is to identify the key performance variables to measure cost, quality, and efficiency for the functions you have selected. The third step is to choose the best-in-class organizations for each benchmark item. Note that different organizations can be identified for each item. The idea is to improve *functions* within the organization, thus improving overall quality and effectiveness. The fourth step is to measure the performance of the best-in-class entities for each benchmarked function. You can get these data from published sources, by sharing data, or by interviewing experts on the function.

The fifth step involves measuring your own performance for each benchmarked item and identifying the gaps between your organization's performance and that of the best-in-class organization. This data collection should be a continuous process with responsibility for gathering the data assigned specifically to someone on staff. The sixth is

to specify actions and programs to close the gaps. In this step, you use the data to make organizational changes that result in program improvement. The final step is to implement and monitor your benchmarking results.

The benchmarking process flows from an organization's goals and objectives and is grounded in the desire for continuous improvement of services. The purposes of benchmarking are to "establish the criteria which underlie performance, to identify problem areas within respective services, and to improve service delivery by importing best practices" (Fischer 1994, S-4). Benchmarking is an integral part of any performance monitoring effort.

Conclusion

The Results Act required that ten performance measurement pilot projects be implemented before the government-wide implementation in 1997. In fact, approximately seventy pilots were launched during that time. Joseph Wholey and Kathryn Newcomer (1997) report some lessons learned from the pilots that are instructive for any agencies considering implementing performance-based management systems.

- Top leadership support is clearly the critical element that can make or break strategic planning and performance efforts. The key finding of virtually all observers of performance measurement efforts is that the support of agency leadership is essential to ensure the success of the system.
- The personal involvement of senior line managers is critical. Senior line managers must be involved in the planning process. Performance measures must be relevant and useful to line management.
- Participation of the relevant stakeholders is needed to develop useful plans and performance measures.
- Technical assistance in the design of useful performance measurement systems is often necessary but may not be available when needed. At that point the agencies are on their own to develop the expertise.
- Uncertainty about how performance data will be used will inhibit the design of useful performance measurement systems. The current fervor for budget cuts has sent an important message to managers coping with demands of performance measurement report cards; systematic performance measurement could provide useful data to inform budget cutting. (94–95)

The next question should be "Is there any evidence that the performance measures were used?" Wholey and Newcomer cite evidence that agreement on performance measures have tightened lines of communication about performance expectations in some agencies (e.g., National Highway Traffic Safety Administration, Environmental Protection Agency, Coast Guard); internal reallocation of resources has occurred (e.g., Coast Guard, Army Audit Agency, Inter-American Foundation); heightened awareness among employees about the need to be outcome oriented has been reported (e.g., National Air and Space Administration, Coast Guard, Energy Information Administration); incentives to reward high performers have been introduced (e.g., Housing and Urban Development, Public Health Service) (1997, 96–97).

As the public seeks more and more accountability from all sectors, it is clear that performance-based management is bound to become the rule and not the exception. Thoughtful implementation of performance-based management systems is good for managers, service deliverers, customers, and decision makers.

Acknowledgment

The authors thank Harry Hatry and Joseph Wholey for sharing their performance measurement materials and allowing license to use their work in this chapter.

References

Bruder, Kenneth A., Jr., and Edward M. Gray. 1994. "Public-Sector Benchmarking: A Practical Approach." *Public Management* 76, no. 9 (September): S-9 S-14.

Fischer, Richard J. 1994. "An Overview of Performance Measurement." *Public Management* 76, no. 9 (September): S-2 S-8.

Hatry, Harry P. 1997. "Where the Rubber Meets the Road." In *Using Performance Measurement to Improve Public and Nonprofit Programs. New Directions for Evaluation* 75, edited by Kathryn E. Newcomer. (Fall). San Francisco: Jossey-Bass Publishers,31–44.

Meszaros, Greg, and C. James Owen. 1997. "Improving Through Benchmarking." *Public Works Management & Policy* 2, no. 1 (July): 11–23.

National Performance Review. 1993. *From Red Tape to Results: Creating a Government that Works Better and Costs Less.* Washington, D.C.: Government Printing Office.

Newcomer, Kathryn E. 1997. "Using Performance Measurement to Improve Programs." In *Using Performance Measurement to Improve Public and Nonprofit Programs. New Directions for Evaluation* 75, edited by Kathryn E. Newcomer. (Fall). San Francisco: Jossey-Bass Publishers, 5–14.

Omohundro, L., and J. Albino. 1993. "The Measure of a Utility." *Electric Perspectives* (May/June): 14–26.

Oregon Progress Board and the Governor's Oregon Shines Task Force. 1997. *Oregon Shines II: Updating Oregon's Strategic Plan.* Salem: Oregon Progress Board.

Osborne, David, and Ted Gaebler. 1992. *Reinventing Government.* Reading, Mass.: Addison-Wesley.

Wholey, Joseph S., and Harry P. Hatry. 1992. "The Case for Performance Monitoring." *Public Administration Review* 5, no. 6: 604–10.

Wholey, Joseph S., and Kathryn E. Newcomer. 1997. "Clarifying Goals, Reporting Results." In *Using Performance Measurement to Improve Public and Nonprofit Programs. New Directions for Evaluation* 75, edited by Kathryn E. Newcomer. (Fall). San Francisco: Jossey-Bass Publishers, 91–98.

PART II

Experimental Designs

There are experimental designs and there are experimental designs. A distinction must be made between experimental designs and the quasi-experimental designs that are discussed in Part III. The concern here is with the most powerful and "truly scientific" evaluation design—the controlled, randomized experiment. Essentially three "true" experimental designs can be found in the literature:

(1) the pretest-posttest control group design,

(2) the Solomon Four-Group Design, and

(3) the posttest-only control group design.

Actually, a fourth variation—the factorial design—was discussed in chapter 2.

As Peter Rossi and Howard Freeman (1985, 263) note, randomized experiments (those in which participants in the experiment are selected for participation strictly by chance) are the "flagships" in the field of program evaluation because they allow program personnel to reach conclusions about program impact (or lack of impact) with a high degree of certainty. These evaluations have much in common with experiments in the physical and biological sciences, particularly as they enable the research results to establish causal effects. The findings of randomized experiments are treated with considerable respect by policymakers, program staff, and knowledgeable publics.

The key is randomization, that is, random assignment. True experimental designs always assign subjects to treatment randomly. As long as the number of subjects is sufficiently large, random assignment more or less guarantees that the characteristics of the subjects in the experimental and control groups are statistically equivalent.

As David Nachmias points out, the classical evaluation design consists of four essential features: comparison, manipulation, control, and generalizability (1979, 23–29). To assess the impact of a policy, some form of comparison must be made. Either a comparison is made of an experimental group with a control group or the experimental group is compared with itself or with some selected group before and after treatment. In a true experimental design, the experimental group is compared to a control group.

The second feature of an evaluation design is manipulation. The idea is that if a program or policy is actually effective, the individuals (or cities, or organizations) should change over the time of participation. If we

are able to hold all other factors in the world constant during the evaluation, then the change in policy (manipulation) should cause a change in the target exposed to it (individuals, cities, or organizations).

The third feature of the experimental design—control—requires that other factors be ruled out as explanations of the observed relationship between a policy and its target. These other factors are the well-known sources of internal invalidity discussed in chapter 2: history, maturation, testing, instrumentation, statistical regression, selection, and experimental mortality. As Nachmias points out, these sources of internal invalidity are controlled through randomization (1979, 27–28).

The final essential feature of the classical design is generalizability, or the extent to which research findings can be generalized to larger populations and in different settings. Unfortunately, the mere use of the controlled, randomized experimental design will not in itself control for sources of external invalidity or the lack of generalizability.

The three chapters in this section examine and critique three studies that illustrate the use of the pretest-posttest control group design, the Solomon Four-Group Design, and the posttest-only control group design.

References

Nachmias, David. 1979. *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's.

Rossi, Peter H., and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach.* 3d ed. Beverly Hills, Calif.: Sage.

CHAPTER 5

Pretest-Posttest Control Group Design

The pretest-posttest control group experiment, the classical experimental design, consists of two comparable groups: an experimental group and a control group. Although a number of authors use the terms "control group" and "comparison group" interchangeably, in a strict sense they are not. True control groups are formed by the process of random assignment. Comparison groups are matched to be comparable in important respects to the experimental group. In this book, the distinction between control groups and comparison groups is strictly maintained.

When the pretest-posttest control group design is used, individuals are randomly assigned to one of the two groups. Random assignment of members of a target population to different groups implies that whether an individual (city, organization) is selected for participation is decided purely by chance. Peter Rossi and Howard Freeman elaborate:

> Because the resulting experimental and control groups differ from one another only by chance, whatever processes may be competing with a treatment to produce outcomes are present in the experimental and control groups to the same extent except for chance fluctuations. For example, given randomization, persons who would be more likely to seek out the treatment if it were offered to them on a free-choice basis are equally likely to be in the experimental as in the control group. Hence, both groups have the same proportion of persons favorably predisposed to the intervention. (1985, 235)

But how is randomization accomplished? Randomization is analogous to flipping a coin, with all of those flipping heads being assigned to one group and all of those flipping tails being assigned to another. The most common ways of affecting random assignment are the following:

1. Actually flipping a coin for each subject (allowing all heads to represent one group and tails the other).
2. Throwing all the names into a hat or some other container, thoroughly mixing them, and drawing them out one at a time, allowing odd draws to represent one group and even draws the other.
3. Using a table of random numbers or

random numbers generated by a computer program.

It is important to distinguish between random assignment and random sampling. At first glance, they may appear to be identical, and in some instances they may be identical. Major differences exist, however, between the two techniques. Random sampling ensures representativeness between a sample and the population from which it is drawn. Random selection (sampling) is thus an important factor in the external validity of a study—that is, the extent to which a study's results can be generalized beyond the sample drawn. For example, in the study presented in this chapter, Harrison McKay and colleagues evaluated a program of treatment combining nutrition, health care, and education on the cognitive ability of chronically undernourished children from around the world. Thus, the question is this (assuming that the study itself is reliable): To what degree can the impact of this program conducted in Colombia be generalized to the probable impact of similar programs on other children throughout the world?

Random assignment, as opposed to random selection, is related to the evaluation's internal validity—that is, the extent to which the program's impact is attributed to the treatment and no other factors.

Obviously, the best course would be to select subjects randomly and then to assign them randomly to groups once they were selected. (This discussion has digressed a bit from the discussion of the pretest-posttest control group design, but the digression is important.)

Returning to the design: Subjects are randomly assigned to an experimental group and a control group. To evaluate the effectiveness of the program, measurements are taken twice for each group. A preprogram measure is taken for each group before the introduction of the program to the experimental group. A postprogram measure is then taken after the experimental group has been exposed to (or has completed) the program.

Preprogram scores are then subtracted from postprogram scores. If the gain made by the experimental group is significantly larger than the gain made by the control group, then the researchers can conclude that the program is effective. The pretest-posttest control group design is illustrated in table 5.1. Group E is the experimental group, or the group receiving the program. Group C is the control group. The "O" indicates a test point and the "X" represents the program.

TABLE 5.1
Pretest-Posttest Control Group Design

	Pretest	Program	Posttest
Group E	O	X	O
Group C	O		O

The critical question is this: When should such a design be used? Although it is extremely powerful, this design is also costly and difficult to implement. It is not possible, for example, to withhold treatment purposely from some groups and to assign them randomly to control groups (in matters of life and death, for example). And in many cases, program participants are volunteers; there is no comparable control group (those persons who had the desire and motivation to participate in the program but who did not volunteer). Then there is the matter of cost. Experimental evaluation designs are generally more costly than other designs because of the greater amount of time required to plan and conduct the experiment and the higher level of analytical skills required for planning and undertaking the evaluation and analyzing the results.

This design is frequently used in health or employment programs. A popular, but now somewhat dated, Urban Institute publication, *Practical Program Evaluation for State and Local Governments*, documents conditions under which such experimental designs are likely to be appropriate for state and local governments (Hatry, Winnie, and Fisk 1981,

107–15). The more significant conditions include the following:

1. *There is likely to be a high degree of ambiguity as to whether outcomes were caused by the program if some other evaluation design is used.* The design is appropriate when the findings obtained through the use of a less powerful design may be criticized for not causing the results. Take a hypothetical example of a medical experiment involving a cold remedy. If no control group was used, would not critics of the experiment ask, "How do we know that the subject would not have recovered from the cold in the same amount of time without the pill?"
2. *Some citizens can be given different services from others without significant danger or harm.* The experimental design may be used if public officials and the evaluators agree that the withdrawn or nonprovision of a service or program is not likely to have harmful effects. An example might be discontinuing evening hours at a local branch library, which is not likely to harm many individuals.
3. *Some citizens can be given different services from others without violating moral and ethical standards.* Some programs, although not involving physical danger, may call for not providing services to some groups. For example, an experiment to assess the effectiveness of a counseling program for parolees might be designed in such a way that certain parolees do not receive counseling (and thus might be more likely to commit a crime and be returned to prison). This could be seen by some as unethical or immoral.
4. *There is substantial doubt about the effectiveness of a program.* If a program is not believed to be working or effective, controlled, randomized experimentation is probably the only way to settle the issue once and for all.
5. *There are insufficient resources to provide the program to all clients.* Even when a program is expected to be helpful, the resources necessary to provide it to all eligible clients may not be available. In the article in this chapter, McKay and colleagues did not have sufficient financial resources to provide the program to all chronically undernourished children in Cali, Colombia. They thus were able to evaluate the program by comparing children who had received the program with those who had not—even though it would have been desirable to provide the program to all children.
6. *The risk in funding the program without a controlled experiment is likely to be substantially greater than the cost of the experiment; the new program involves large costs and a large degree of uncertainty.* The income maintenance experiments described in chapter 2 were designed to test a new form of welfare payment. These evaluations are among the most expensive ever funded by the federal government. Yet the millions of dollars spent on the experiment are insignificant when compared to the cost of a nationwide program that did not provide the desired results.
7. *A decision to implement the program can be postponed until the experiment is completed.* Most experiments take a long time—a year or more. If there is considerable pressure (usually political) to fully implement a program, experimentation may be difficult to apply.

What all this means is that there are probably many occasions when an experimental design is not appropriate. In contrast, there are times when such a design is clearly needed. The following article, "Improving Cognitive Ability in Chronically Deprived Children," is an example of the pretest-posttest control group design.

READING

Improving Cognitive Ability in Chronically Deprived Children

Harrison McKay • Leonardo Sinisterra • Arlene McKay
Hernando Gomez • Pascuala Lloreda

IN RECENT YEARS, social and economic planning in developing countries has included closer attention than before to the nutrition, health, and education of children of preschool age in low-income families. One basis for this, in addition to mortality and morbidity studies indicating high vulnerability at that age, (*1*) is information suggesting that obstacles to normal development in the first years of life, found in environments of such poverty that physical growth is retarded through malnutrition, are likely also to retard intellectual development permanently if early remedial action is not taken (*2*). The loss of intellectual capability, broadly defined, is viewed as especially serious because the technological character of contemporary civilization makes individual productivity and personal fulfillment increasingly contingent upon such capability. In tropical and subtropical zones of the world between 220 and 250 million children below 6 years of age live in conditions of environmental deprivation extreme enough to produce some degree of malnutrition (*3*); failure to act could result in irretrievable loss of future human capacity on a massive scale.

Although this argument finds widespread agreement among scientists and planners, there is uncertainty about the effectiveness of specific remedial actions. Doubts have been growing for the past decade about whether providing food, education, or health care directly to young children in poverty environments can counteract the myriad social, economic, and biological limitations to their intellectual growth. Up to 1970, when the study reported here was formulated, no definitive evidence was available to show that food and health care provided to malnourished or "at risk" infants and young children could produce lasting increases in intellectual functioning. This was so in spite of the ample experience of medical specialists throughout the tropical world that malnourished children typically responded to nutritional recuperation by being more active physically, more able to assimilate environmental events, happier, and more verbal, all of which would be hypothesized to create a more favorable outlook for their capacity to learn (*4*).

In conferences and publications emphasis was increasingly placed upon the inextricable relation of malnutrition to other environmental factors inhibiting full mental development of preschool age children in poverty environments (*5*). It was becoming clear that, at least after the period of rapid brain growth in the first 2 years of life, when protein-calorie malnutrition could have its maximum deleterious physiological effects (*6*), nutritional rehabilitation and health care programs should be accompanied by some form of environmental modification of children at risk. The largest amount of available information about the potential effects of environmental modification among children from poor families pertained to the United States, where poverty was not of such severity as to make malnutrition a health issue of marked proportions. Here a large literature showed that the low intellectual performance found among disadvantaged children was environmentally based and probably was largely fixed during the preschool years (*7*). This information gave impetus to the belief that direct treatments, carefully designed and properly delivered to children dur-

ing early critical periods, could produce large and lasting increases in intellectual ability. As a consequence, during the 1960s a wide variety of individual, research-based preschool programs as well as a national program were developed in the United States for children from low-income families (*8*). Several showed positive results but in the aggregate they were not as great or as lasting as had been hoped, and there followed a widespread questioning of the effectiveness of early childhood education as a means of permanently improving intellectual ability among disadvantaged children on a large scale (*9*).

From pilot work leading up to the study reported here, we concluded that there was an essential issue that had not received adequate attention and the clarification of which might have tempered the pessimism: the relation of gains in intellectual ability to the intensity and duration of meliorative treatment received during different periods in the preschool years. In addition to the qualitative question of what kinds of preschool intervention, if any, are effective, attention should have been given to the question of what amount of treatment yields what amount of gain. We hypothesized that the increments in intellectual ability produced in preschool programs for disadvantaged children were subsequently lost at least in part because the programs were too brief. Although there was a consensus that longer and more intensive preschool experience could produce larger and more lasting increases, in only one study was there to be found a direct attempt to test this, and in that one sampling problems caused difficulties in interpretation (*10*).

As a consequence, the study reported here was designed to examine the quantitative question, with chronically undernourished children, by systematically increasing the duration of multidisciplinary treatments to levels not previously reported and evaluating results with measures directly comparable across all levels (*11*). This was done not only to test the hypothesis that greater amounts of treatment could produce greater and more enduring intellectual gains but also to develop for the first time an appraisal of what results could be expected at different points along a continuum of action. This second objective, in addition to its intrinsic scientific interest, was projected to have another benefit: that of being useful in the practical application of early childhood services. Also unique in the study design was the simultaneous combination of health, nutrition, and educational components in the treatment program. With the exception of our own pilot work (*12*), prior studies of preschool nutritional recuperation programs had not included educational activities. Likewise, preschool education studies had not included nutritional recuperation activities, because malnutrition of the degree found in the developing countries was not characteristic of disadvantaged groups studied in the United States (*13*), where most of the modern early-education research had been done.

Experimental Design and Subjects

The study was carried out in Cali, Colombia, a city of nearly a million people with many problems characteristic of rapidly expanding cities in developing countries, including large numbers of families living in marginal economic conditions. Table 1 summarizes the experimental design employed. The total time available for the experiment was 3˚ years, from February 1971 to August 1974. This was divided into four treatment periods of 9 months each plus interperiod recesses. Our decision to begin the study with children as close as possible to 3 years of age was based upon the 2 years of pilot studies in which treatment and measurement systems were developed for children starting at that age (*14*). The projected 180 to 200 days of possible attendance at treatment made each projected period similar in length to a school

TABLE 1
Basic selection and treatment variables of the groups of children in the study.

Group	N		Characteristic
	In 1971	In 1975	
T1(a)	57	49	Low SES, subnormal weight and height. One treatment period, between November 1973 and August 1974 (75 to 84 months of age)[a]
T1(b)	56	47	Low SES, subnormal weight and height. One treatment period, between November 1973 and August 1974 (75 to 84 months of age), with prior nutritional supplementation and health care
T2	64	51	Low SES, subnormal weight and height. Two treatment periods, between November 1972 and August 1974 (63 to 84 months of age)
T3	62	50	Low SES, subnormal weight and height. Three treatment periods, between December 1971 and August 1974 (52 to 84 months of age)
T4	62	51	Low SES, subnormal weight and height. Four treatment periods, between February 1971 and August 1974 (42 to 84 months of age)
HS	38	30	High SES. Untreated, but measured at the same points as groups T1–T4
T0	116	72	Low SES, normal weight and height. Untreated

a. SES is family socioeconomic status.

year in Colombia, and the end of the fourth period was scheduled to coincide with the beginning of the year in which the children were of eligible age to enter first grade.

With the object of having 60 children initially available for each treatment group (in case many should be lost to the study during the 3° year period), approximately 7500 families living in two of the city's lowest-income areas were visited to locate and identify all children with birth dates between 1 June and 30 November 1967, birth dates that would satisfy primary school entry requirements in 1974. In a second visit to the 733 families with such children, invitations were extended to have the children medically examined. The families of 518 accepted, and each child received a clinical examination, anthropometric measurement, and screening for serious neurological dysfunctions. During a third visit to these families, interviews and observations were conducted to determine living conditions, economic resources, and age, education, and occupations of family members. At this stage the number of potential subjects was reduced by 69 (to 449), because of errors in birth date, serious neurological or sensory dysfunctions, refusal to participate further, or removal from the area.

Because the subject loss due to emigration during the 4 months of preliminary data gathering was substantial, 333 children were selected to assure the participation of 300 at the beginning of treatment; 301 were still available at that time, 53 percent of them male. Children selected for the experiment from among the 449 candidates were those having, first, the lowest height and weight for age; second, the highest number of clinical signs of malnutrition (*15*); and third, the lowest per capita family income. The second and third criteria were employed only in those regions of the frequency distributions where differences among the children in height and weight for age were judged by the medical staff to lack biological significance. Figure 1 shows these frequency distributions

and includes scales corresponding to percentiles in a normal population (*16*).

The 116 children not selected were left untreated and were not measured again until 4 years later, at which point the 72 still living in the area and willing once again to collaborate were reincorporated into the longitudinal study and measured on physical growth and cognitive development at the same time as the selected children, beginning at 7 years of age. At 3 years of age these children did not show abnormally low weight for age or weight for height.

In order to have available a set of local reference standards for "normal" physical and psychological development, and not depend solely upon foreign standards, a group of children (group HS) from families with high socioeconomic status, living in the same city and having the same range of birth dates as the experimental group, was included in the study. Our assumption was that, in regard to available economic resources, housing, food, health care, and educational opportunities, these children had the highest probability of full intellectual and physical development of any group in the society. In relation to the research program they remained untreated, receiving only medical and psychological assessment at the same intervals as the treated children, but the majority were attending the best private preschools during the study. Eventually 63 children were recruited for group HS, but only the 38 noted in Table 1 were available at the first psychological testing session in 1971.

Nearly all the 333 children selected for treatment lived in homes distributed throughout an area of approximately 2 square kilometers. This area was subdivided into 20 sectors in such a way that between 13 and 19 children were included in each sector. The sectors were ranked in order of a standardized combination of average height and weight for age and per capita family income of the children. Each of the first five sectors in the ranking was assigned randomly to one of five groups. This procedure was followed for the next three sets of five sectors, yielding four sectors for each group, one from each of four strata. At this point the groups remained unnamed; only as each new treatment period was to begin was a group assigned to it and families in the sectors chosen so informed. The children were assigned by sectors instead of individually in order to minimize social interaction between families in different treatment groups and to make daily transportation more efficient (*17*). Because this "lottery" system was geographically based, all selected children who were living in a sector immediately prior to its assignment were included in the assignment and remained in the same treatment group regardless of further moves. In view of this process, it must be noted that the 1971 N's reported for the treatment groups in Table 1 are retrospective figures, based upon a count of children then living in sectors assigned later to treatment groups. Table 1 also shows the subject loss, by group, between 1971 and 1975. The loss of 53 children—18 percent—from the treatment groups over 4 years was considerably less than expected. Two of these children died and 51 emigrated from Cali with their families; on selection variables they did not differ to a statistically significant degree from the 248 remaining.

A longitudinal study was begun, then, with groups representing extreme points on continua of many factors related to intellectual development, and with an experimental plan to measure the degree to which children at the lower extreme could be moved closer to those of the upper extreme as a result of combined treatments of varying durations. Table 2 compares selected (T1-T4), not selected (T0), and reference (HS) groups on some of the related factors, including those used for selecting children for participation in treatment.

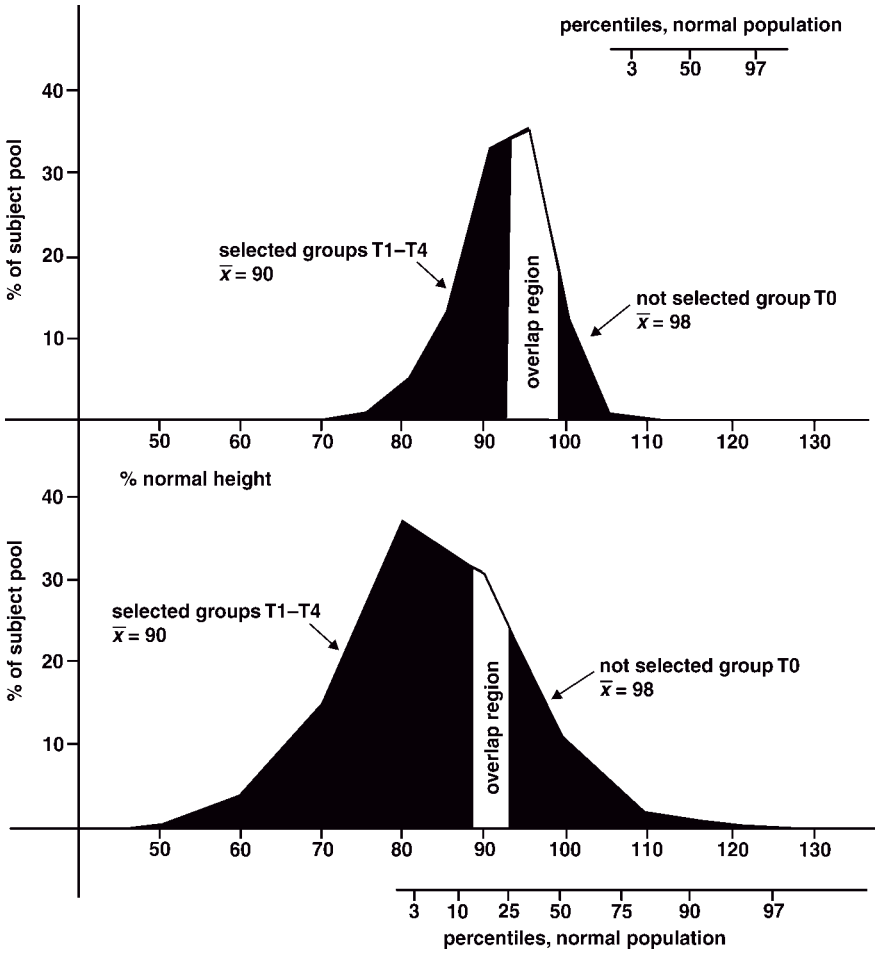


FIGURE 1

Frequency distributions of height and weight (as percent of normal for age) of the subject pool of 449 children available in 1970, from among whom 333 were selected for treatment groups. A combination of height and weight was the first criterion; the second and third criteria, applied to children in the overlap regions, were clinical signs of malnutrition and family income. Two classification systems for childhood malnutrition yield the following description of the selected children: 90 percent nutritionally "stunted" at 3 years of age and 35 percent with evidence of "wasting"; 26 percent with "second degree" malnutrition, 54 percent with "first degree," and 16 percent "low normal." (16)

Treatments

The total number of treatment days per period varied as follows: period 1, 180 days; period 2, 185; period 3, 190; period 4, 172. A fire early in period 4 reduced the time available owing to the necessity of terminating the study before the opening of primary school. The original objective was to have each succeeding period at least as long as the preceding one in order to avoid reduction in intensity of treatment. The programs occupied 6 hours a day 5 days a week, and attendance was above 95 percent for all groups; hence there were approximately 1040, 1060, 1080, and 990 hours of treatment per child per period from period 1 to period 4, respectively. The total number of hours of treatment per group, then, were as follows: T4, 4170 hours;

T3, 3130 hours; T2, 2070 hours; T1 (a and b), 990 hours.

In as many respects as possible, treatments were made equivalent between groups within each period. New people, selected and trained as child-care workers to accommodate the periodic increases in numbers of children, were combined with existing personnel and distributed in such a way that experience, skill, and familiarity with children already treated were equalized for all groups, as was the adult-child ratio. Similarly, as new program sites were added, children rotated among them so that all groups occupied all sites equal lengths of time. Except for special care given to the health and nutritional adaptation of each newly entering group during the initial weeks, the same systems in these treatments were applied to all children within periods.

An average treatment day consisted of 6 hours of integrated health, nutritional, and educational activities, in which approximately 4 hours were devoted to education and 2 hours to health, nutrition, and hygiene. In practice, the nutrition and health care provided opportunities to reinforce many aspects of the educational curriculum, and time in the education program was used to reinforce recommended hygienic and food consumption practices.

The nutritional supplementation program was designed to provide a minimum of 75 percent of recommended daily protein and calorie allowances, by means of low-cost foods available commercially, supplemented with vitamins and minerals, and offered ad libitum three times a day. In the vitamin and mineral supplementation, special attention was given to vitamin A, thiamin, riboflavin, niacin, and iron, of which at least 100 percent of recommended dietary allowance was provided (*18*).

The health care program included daily observation of all children attending the treatment center, with immediate pediatric attention to those with symptoms reported by the parents or noted by the health and education personnel. Children suspected of an infectious condition were not brought into contact with their classmates until the danger of contagion had passed. Severe health problems occurring during weekends or holidays were attended on an emergency basis in the local university hospital.

The educational treatment was designed to develop cognitive processes and language, social abilities, and psychomotor skills, by means of an integrated curriculum model. It was a combination of elements developed in pilot studies and adapted from other programs known to have demonstrated positive effects upon cognitive development (*19*). Adapting to developmental changes in the children, its form progressed from a structured day divided among six to eight different directed activities, to one with more time available for individual projects. This latter form, while including activities planned to introduce new concepts, stimulate verbal expression, and develop motor skills, stressed increasing experimentation and decision taking by the children. As with the nutrition and health treatments during the first weeks of each new period, the newly entering children received special care in order to facilitate their adaptation and to teach the basic skills necessary for them to participate in the program. Each new period was conceptually more complex than the preceding one, the last ones incorporating more formal reading, writing, and number work.

Measures of Cognitive Development

There were five measurement points in the course of the study: (i) at the beginning of the first treatment period; (ii) at the end of the first treatment period; (iii) after the end of the second period, carrying over into the beginning of the third; (iv) after the end of the third period, extending into the fourth; and (v) following the fourth treatment pe-

TABLE 2

Selection variables and family characteristics of study groups in 1970 (means). All differences between group HS and groups T1–T4 are statistically significant ($P < .01$) except age of parents. There are no statistically significant differences among groups T1–T4. There are statistically significant differences between group T0 and combined groups T1–T4 in height and weight (as percent of normal), per capita income and food expenditure, number of family members and children, and rooms per child; and between group T0 and group HS on all variables except age of parents and weight.

	Group		
Variable	**T1–T4**	**T0**	**HS**
Height as percent of normal for age	90	98	101
Weight as percent of normal for age	79	98	102
Per capita family income as percent of group HS	5	7	100
Per capita food expenditure in family as percent of group HS	15	22	100
Number of family members	7.4	6.4	4.7
Number of family under 15 years of age	4.8	3.8	2.4
Number of play/sleep rooms per child	0.3	0.5	1.6
Age of father	37	37	37
Age of mother	31	32	31
Years of schooling, father	3.6	3.7	14.5
Years of schooling, mother	3.5	3.3	10.0

riod. For the purpose of measuring the impact of treatment upon separate components of cognitive development, several short tests were employed at each measurement point, rather than a single intelligence test. The tests varied from point to point, as those only applicable at younger ages were replaced by others that could be continued into primary school years. At all points the plan was to have tests that theoretically measured adequacy of language usage, immediate memory, manual dexterity and motor control, information and vocabulary, quantitative concepts, spatial relations, and logical thinking, with a balance between verbal and nonverbal production. Table 3 is a list of tests applied at each measurement point. More were applied than are listed; only those employed at two or more measurement points and having items that fulfilled the criteria for the analysis described below are included.

Testing was done by laypersons trained and supervised by professional psychologists. Each new test underwent a 4 to 8 month developmental sequence which included an initial practice phase to familiarize the examiners with the format of the test and possible difficulties in application. Thereafter, a series of pilot studies were conducted to permit the modification of items in order to attain acceptable levels of difficulty, reliability, and ease of application. Before each measurement point, all tests were applied to children not in the study until adequate inter-tester reliability and standardization of application were obtained. After definitive application at each measurement point, all tests were repeated on a 10 percent sample to evaluate test-retest reliability. To protect against examiner biases, the children were assigned to examiners randomly and no information was provided regarding treatment group or nutritional or socioeconomic level. (The identification of group HS children was, however, unavoidable even in the earliest years, not only because of their dress and speech but also because of the differences in their interpersonal behavior.) Finally, in order to prevent children from being trained specifically to perform well on test items, the two functions of intervention and evaluation were separated as far as possible. We intention-

ally avoided, in the education programs, the use of materials or objects from the psychological tests. Also, the intervention personnel had no knowledge of test content or format, and neither they nor the testing personnel were provided with information about group performance at any of the measurement points.

Data Analysis

The data matrix of cognitive measures generated during the 44-month interval between the first and last measurement points entailed evaluation across several occasions by means of a multivariate vector of observations. A major problem in the evaluation procedure, as seen in Table 3, is that the tests of cognitive development were not the same at every measurement point. Thus the response vector was not the same along the time dimension. Initially, a principal component approach was used, with factor scores representing the latent variables (*20*). Although this was eventually discarded because there was no guarantee of factor invariance across occasions, the results were very similar to those yielded by the analyses finally adopted for this article. An important consequence of these analyses was the finding that nearly all of the variation could be explained by the first component (*21*), and under the assumption of unidimensionality cognitive test items were pooled and calibrated according to the psychometric model proposed by Rasch (*22*) and implemented computationally by Wright (*23*). The technique employed to obtain the ability estimates in Table 4 guarantees that the same latent trait is being reflected in these estimates (*24*). Consequently, the growth curves in Fig. 2 are interpreted as representing "general cognitive ability" (*25*).

Table 5 shows correlations between pairs of measurement points of the ability estimates of all children included in the two points. The correspondence is substantial, and the matrix exhibits the "simplex" pattern expected in psychometric data of this sort (*26*). As the correlations are not homogeneous, a test for diagonality in the transformed error covariance matrix was carried out, and the resulting chi-square value led to rejection of a mixed model assumption. In view of this, Bock's multivariate procedure (*27*), which does not require constant correlations, was employed to analyze the differences among groups across measurement points. The results showed a significant groups-by-occasions effect, permitting rejection of the hypothesis of parallel profiles among groups. A single degree-of-freedom decomposition of this effect showed that there were significant differences in every possible Helmert contrast. Stepdown tests indicated that all components were required in describing profile differences.

TABLE 3
Tests of cognitive ability applied at different measurement points (see text) between 43 and 87 months of age. Only tests that were applied at two adjacent points and that provided items for the analysis in Table 4 are included. The unreferenced tests were constructed locally.

Test	Measurement points
Understanding complex commands	1,2
Figure tracing	1, 2, 3
Picture vocabulary	1, 2, 3
Intersensory perception (*33*)	1, 2, 3
Colors, numbers, letters	1, 2, 3
Use of prepositions	1, 2, 3
Block construction	1, 2, 3
Cognitive maturity (*34*)	1, 2, 3, 4
Sentence completion (*35*)	1, 2, 3, 4
Memory for sentences (*34*)	1, 2, 3, 4, 5
Knox cubes (*36*)	1, 2, 3, 4, 5
Geometric drawings (*37*)	3, 4
Arithmetic (*38, 39*)	3, 4, 5
Mazes (*40*)	3, 4, 5
Information (*41*)	3, 4, 5
Vocabulary (*39*)	3, 4, 5
Block design (*42*)	4, 5
Digit memory (*43*)	4, 5
Analogies and similarities (*44*)	4, 5
Matrices (*45*)	4, 5
Visual classification	4, 5

TABLE 4

Scaled scores on general cognitive ability, means and estimated standard errors, of the four treatment groups and group HS at five testing points.

Group	*N*	Average age at testing (months)				
		43	**49**	**63**	**77**	**87**
			Mean score			
HS	28	–0.11	.39	2.28	4.27	4.89
T4	50	–1.82[a]	.21	1.80	3.35	3.66
T3	47	–1.72	–1.06	1.64	3.06	3.35
T2	49	–1.94	–1.22	.30[b]	2.61	3.15
T1	90	–1.83	–1.11	.33	2.07	2.73
			Estimated standard error			
HS	28	.192	.196	.166	.191	.198
T4	50	.225	.148	.138	.164	.152
T3	47	.161	.136	.103	.123	.120
T2	49	.131	.132	.115	.133	.125
T1	90	.110	.097	.098	.124	.108
			Standard deviation			
All groups		1.161	1.153	1.169	1.263	1.164

a. Calculated from 42 percent sample tested prior to beginning of treatment.
b. Calculated from 50 percent sample tested prior to beginning of treatment.

The data in Table 4, plotted in Fig. 2 with the addition of dates and duration of treatment periods, are based upon the same children at all measurement points. These are children having complete medical, socioeconomic, and psychological test records. The discrepancies between the 1975 *N*'s in Table 1 and the *N*'s in Table 4 are due to the fact that 14 children who were still participating in the study in 1975 were excluded from the analysis because at least one piece of information was missing, a move made to facilitate correlational analyses. Between 2 percent (T4) and 7 percent (HS) were excluded for this reason.

TABLE 5

Correlation of ability scores across measurement points

Measurement points	1	2	3	4	5
1	—	.78	.68	.54	.48
2		—	.80	.66	.59
3			—	.71	.69
4				—	.76
5					—

For all analyses, groups T1(a) and T1(b) were combined into group T1 because the prior nutritional supplementation and health care provided group T1(b) had not been found to produce any difference between the two groups. Finally, analysis by sex is not included because a statistically significant difference was found at only one of the five measurement points.

Relation of Gains to Treatment

The most important data in Table 4 and Fig. 2 are those pertaining to cognitive ability scores at the fifth testing point. The upward progression of mean scores from T1 to T4 and the nonoverlapping standard errors, except between T2 and T3, generally confirm that the sooner the treatment was begun the higher the level of general cognitive ability reached by age 87 months. Another interpretation of the data could be that the age at which treatment began was a determining factor independent of amount of time in treatment.

It can be argued that the level of cognitive development which the children reached at 7

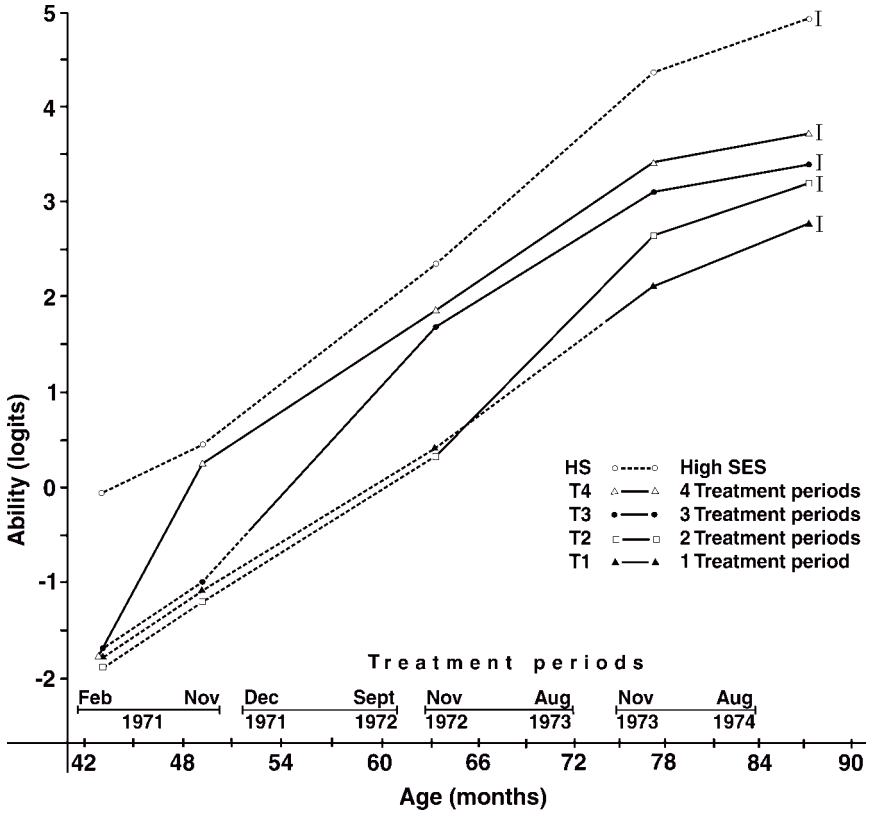


FIGURE 2

Growth of general cognitive ability of the children from age 43 months to 87 months, the age at the beginning of primary school. Ability scores are scaled sums of test items correct among items common to proximate testing points. The solid lines represent periods of participation in a treatment sequence, and brackets to the right of the curves indicate ± 1 standard error of the corresponding group means at the fifth measurement point. At the fourth measurement point there are no overlapping standard errors; at earlier measurement points there is overlap only among obviously adjacent groups (see Table 4). Group T0 was tested at the fifth measurement point but is not represented in this figure, or in Table 2, because its observed low level of performance could have been attributed to the fact that this was the first testing experience of the group T0 children since the neurological screening 4 years earlier.

years of age depended upon the magnitude of gains achieved during the first treatment period in which they participated, perhaps within the first 6 months, although the confounding of age and treatment duration in the experimental design prohibits conclusive testing of the hypothesis. The data supporting this are in the declining magnitude of gain, during the first period of treatment attended, at progressively higher ages of entry into the program. Using group T1 as an untreated baseline until it first entered treatment, and calculating the difference in gains (28) between it and groups T4, T3, and T2 during their respective first periods of treatment, we obtain the following values: group T4, 1.31; group T3, 1.26; and group T2, .57. When calculated as gains per month between testing periods, the data are the following: T4, .22; T3, .09; and T2, .04. This suggests an exponential relationship. Although, because of unequal intervals between testing points and the overlapping of testing durations with treatment periods, this latter relationship must be

viewed with caution, it is clear that the older the children were upon entry into the treatment programs the less was their gain in cognitive development in the first 9 months of participation relative to an untreated baseline.

The lack of a randomly assigned, untreated control group prevents similar quantification of the response of group T1 to its one treatment period. If group HS is taken as the baseline, the observed gain of T1 is very small. The proportion of the gap between group HS and group T1 that was closed during the fourth treatment period was 2 percent, whereas in the initial treatment period of each of the other groups the percentages were group T4, 89; group T3, 55; and group T2, 16. That the progressively declining responsiveness at later ages extends to group T1 can be seen additionally in the percentages of gap closed between group T4 and the other groups during the first treatment period of each of the latter: group T3, 87; group T2, 51; and group T1, 27.

Durability of Gains

Analysis of items common to testing points five and beyond has yet to be done, but the data contained in Fig. 3, Stanford-Binet intelligence quotients at 8 years of age, show that the relative positions of the groups at age 7 appear to have been maintained to the end of the first year of primary school. Although the treated groups all differ from each other in the expected direction, generally the differences are not statistically significant unless one group has had two treatment periods more than another. A surprising result of the Stanford-Binet testing was that group T0 children, the seemingly more favored among the low-income community (see Table 2), showed such low intelligence quotients; the highest score in group T0 (IQ = 100) was below the mean of group HS, and the lowest group HS score (IQ = 84) was above the mean of group T0. This further confirms that the obstacles to normal intellectual growth found in conditions of poverty in which live large segments of the population are very strong. It is possible that this result is due partly to differential testing histories, despite the fact that group T0 had participated in the full testing program at the preceding fifth measurement point, and that this was the first Stanford-Binet testing for the entire group of subject children.

The difference between groups T0 and T1 is in the direction of superiority of group T1 ($t = 1.507$, $P < .10$). What the IQ of group T1 would have been without its one treatment period is not possible to determine except indirectly through regression analyses with other variables, but we would expect it to have been lower than T0's, because T0 was significantly above T1 on socioeconomic and anthropometric correlates of IQ (Table 2). Also, T1 was approximately .30 standard deviation below T0 at 38 months of age on a cognitive development factor of a preliminary neurological screening test applied in 1970, prior to selection. Given these data and the fact that at 96 months of age there is a difference favoring group T1 that approaches statistical significance, we conclude not only that group T1 children increased in cognitive ability as a result of their one treatment period (although very little compared to the other groups) but also that they retained the increase through the first year of primary school.

An interesting and potentially important characteristic of the curves in Fig. 3 is the apparent increasing bimodality of the distribution of the groups with increasing length of treatment, in addition to higher means and upward movement of both extremes. The relatively small sample sizes and the fact that these results were found only once make it hazardous to look upon them as definitive. However, the progression across groups is quite uniform and suggests that the issue of individual differential response to equivalent treatment should be studied more carefully.

Social Significance of Gains

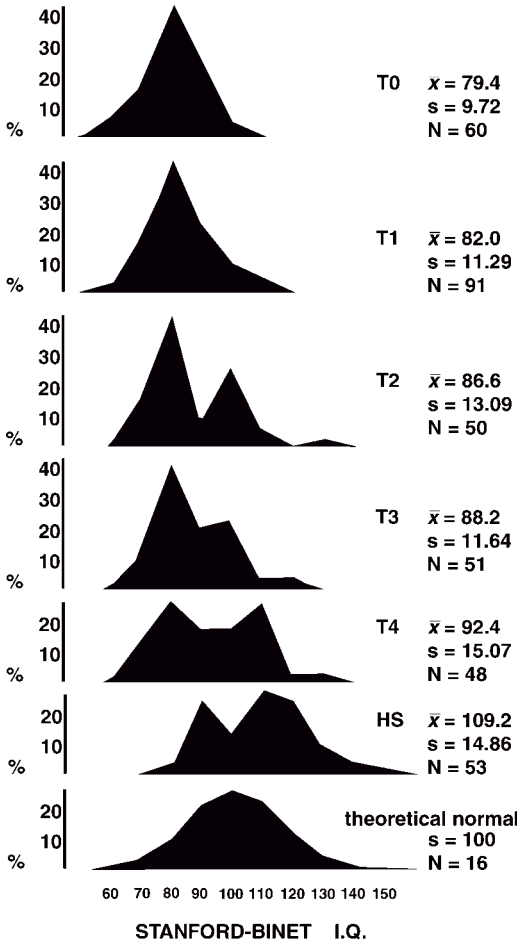
Group HS was included in the study for the purpose of establishing a baseline indicating what could be expected of children when conditions for growth and development were optimal. In this way the effectiveness of the treatment could be evaluated from a frame of reference of the social ideal. It can be seen in Table 4 that group HS increased in cognitive ability at a rate greater than the baseline group T1 during the 34 months before T1 entered treatment. This is equivalent to, and confirms, the previously reported relative decline in intelligence among disadvantaged children (29, p. 258). Between the ages of 4 and 6 years, group HS children passed through a period of accelerated development that greatly increased the distance between them and all the treatment groups. The result, at age 77 months, was that group T4 arrived at a point approximately 58 percent of the distance between group HS and the untreated baseline group T1, group T3 arrived at 45 percent, and group T2 at 24 percent. Between 77 and 87 months, however, these differences appear to have diminished, even taking into account that group T1 entered treatment during this period. In order for these percentages to have been maintained, the baseline would have had to remain essentially unchanged. With respect to overall gains from 43 months to 87 months, the data show that reduction of the 1.5 standard deviation gap found at 43 months of age between group HS and the treated children required a duration and intensity of treatment at least equal to that of group T2; the group HS overall growth of 5.00 units of ability is less than that of all groups except T1.

As noted, group HS was not representative of the general population, but was a sample of children intentionally chosen from a subgroup above average in the society in characteristics favorable to general cognitive development. For the population under study, normative data do not exist; the "theoretical normal" distribution shown in Fig. 3 represents the U.S. standardization group of 1937 (30). As a consequence, the degree to which the treatments were effective in closing the gap between the disadvantaged children and what could be de-

FIGURE 3

Mean scores on the Stanford-Binet Intelligence Test at 8 years of age. Groups T0–T4 had had one year of primary school. Group HS children had attended preschool and primary schools for up to five consecutive years prior to this testing point. Mental age minus chronological age is as follows:

Group T0	- 18 months
Group T1	- 15 months
Group T2	- 11 months
Group T3	- 9 months
Group T4	- 5 months
Group HS	+ 10 months



scribed as an acceptable level cannot be judged. It is conceivable that group HS children were developing at a rate superior to that of this hypothetical normal. If that was the case, the gains of the treated children could be viewed even more positively.

Recent studies of preschool programs have raised the question whether differences between standard intellectual performance and that encountered in disadvantaged children represent real deficits or whether they reflect cultural or ethnic uniquenesses. This is a particularly relevant issue where disadvantaged groups are ethnically and linguistically distinct from the dominant culture (29, pp. 262–72; 31). The historical evolution of differences in intellectual ability found between groups throughout the world is doubtless multidimensional, with circumstances unique to each society or region, in which have entered religious, biological, and other factors in different epochs, and thus the simple dichotomy of culture uniqueness versus deprivation is only a first approximation to a sorely needed, thorough analysis of antecedents and correlates of the variations. Within the limits of the dichotomy, however, the evidence with regard to the children in our study suggests that the large differences in cognitive ability found between the reference group and the treated groups in 1971 should be considered as reflecting deficits rather than divergent ethnic identities. Spanish was the language spoken in all the homes, with the addition of a second language in some group HS families. All the children were born in the same city sharing the same communication media and popular culture and for the most part the same religion. Additionally, on tests designed to maximize the performance of the children from low-income families by the use of objects, words, and events typical in their neighborhoods (for example, a horse-drawn cart in the picture vocabulary test), the difference between them and group HS was still approximately 1.50 standard deviations at 43 months of age. Thus it is possible to conclude that the treated children's increases in cognitive ability are relevant to them in their immediate community as well as to the ideal represented by the high-status reference group. This will be more precisely assessed in future analyses of the relation of cognitive gains to achievement in primary school.

Conclusions

The results leave little doubt that environmental deprivation of a degree severe enough to produce chronic undernutrition manifested primarily by stunting strongly retards general cognitive development, and that the retardation is less amenable to modification with increasing age. The study shows that combined nutritional, health, and educational treatments between 3° and 7 years of age can prevent large losses of potential cognitive ability, with significantly greater effect the earlier the treatments begin. As little as 9 months of treatment prior to primary school entry appears to produce significant increases in ability, although small compared to the gains of children receiving treatment lasting two, three, and four times as long. Continued study will be necessary to ascertain the long-range durability of the treatment effects, but the present data show that they persist at 8 years of age.

The increases in general cognitive ability produced by the multiform preschool interventions are socially significant in that they reduce the large intelligence gap between children from severely deprived environments and those from favored environments, although the extent to which any given amount of intervention might be beneficial to wider societal development is uncertain (32). Extrapolated to the large number of children throughout the world who spend their first years in poverty and hunger, however, even the smallest increment resulting from one 9-month treatment period could constitute an

important improvement in the pool of human capabilities available to a given society.

References and Notes

1. D. B. Jelliffe, *Infant Nutrition in the Tropics and Subtropics* (World Health Organization, Geneva, 1968); C. D. Williams, in *Preschool Child Malnutrition* (National Academy of Sciences, Washington, D. C., 1966), pp. 3–8; N. S. Scrimshaw, in ibid., pp. 63–73.
2. N. S. Scrimshaw and J. E. Gordon, Eds., *Malnutrition, Learning and Behavior* (MIT Press, Boston, 1968), especially the articles by J. Cravioto and E. R. DeLicardie, F. Monckeberg, and M. B. Stoch and P. M. Smythe; J. Carvioto, in *Preschool Child Malnutrition* (National Academy of Sciences, Washington, D.C., 1966), pp. 74–84; M. Winick and P. Rosso, *Pediatr. Res.* **3**, 181 (1969); L. M. Brockman and H. N. Ricciuti, *Dev. Psychol.* **4**, 312 (1971). As significant as the human studies was the animal research of R. Barnes, J. Cowley, and S. Franková, all of whom summarize their work in Malnutrition, Learning and Behavior.
3. A. Berg, *The Nutrition Factor* (Brookings Institution, Washington, D.C., 1973), p. 5.
4. In addition to the authors' own experience, that of pediatricians and nutrition specialists in Latin America, Africa, and Asia is highly uniform in this respect. In fact, a generally accepted rule in medical care of malnourished children is that improvement in psychological response is the first sign of recovery.
5. D. Kallen, Ed., *Nutrition, Development and Social Behavior*. DHEW Publication No. (NIH) 73–242 (National Institutes of Health, Washington, D.C., 1973); S. L. Manocha, *Malnutrition and Retarded Human Development* (Thomas, Springfield, Ill., 1972). pp. 132–165.
6. J. Dobbing, in *Malnutrition, Learning and Behavior*, N. S. Scrimshaw and J. E. Gordon, Eds. (MIT Press, Boston, 1968), p. 181.
7. B. S. Bloom, *Stability and Change in Human Characteristics* (Wiley, New York, 1964); J. McV. Hunt, *Intelligence and Experience* (Ronald, New York, 1961); M. Deutsch et al., *The Disadvantaged Child* (Basic Books, New York, 1967). As with nutrition studies, there was also a large amount of animal research on early experience, such as that of J. Denenberg and G. Karas [*Science* **130**, 629 (1959)] showing the vulnerability of the young organism and that early environmental effects could last into adulthood.
8. M. Pines, *Revolution in Learning* (Harper & Row, New York, 1966); R. O. Hess and R. M. Bear, Eds., *Early Education* (Aldine, Chicago, 1968).
9. The 1969–1970 debate over the effectiveness of early education can be found in M. S. Smith and J. S. Bissel, *Harv. Educ. Rev.* **40**, 51 (1970); V. G. Cicirelli, J. W. Evans, J. S. Schiller, ibid. p. 105; D. T. Campbell and A. Erlebacher, in *Disadvantaged Child*, J. Hellmuth, Ed. (Brunner/Mazel, New York, 1970), vol. 3; *Environment, Heredity and Intelligence* (Harvard Educational Review, Cambridge, Mass., 1969); F. D. Horowitz and L. Y. Paden, in *Review of Child Development Research*, vol. 3, *Child Development and Social Policy*, B. M. Caldwell and H. N. Ricciuti, Eds. (Univ. of Chicago Press, Chicago, 1973), pp. 331–402; T. Kellaghan, *The Evaluation of a Preschool Programme for Disadvantaged Children* (Educational Research Centre, St. Patrick's College, Dublin, 1975), pp. 21–33; U. Bronfenbrenner, *Is Early Intervention Effective?*, DHEW Publication No. (OHD) 74–25 (Office of Human Development, Department of Health, Education, and Welfare, Washington, D.C., 1974).
10. S. Gray and R. Klaus, in *Intelligence: Some Recurring Issues*, L. E. Tyler, Ed. (Van Nostrand Reinhold, New York, 1969), pp. 187–206.
11. This study sprang from prior work attempting to clarify the relationship of moderate malnutrition to mental retardation, in which it became evident that malnutrition of first and second degree might be only one of many correlated environmental factors found in poverty contributing to deficit in cognitive performance. We selected children for treatment with undernutrition as the first criterion not to imply that this was the major critical factor in cognitive development, but to be assured that we were dealing with children found typically in the extreme of poverty in the developing world, permitting generalization of our results to a population that is reasonably well defined in pediatric practice universally. Figure 1 allows scientists anywhere to directly compare their population with ours on criteria that, in our experience, more reliably reflect the sum total of chronic environment deprivation than any other measure available.
12. H. McKay, A. McKay, L. Sinisterra, in *Nutrition, Development and Social Behavior*, D. Kallen, Ed., DHEW Publ. No. (NIH) 73–242 (National Institutes of Health, Washington, D.C., 1973); S. Franková, in *Malnutrition, Learning and Behavior*, N.S. Scrimshaw and J. E. Gordon, Eds. (MIT Press, Cambridge, Mass., 1968). Franková, in addition to showing in her studies with rats that malnutrition caused behavioral abnormalities, also was the first to suggest that the effects of malnutrition could be modified through stimulation.
13. *First Health and Nutrition Examination Survey, United States*, 1971–72, DHEW Publication No. (HRA) 75–1223 (Health Resources Administra-

tion, Department of Health, Education, and Welfare, Rockville, Md., 1975).

14. H. McKay, A. McKay, L. Sinisterra, *Stimulation of Cognitive and Social Competence in Preschool Age Children Affected by the Multiple Deprivations of Depressed Urban Environments* (Human Ecology Research Foundation, Cali, Colombia, 1970).
15. D. B. Jelliffe, *The Assessment of the Nutritional Status of the Community* (World Health Organization, Geneva, 1966), pp. 10–96.
16. In programs that assess community nutritional conditions around the world, standards widely used for formal growth of preschool age children have been those from the Harvard School of Public Health found in *Textbook of Pediatrics*, W. E. Nelson, V. C. Vaughan, R. J. McKay, Eds. (Saunders, Philadelphia, ed. 9, 1969), pp. 15–57. That these were appropriate for use with the children studied here is confirmed by data showing that the "normal" comparison children (group HS) were slightly above the median values of the standards at 3 years of age. Height for age, weight for age, and weight for height of all the study children were compared with this set of standards. The results, in turn, were the basis for the Fig. 1 data, including the "stunting" (indication of chronic, or past, undernutrition) and "wasting" (indication of acute, or present, malnutrition) classifications of J. C. Waterlow and R. Rutishauser [in *Early Malnutrition and Mental Development*, J. Cravioto, L. Hambreaus, B. Vahlquist, Eds. (Almqvist & Wiksell, Uppsala, Sweden, 1974), pp. 13–26]. The classification "stunted" included the children found in the range from 75 to 95 percent of height for age. The "wasting" classification included children falling between 75 and 90 percent of weight for height. The cutoff points of 95 percent height for age and 90 percent weight for height were, respectively, 1.8 and 1.2 standard deviations (S.D.) below the means of group HS, while the selected children's means were 3 S.D. and 1 S.D. below group HS. The degree of malnutrition, in accordance with F. Gomez, R. Ramos-Galvan, S. Frenk, J. M. Cravioto, J. M. Chavez, and J. Vasquez [J. Trop. *Pediatr.* 2, 77 (1956)] was calculated with weight for age less than 75 percent as "second degree" and 75 to 85 percent as "first degree." We are calling low normal a weight for age between 85 and 90 percent. The 75, 85, and 90 percent cutoff points were 2.6, 1.7, and 1.3 S.D. below the mean of group HS, respectively, while the selected children's mean was 2.3 S.D. below that of group HS children. Examination of combinations of both height and weight to assess nutritional status follows the recommendations of the World Health Organization expert committee on nutrition, found in *FAO/WHO Technical Report Series No. 477* (World Health Organization, Geneva, 1972), pp. 36–48 (Spanish language edition). In summary, from the point of view of both the mean values and the severity of deficit in the lower extreme of the distributions, it appears that the group of selected children, at 3 years of age, can be characterized as having had a history of chronic undernutrition rather than suffering from acute malnutrition at the time of initial examination.
17. The data analyses in this article use individuals as the randomization unit rather than sectors. To justify this, in view of the fact that random assignment of children was actually done by sectors, a nested analysis of variance was performed on psychological data at the last data point, at age 7, to examine the difference between mean-square for variation (MS) between sectors within treatment and MS between subjects within treatments within sectors. The resulting insignificant F-statistic ($F = 1.432$, d.f. 15,216) permits such analyses.
18. Food and Nutrition Board, National Research Council, *Recommended Dietary Allowances*, (National Academy of Sciences, Washington, D.C., 1968).
19. D. B. Weikart acted as a principal consultant on several aspects of the education program; D. B. Weikart, L. Rogers, C. Adcock, D. McClelland [*The Cognitively Oriented Curriculum* (ERIC/National Association for the Education of Young Children, Washington, 1971)] provided some of the conceptual framework. The content of the educational curriculum included elements described in the works of C. S. Lavatelli, *Piaget's Theory Applied to an Early Childhood Curriculum* (Center for Media Development, Boston, 1970); S. Smilansky, *The Effects of Sociodramatic Play on Disadvantaged Preschool Children* (Wiley, New York, 1968); C. Bereiter and S. Englemann, *Teaching Disadvantaged Children in the Preschool* (Prentice-Hall, Englewood Cliffs, N.J., 1969); R. G. Stauffer, *The Language-Experience Approach to the Teaching of Reading* (Harper & Row, New York, 1970); R. Van Allen and C. Allen, *Language Experiences in Early Childhood* (Encyclopaedia Britannica Press, Chicago, 1969); S. Ashton-Warner, Teacher (Bantam, New York, 1963); R. C. Orem, *Montessori for the Disadvantaged Children in the Preschool* (Capricorn, New York, 1968); and M. Montessori, *The Discovery of the Child* (Ballantine, New York, 1967).
20. M. McKay, L. Sinisterra, A. McKay, H. Gomez, P. Lloreda, J. Korgi, A. Dow, in *Proceedings of the Tenth International Congress of Nutrition* (International Congress of Nutrition, Kyoto, 1975), chap. 7.
21. Although the "Scree" test of R. B. Cattell [*Multivar. Behav. Res.* 2, 245 (1966)] conducted

on the factor analyses at each measurement point clearly indicated a single factor model, there does exist the possibility of a change in factorial content that might have affected the differences between group HS children and the others at later measurement periods. New analysis procedures based upon a linear structural relationship such as suggested by K. G. Jöreskog and D. Sörbom [in *Research Report 76–1* (Univ. of Uppsala, Uppsala, Sweden, 1976)] could provide better definition of between-occasion factor composition, but the number of variables and occasions in this study still surpass the limits of software available.

22. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danmarks Paedagogiske Institut, Copenhagen, 1960); in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics* (Univ. of California Press, Berkeley, 1961), vol. 4, pp. 321–33; *Br. J. Math. Stat. Psychol.* 19, 49 (1966).

23. B. Wright and N. Panchapakesan, *Educ. Psychol. Meas.* 29, 23 (1969); B. Wright and R. Mead, *CALFIT: Sample-Free Item Calibration with a Rasch Measurement Model* (Res. Memo. No. 18, Department of Education, Univ. of Chicago, 1975). In using this method, analyses included four blocks of two adjacent measurement points (1 and 2, 2 and 3, 3 and 4, 4 and 5). All test items applied to all children that were common to the two proximate measurement points were included for analysis. Those items that did not fit the theoretical (Rasch) model were not included for further analysis at either measurement point, and what remained were exactly the same items at both points. Between measurement points 1 and 2 there were 126 common items included for analysis; between 2 and 3, 105 items; between 3 and 4, 82 items; and between 4 and 5, 79 items. In no case were there any perfect scores or zero scores.

24. Let M_W = total items after calibration at measurement occasion W; $C_{W, L+1}$ = items common to both occasion W and occasion $W + 1$; $C_{W, L-1}$ items common to both occasion W and occasion $W - 1$. Since $C_{W, L+1}$ and $C_{W, L-1}$ are subsets of M_W they estimate the same ability. However, a change in origin is necessary to equate the estimates because the computational program centers the scale in an arbitrary origin. Let $X_{W, L-1}$ and $X_{W, L+1}$ be the abilities estimated by using tests of length $C_{W, L-1}$ and $C_{W, L+1}$, respectively.

    Then:

    $$X_{W, L-1} = ß0 + ßX_{W, L+1} \quad (1)$$

    Since the abilities estimated are assumed to be item-free, then the slope in the regression will be equal to 1, and ß0 is the factor by which one ability is shifted to equate with the other. $X_{W, L+1}$ and $X_{W+1, L+1}$ are abilities estimated with one test at two different occasions (note that $C_{W, L+1} = C_{W+1, L-1}$); then by Eq. 1 it is seen that $X_{W, L-1}$ and $X_{W+1, L-1}$ are measuring the same latent trait. Because the scales have different origins, $X_{W+1, L-1}$ is shifted by an amount ß0 to make them comparable.

25. It must be acknowledged here that with this method the interpretability of the data depends upon the comparability of units (ability scores) throughout the range of scores resulting from the Rasch analysis. Although difficult to prove, the argument for equal intervals in the data is strengthened by the fact that the increase in group means prior to treatment is essentially linear. Further discussion of this point may be found in H. Gomez, paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

26. T. W. Anderson, in *Mathematical Methods in the Social Sciences*, K. J. Arrow, S. Karlin, P. Suppes, Eds. (Stanford Univ. Press, Stanford, Calif., 1960).

27. R. D. Bock, *Multivariate Statistical Methods in Behavioral Research* (McGraw-Hill, New York, 1975); in *Problems in Measuring Change*, C. W. Harris, Ed. (Univ. of Wisconsin Press, Madison, 1963), pp. 85–103.

28. Gain during treatment period is defined here as the mean value of a group at a measurement occasion minus the mean value of that group on the previous occasion. Thus the group T1 gains that form the baseline for this analysis are the following: treatment period 1 = .72; period 2 = 1.44; period 3 = 1.74.

29. C. Deutsch, in *Review of Child Development Research*, vol. 3, *Child Development and Social Policy*, B. M. Caldwell and H. N. Ricciuti, Eds. (Univ. of Chicago Press, Chicago, 1973).

30. L. M. Terman and M. A. Merrill, *Stanford-Binet Intelligence Scale, Form L-M* (Houghton, Mifflin, Boston, 1960), adapted for local use.

31. F. Horowitz and L. Paden, in *Review of Child Development Research*, B. M. Caldwell and H. N. Ricciuti, Eds. (Univ. of Chicago Press, Chicago, 1973), vol. 3, pp. 331–335; S. S. Baratz and J. C. Baratz, *Harv. Edu. Rev.* 40, 29 (1970); C. B. Cazden, *Merrill-Palmer Q.* 12, 185 (1966); *Curriculum in Early Childhood Education* (Bernard van Leer Foundation, The Hague, 1974).

32. Colombia has now begun to apply this concept of multiform, integrated attention to its preschool age children in a nationwide government program in both rural and urban areas. This is, among developing countries, a rarely encountered confluence of science and political decision, and the law creating this social action must be viewed as a very progressive one for Latin

America. Careful documentation of the results of the program could give additional evidence of the social validity of the scientific findings presented in this article, and could demonstrate the potential value of such programs in the other regions of the world.

33. Adapted from a procedure described by H. G. Birch and A. Lefford, in *Brain Damage in Children: The Biological and Social Aspects*, H. G. Birch, Ed. (Williams & Wilkins, Baltimore, 1964). Only the visual-haptic modality was measured.
34. The measure was constructed locally using some of the items and format found in C. Bereiter and S. Englemann, *Teaching Disadvantaged Children in the Preschool* (Prentice-Hall, Englewood Cliffs, N. J., 1969), pp. 74–75.
35. This is a locally modified version of an experimental scale designed by the Growth and Development Unit of the Instituto de Nutrición de Centro América y Panamá, Guatemala.
36. G. Arthur, *A Point Scale of Performance* (Psychological Corp., New York, 1930). Verbal instructions were developed for the scale and the blocks were enlarged.
37. D. Wechsler, *WPPSI: Wechsler Preschool and Primary Scale of Intelligence* (Psychological Corp., New York, 1963).
38. At measurement points 3 and 4, this test is a combination of an adapted version of the arithmetic subscale of the WPPSI and items developed locally. At measurement point 5, the arithmetic test included locally constructed items and an adaptation of the subscale of the WISC-R (39).
39. D. Wechsler, *WISC-R: Wechsler Intelligence Scale for Children-Revised* (Psychological Corp., New York, 1974).
40. At measurement points 3 and 4 the mazes test was taken from the WPPSI and at point 5 from the WISC-R.
41. Taken from (*39*). The information items in some instances were rewritten because the content was unfamiliar and the order had to be changed when pilot work demonstrated item difficulty levels at variance with the original scale.
42. At measurement point 4 the test came from the WPPSI, at point 5 from the WISC-R.
43. At measurement point 4 this was from *WISC: Wechsler Intelligence Scale for Children* (Psychological Corp., New York, 1949); at point 5 the format used was that of the WISC-R.
44. At measurement point 4 this test was an adaptation from the similarities subscale of the WPPSI. At point 5 it was adapted from the WISC-R. Modifications had to be made similar to those described in (*38*).
45. B. Inhelder and J. Piaget, *The Early Growth of Logic in the Child* (Norton, New York, 1964), pp. 151–165. The development of a standardized format for application and scoring was done locally.
46. This research was supported by grants 700-0634 and 720-0418 of the Ford Foundation and grant 5R01HD07716-02 of the National Institute for Child Health and Human Development. Additional analyses were done under contract No. C-74–0115 of the National Institute of Education. Early financial support was also received from the Medical School of the Universidad del Valle in Cali and the former Council for Intersocietal Studies of Northwestern University, whose members, Lee Sechrest, Donald Campbell, and B. J. Chandler, have provided continual encouragement. Additional financial support from Colombian resources was provided by the Ministerio de Salud, the Ministerio de Educación, and the following private industries: Miles Laboratories de Colombia, Carvajal & Cía., Cementos del Valle, Cartón de Colombia, Colgate-Palmolive de Colombia, La Garantía, and Molinos Pampa Rita.

Explanation and Critique

Students are typically critical of this article. First, they argue that the design is not experimental because the children participating in the program were not randomly selected and because there was no randomly assigned control group. In fact, children participating in the program were randomly selected, and although there was no completely untreated randomly selected control group, T1 served this purpose until the time it was subject to treatment (as did T2 and T3).

First, let us examine the random selection of children. The researchers initially identified virtually all the children (333) in the target area (2 square kilometers) below normal in terms of (1) height and weight for age, (2) highest in number of signs of malnutrition, and (3) lowest in per capita family income. The remaining 116 children of the original 449 candidates identified did not show abnormally low weight for age or weight for height. Thus, the *population* for the study became the 333 deprived children in the target area whose parents volunteered them for the program.

This population was then randomly assigned to groups T1 to T4 through a perfectly valid geographically random assignment process. The researchers describe the process and the reasons behind geographic random selection:

> Nearly all the 333 children selected for treatment lived in homes distributed throughout an area of approximately 2 square kilometers. This area was subdivided into 20 sectors in such a way that between 13 and 19 children were included in each sector. The sectors were ranked in order of a standardized combination of average height and weight for age and per capita family income of the children. Each of the first five sectors in the ranking was assigned randomly to one of five groups. This procedure was followed for the next three sets of five sectors, yielding four sectors for each group, one from each of the four strata. At this point the groups remained unnamed; only as each new treatment period was to begin was a group assigned to it and families in the sectors chosen so informed. The children were assigned by sectors instead of individually in order to minimize social interaction between families in different treatment groups and to make daily transportation more efficient. Because this "lottery" system was geographically based, all selected children who were living in a sector immediately prior to its assignment were included in the assignment and remained in the same treatment group regardless of further moves.

Is this an experimental design? Yes, so long as the population consists of *below normal* children. Were children assigned to the groups T1–T4 randomly? Yes, through random assignment of geographic areas.

Then there is the issue of the control groups. Since all of the children eventually receive the treatment, many students wonder where the control group is. Controls are established through the sequencing of the treatment groups through the program. The first group, T4, received its initial treatment from February through November 1971. During this time period, groups T1 – T3 were the control group. From December 1971 through September 1972, T3 received its initial treatment and T2 and T1 constituted the control group. Then from November 1972 through August 1973, T2 received its initial treatment while T1 was the control. Finally T1 received its treatment from November 1973 through August 1974, but for them there was no control group.

But what about the High SES group? This group had little real meaning and was included only to provide additional information.

Another important strength of the McKay article is that it vividly portrays the impact of maturation. The clearest impact of maturation is shown in figure 2 of the article. The figure shows the growth of general cognitive ability of the children from age 43 months to 87 months, the age at the beginning of primary schools. Ability scores are scaled sums of the test items correct among the items common to proximate testing points. Table 4 in the article shows the same information as figure 2 (in the article) but in slightly different form. The preprogram mean cognitive ability scores for the four groups are taken from table 4 and reproduced below (average age at testing 43 months).

	T1	*T2*	*T3*	*T4*
Ability score (43 months)	–1.83	–1.94	–1.72	–1.82

During the next six months, treatment was given only to group T4. At the end of the six month period, all four groups were tested again. The change in the mean test scores for each group is shown below:

	T1	*T2*	*T3*	*T4*
Ability score (49 months)	–1.11	–1.22	–1.06	.21
Ability score (43 months)	–1.83	–1.94	–1.72	–1.82
CHANGE	+.72	+.72	+.66	+2.03

For group T4, the treatment improved the children's mean cognitive ability dramatically from –1.82 to +.21 for a gain of 2.03. The treatment has been successful! But the three untreated groups also improved their cognitive ability by an average of .70. Thus some of the improvement in groups T4 is explained by maturation and not the program. If one assumes that the children in the program would have shown the same growth in cognitive ability as the children not receiving treatment (.70), then it is probably safe to conclude that the program accounted for about two-thirds of the children's change in cognitive ability (2.03 – .70/2.03) and maturation the other one-third (.70/2.03).

These computations are not shown to downgrade the article or the substantial program impact. They merely show how the authors' design allows us to estimate both the program impacts and the influence of maturation on the children's performance.

Overall, this is an elegant article. The sampling design was quite satisfactory. The use of staggered treatment groups allowed for later participants to act as controls for the earlier participants, and it was even possible to isolate the effect of maturation from the effect of the program. A truly elegant article.

References

Hatry, Harry P., Richard E. Winnie, and Donald M. Fisk. 1986. *Practical Program Evaluation for State and Local Governments.* 2d ed. Washington, D.C.: Urban Institute.

Rossi, Peter H., and Howard E. Freeman. *1985. Evaluation: A Systematic Approach.* 3d ed. Beverly Hills, Calif.: Sage.

Supplementary Readings

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* Cambridge, England: Cambridge University Press, chap. 4.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 4.

CHAPTER 6

The Solomon Four-Group Design

The Solomon Four-Group Design takes its name from its originator, Richard L. Solomon. In 1949 Solomon pointed out that a two-group, pretest-posttest design confounds or confuses training effects with the interaction between pretesting and training (Solomon 1949). That is, pretesting may sensitize the subjects to the treatment.[1] The Solomon Four-Group Design depicted in table 6.1 provides a method that shows whether groups are sensitized by the pretest, and if so, by how much. David Nachmias describes the design as follows:

> The Solomon Four-Group design . . . contains the same features as the classic [the pretest-posttest control group design], plus an additional set of comparison and experimental groups that are not measured prior to the introduction of the program. Therefore, the reactive effect of measurement can be directly assessed by comparing the two experimental groups and the two comparison groups. The comparisons will indicate whether X (the policy) had an independent effect on the groups that were not sensitized by the preprogram measurement procedures. (1979, 30)[2]

When is such a complicated design appropriate? Fortunately, only rarely. Because the design documents only the interactive effects of testing, or pretest sensitization, it is not recommended for most evaluations (see Bausell 1986, 102). The design is ultimately appropriate in cases in which a pretest is used and in which the independent effects of this pretest are thought to be substantial (perhaps more significant than the program itself). This is sometimes the case in educational programs (e.g., Lana and King 1960).

The following article, "Evaluation of a Multimethod Undergraduate Management Skills Development Program" (Extejt, Forbes, and Smith 1996), is an example of an adaptation of the Solomon Four-Group Design. Not surprisingly, the evaluation is of an educational program.

TABLE 6.1
Solomon Four-Group Design

	Pretest	Program	Posttest
Group A	O	X	O
Group B	O		O
Group C		X	O
Group D			O

Notes

1. Research has shown that the actual impact of the pretest on the posttest is small to nonexistent. See, for example, Lana 1969.
2. Nachmias uses the terms *control group* and *comparison group* interchangeably.

READING

Evaluation of a Multimethod Undergraduate Management Skills Development Program

Marian M. Extejt J. Benjamin Forbes Jonathan E. Smith
John Carroll University, University Heights, Ohio

ABSTRACT. A Solomon four-group experimental design was used to measure the impact of a multimethod development program on the managerial skills of undergraduate business majors. Skill levels were measured through use of an assessment center. Participation in assessment center exercises had a significant impact on more skills than did participation in the skills development program. The study has implications for management development program design and evaluation.

MANAGEMENT DEVELOPMENT AND training activities account for a significant portion of corporate expense. In 1992, U.S. businesses with 100 or more employees spent $45 billion on training ("Training rebounded," 1993). Yet Georgenson (1982) estimated that only 10% of this expense transfers to actual changes in job performance. Major reviews of training and management development call for greater attention and research into the issues of effectiveness and skill transference to the job (e.g., Goldstein, 1980; Wexley, 1984).

As these costs to business rise, business leaders are voicing their dissatisfaction with the business education that universities and colleges provide. Porter and McKibbin (1988) listed many criticisms of business school curricula, including inadequate emphasis on generating "vision" in students and insufficient attention to managing people and communication skills. In response, educators are examining whether and how behavioral skills and competencies similar to management skills should be developed in undergraduate and graduate students. Recently there has been an explosion of programs, courses, and modules that seek to develop these "soft skills" within traditional educational programs. Several innovative programs that have introduced managerial competencies in their curricula have received widespread publicity. The University of Pennsylvania's Wharton School, cited as having launched the most innovative MBA curriculum in years, focuses on leadership skills training (Byrne & Bongiorno, 1994). University of Tennessee MBA students manage a hypothetical firm, with an emphasis on oral and written communications and problem-solving skills (Massingale & Dewhirst, 1992). At Indiana University, the MBA curriculum emphasizes team building and leadership (Hotch, 1992).

In spite of these major efforts to develop managerial skills and competencies in the

educational system, the effectiveness of such education has received little attention. Considering the growing investment of resources by the academic community, this paucity is troubling.

Effectiveness of Skills Development Courses

Studies to date have focused on the ability of programs to add value to students' abilities. Most have used some type of pre-post experimental design. Outcome studies conducted with MBA students have varied from those assessing the effects of specific skill training programs (e.g., Development Dimensions International [DDI], 1985; Keys & Wolfe, 1990; McConnell & Seybolt, 1991) to those attempting to assess the impact of the entire MBA program (e.g., Boyatzis, Cowen, Kolb, & Associates, 1995).

McConnell and Seybolt (1991) reported what may be significant improvements in communications skills resulting from their intensive, one-on-one, extracurricular program. DDI's (1985) study showed significant improvements on assessment center measures associated with participation in an interaction management workshop. This study also showed that the preassessment with feedback led to improvements in several skill areas.

Preliminary evaluation of the competency-based, outcome-oriented MBA program at Case Western Reserve University has been recently reported (Boyatzis et al., 1995). Preprogram versus postprogram behavioral measures showed improvements in oral communication and group discussion skills. Analysis of these results plus self-report and in-depth interview measures led to the conclusion that there was at least some improvement in 86% of the abilities assessed.

Mullin, Shaffer, and Grelle (1991) conducted two evaluation studies with undergraduates: a one-group, pretest-posttest design and a Solomon four-group design. The single-group study assessed the impact of a one-semester skills development course. Precourse assessments were fed back to the students in individual goal planning conferences. Mullins et al. noted that there were significant improvements; however, the design did not allow an analysis of whether the improvements were a result of the program or of the pretesting with feedback. The Solomon four-group study of Mullin et al. addressed this question. An experimental skills development course was compared with a traditional principles of management course. The experimental program consisted of six learning and assessment exercises and a team learning approach. DDI assessment center instruments were administered before and after the course to both groups. No feedback was given on the precourse assessments. Results showed that the experimental program affected skill performance, although the differences were described as "minor, and the results … mixed" (p. 137). Pretesting did not have a significant effect.

In the programs assessed by DDI (1985) and Mullins et al. (1991), the learning exercises were very similar in structure and content to the DDI assessment exercises used in the pre- and posttreatment assessments. Are such programs successful because they "teach to the test?" Did management skills increase because the experimental students had been specifically coached on how to perform in assessment-center exercises?

Hypotheses

The focus in this study was on the efficacy of management development activities in the traditional university setting. Several researchers have concluded that traditional courses in management do not prepare students to deal with the on-the-job demands faced by practicing managers (Porter & McKibbin, 1988). Dissatisfied with their new employees' lack of practical skills in dealing

with these types of problems, industry has significantly expanded its use of management development programs. Most of these industry programs use techniques based on social learning theory. Thus, Hypothesis 1:

> Undergraduates who participate in a comprehensive management development program based on social learning theory techniques will exhibit a higher level of management skills than those who are not exposed to such a program.

Mullin et al. (1991) suggested that pretesting without feedback does not affect skill improvement. Whetten and Cameron (1991) proposed that one of the problems in developing management skills in students who lack managerial experience in organizations is the students' lack of awareness of their current level of competence in management skills. Other findings have consistently shown that the presence of knowledge of results leads to improved performance (Ilgen, Fisher, & Taylor, 1979). However, Cascio (1991) stated that feedback is often misperceived or not accepted by the recipient. Stone, Gueutal, and McIntosh (1984) found that a main determinant of perceived accuracy of feedback from an in-basket exercise was the rater's level of expertise. It follows then that feedback from a professionally developed assessment center is likely to be perceived as more accurate—and thus result in greater learning and higher motivation to improve—than is self-assessment feedback. In fact, several leading undergraduate (e.g., Alverno College) and graduate (e.g., Case Western Reserve University's Weatherhead School) programs are based on this notion. Thus, Hypothesis 2:

> Undergraduates who participate in a comprehensive developmental assessment center with professional assessor feedback will exhibit a higher level of management skills than those who do not participate in an assessment center.

Finally, Cook and Campbell (1979) warned that one threat to valid interpretation of findings from field studies in which a pretest is used is that an interaction of the pretest with the experimental variables may occur. Because the purpose of a management skills development program is to increase skill levels, we were particularly interested in any factor that could make such a program more effective. Participation in an assessment center may be a positive learning experience in itself; however, it may also lead to enhanced motivation for the program that follows. Thus, Hypothesis 3:

> Undergraduates who participate in a comprehensive management development program and a comprehensive management assessment center will demonstrate a higher level of management skills than those who participate in only one experience or in neither.

Method

Program Description

The School of Business at John Carroll University (a medium-sized, private institution in the midwestern United States, accredited by the American Assembly of Collegiate Schools of Business [AACSB]), conducted a three-semester program designed to develop management skills among a group of undergraduate business majors. Selection of skills to include in the program was based on a review of the leadership and management competencies literature (e.g., Bass, 1981; Boyatzis, 1982; Mintzberg, 1975) and interviews with top executives. The model followed in this undergraduate management skills development program is best described as a comprehensive, multimethod, social learning theory approach. It is similar to Whetten and Cameron's (1991) model but with more emphasis on several key areas: preprogram assessment, exposure to real-life role models, and opportunities for application.

Multiple development techniques were used. Forty-two hours were devoted to in-class activities, primarily lectures by faculty and visiting corporate executives, case studies, role playing, and films. Other activities included participation in two management simulations and individual mentoring by a midlevel corporate executive. Time commitment for activities outside the classroom varied, ranging between 50 and 100 hours.

Different skills were emphasized and different techniques used during each semester. The first semester emphasized modeling and motivation and featured guest lectures by and videotapes of successful business leaders. In the second semester, more emphasis was placed on specific skill learning from role playing, experiential exercises, and related readings. Application and integration of managerial skills began with strategic planning in a small business simulation toward the end of the first semester and continued through the rest of the program. The *Looking Glass, Inc.* (Lombardo, McCall, & De Vries, 1985) management simulation provided further opportunity for application and integration of skills during the final semester. Finally, interaction through shadowing and role playing with mentors allowed students to observe application of skills and to practice them in a relevant setting.

Experimental Design

The impact of the management development program was assessed with a Solomon four-group design (Campbell & Stanley, 1966). This design, outlined in Figure 1, was chosen because it allowed a careful evaluation of the impact of the skills development program, exclusive of threats to internal validity, and because it allowed determination of the impact of the initial assessment center experience (pretesting) on skills development and of the interaction between the program and pretesting. This method is considered one of the most rigorous experimental techniques; however, because of its complexity it is rarely used.

FIGURE 1
Solomon Four-Group Experimental Design Used to Assess Program

	Time period		
	T^1		T^2
Group A	A_1	X	A_2
Group B	B_1		B_2
Group C		X	C_2
Group D			D_2

Note: X stands for skills development program.

Participants in Group A underwent skills testing before (T1) and after (T2) participation in the skills program. Group B's participants also underwent skills measurement at T1 and T2 but did not participate in the program. Group C participated in the development program, but their skills were measured only at T2. Group D received skills testing at T2 but did not participate in the development program.

Participants in Groups A and B received assessment center feedback approximately 6 weeks after participation (T1). Feedback was delivered in both a written and personal counseling framework. The written feedback especially emphasized behavioral changes needed to improve managerial skills. Group C participated in a self-assessment exercise at the start of the program (T1). Participants reviewed their self-assessments with a faculty counselor, and suggestions for improving skills were discussed, but no written feedback was provided. Participants in Group D underwent no assessment of skill levels at T1.

Participants in all experimental groups were required to take courses in the university's liberal arts core curriculum (approximately 60 semester hours) and were required to take 15 business core courses plus any courses required for their chosen majors. One curricular difference existed between

those participating in the development program (Groups A and C) and those in the control groups (Groups B and D). Control group students were required to take a junior-level course in business communications. Students who successfully completed the three-semester development program had the business communications course requirement waived.

Participants

Seventy-five undergraduate business majors were recruited to participate in this experiment. Because of the small number of students available in each condition, assignment to groups was made with the matched pairs method (Kerlinger, 1973). No statistically significant differences ($p < .05$) regarding age, sex, race, academic major, grade point average, or work experience were found among the four groups. In addition, membership in each group reflected the school's population on these same variables.

Because of attrition over the program's three semesters, only 64 students completed the experiment. Attrition was attributed to two factors: (a) Six students were no longer at the university for the final evaluation because of study abroad, early graduation, or withdrawal from school, and (b) 5 students declined to participate in the posttest evaluation because of the extensive time commitment.

Measurement

Sixteen management skills were behaviorally measured with five activity-based instruments, presented in the format of an assessment center. This assessment center was developed by DDI, a Pittsburgh-based consulting firm that specializes in training and development and whose employees served as the consultants for the AACSB Outcome Measures Study (AACSB, 1987). All instruments had been designed for use with undergraduate and graduate business students, and all had been pretested by DDI for administrative and student reaction. Data on the reliability and validity of these instruments are reported in AACSB's (1987) report. These five instruments consist of an in-basket exercise, an analysis and oral presentation exercise, a planning exercise, a group discussion (leaderless group) exercise, and an interview simulation. Students participated in each of these exercises (a 12-hour commitment). Their behaviors were recorded—both in written and video format—and rated by DDI with standard assessment center methods.

Ratings consisted of a numeric score on each of the 16 dimensions defined in the Appendix. Dimension ratings were made relative to job requirements that are typical for supervisors and entry-level managers. Ratings were scaled 1 to 5, with the following anchors: (1) *much less than acceptable,* (2) *less than acceptable,* (3) *acceptable,* (4) *more than acceptable,* and (5) *much more than acceptable.*

Data Analysis

A 2 x 2 x 16 multivariate analysis of variance was used to test for both the main effects of the program and participation in the first assessment center (pretest) and for possible interaction effects on each of the 16 behavioral dependent variables. First, all dependent variables were analyzed with a full-experimental design (main effects and interactions). For those dependent variables that showed no significant interaction effects (14 of 16), univariate analyses of variance (ANOVAs) were conducted, suppressing the interaction effects.

An analysis of covariance (ANCOVA) was conducted on each of the 16 behavioral dimensions, analyzing performance scores at T2 by program participation, with performance scores at T1 as the covariate. In addition, *t* tests were conducted on pre- versus post-program scores for the pretested program group and for the pretested control group.

Results

Results for the multivariate tests for main or interaction effects, shown in Table 1, were not significant at the .05 level. However, a univariate ANOVA showed a main effect for development program participation for two variables: general written communication and judgment. Judgment was positively influenced by the program. General written communication was negatively influenced. Thus, Hypothesis 1 was not well supported.

Main effects for the pretest (participation in the assessment center and subsequent feedback at T1) were evident for oral communication, oral presentation, formal written communication, managing the job, and project planning. Hypothesis 2 was supported by the data (see Table 1).

Two significant interaction effects were found. Students who participated in the program and were pretested showed the highest scores on judgment. Students who participated in both the pretesting and the development program had significantly lower scores on decisiveness at T2 than those who were pretested but did not participate in the program. The interaction hypothesis (Hypothesis 3) thus was not well supported.

When the covariate of initial skill level was accounted for, three dimensions were significantly affected ($p < .05$) by the development program: general written communication, judgment, and decisiveness (see Table 2). Program participation again had a positive effect on judgment and a negative effect on written communication. Although scores on decisiveness increased over time for both

TABLE 1
Behavioral Measure Results

Measure	Pretest Program Group A1 ($n = 20$)	Posttest Program Group A2 ($n = 20$)	Pretest Control Group B1 ($n = 26$)	Posttest Control Group B2 ($n = 11$)	Posttest Program Group C2 ($n = 17$)	Posttest Control Group D2 ($n = 16$)	Program F	Pretest F	Interaction Program x Pretest F
Meeting leadership	1.94	2.13	2.00	1.85	2.00	1.90	1.12	0.00	
Individual leadership	2.10	2.58	2.04	2.30	2.36	2.22	1.47	0.85	
Impact	2.72	2.71	2.72	2.95	2.56	2.56	0.47	2.53	
Oral communication	3.12	3.29	3.09	3.15	2.80	2.90	0.02	6.75**	
Oral presentation	3.30	3.03	3.55	2.95	2.82	2.47	2.22	4.14**	
General written communication	3.27	2.87	3.27	3.40	2.79	3.12	4.41**	0.41	
Formal written communication	3.25	2.83	3.10	2.85	2.47	2.53	0.11	3.59*	
Managing the job	3.42	3.92	3.09	4.05	3.31	3.62	0.84	4.79**	
Project planning	2.65	3.20	2.72	3.18	2.76	2.87	0.03	3.57*	
Operation analysis	2.00	2.25	1.77	2.10	2.06	2.19	0.00	0.25	
Quantitative analysis	1.77	2.35	1.59	2.30	2.44	2.22	0.41	0.12	
Judgment	1.92	2.32	1.86	1.65	1.88	1.96	4.31**	1.30	9.58**
Decisiveness	3.70	3.72	3.86	4.10	3.94	3.75	0.26	2.04	3.43*
Delegation	3.02	2.77	2.82	2.85	2.68	2.48	0.40	2.43	
Follow-up	1.87	2.60	2.27	2.05	2.06	1.81	1.67	1.98	
Information monitoring	1.10	1.20	1.09	1.20	1.06	1.03	0.03	2.01	

Note: Interaction effects are noted only for those variables with significant interactions. All other ANOVA results are for a 2 x 2 model with interaction effects expressed.
* $p < .10$. ** $p < .05$.

groups, there was a significantly higher increase in the control group.

A Comparison of Analytical Methods

The fundamental research question driving this study was "Did the management skills development program have an impact?" A single, simple reply is not possible; the answer depends on how the question is specified, which groups are compared, and the type of analysis conducted.

In Table 3, we present a summary of four methods used to analyze the experiment's results regarding program effect. One method of measuring program impact is to test whether individual participants' skills improve after program participation. A *t* test comparing T1 and T2 assessment center scores for students who participated in the development program (Group A) revealed that 9 of the 16 skills were affected. This limited analysis could generate optimism among program developers, but in-

TABLE 2
Behavioral Measure Results (Analysis of Covariance)

Dimension	Difference score: Group A (program)	Difference score: Group B (program)	Program F
Meeting leadership	0.14	–0.28	2.25
Individual leadership	0.44	0.20	1.11
Impact	–0.03	0.25	1.06
Oral communication	0.16	0.05	0.35
Oral presentation	–0.28	–0.67	0.51
General written communication	–0.40	0.20	5.18**
Formal written communication	–0.43	–0.28	0.03
Managing the job	0.50	0.75	0.13
Project planning	0.55	0.45	0.01
Operational analysis	0.25	0.30	0.27
Quantitative analysis	0.58	0.70	0.00
Judgment	0.40	–0.20	9.88**
Decisiveness	0.03	0.30	4.18*
Delegation	–0.25	0.05	0.16
Follow-up	0.73	–0.10	1.63
Information monitoring	0.10	0.10	0.00

** $p < .05$.

TABLE 3
Comparative Analysis of Program Effects

Skill	*t* tests: Group A, T1 vs. T2: within-subject design	*t* tests: Groups A and C vs. Groups B and D, T2	Analysis of covariance	Analysis of variance
Meeting leadership				
Individual leadership	Positive effect ($p<.05$)			
Impact				
Oral communication				
Oral presentation		Positive effect ($p<.05$)		
General written communication	Negative effect ($p < .05$)	Negative effect ($p < .05$)	Negative effect ($p < .05$)	Negative effect ($p < .05$)
Formal written communication	Negative effect ($p < .05$)			
Managing the job	Positive effect ($p < .05$)			
Project planning	Positive effect ($p < .05$)			
Operational analysis	Positive effect ($p < .05$)			
Quantitative analysis	Positive effect ($p < .01$)			
Judgment	Positive effect ($p < .05$)	Positive effect ($p < .05$)	Positive effect ($p < .05$)	Positive effect ($p < .05$)
Decisiveness			Positive effect ($p < .05$)	
Delegation				
Follow-up	Positive effect ($p < .05$)			
Information monitoring				

ternal threats to the validity of these results abound (Campbell & Stanley, 1966).

When we included the control group in the analysis, the program's impact dropped off dramatically. When the T2 scores of all program participants (Groups A and C) were compared with the T2 scores of all control group members (Groups B and D), only three scores were significantly different. When initial skill level was accounted for, again only three skills were affected. When we conducted an ANCOVA to compare the T2 scores of an experimental group (A) and a control group (B), adjusting for initial skill levels as measured at T1, only judgment and decisiveness were affected in a positive direction. The negative impact of the program on general written communication skills was the same as previously noted. Unfortunately, this analysis does not allow us to isolate the impact of the program exclusive of testing effects.

Comparing T2 scores by using ANOVA to isolate the impact of the program on skill development showed that only two skills were affected: judgment (positively) and general written communication (negatively). This is a far different picture from the one based on the *t*-test scores.

See Table 4 for a summary of three methods of analyzing the impact of the assessment center experience (pretesting) on skill level. A *t*-test analysis of T1 and T2 skill levels for those students who participated only in the testing phases of the experiment (Group B) showed that only 4 of the 16 skills were affected. Managing the job and quantitative analysis were positively affected; participation in the assessment center activities was associated with a decrease in meeting leadership and oral presentation skills.

When we included the control group in analyses, the assessment center's impact increased. When the T2 scores of all pretested participants (Groups A and B) were compared with the T2 scores of all posttest-only members (Groups C and D), six scores were significantly different. Scores on oral communication, oral presentation, formal written communication, managing the job, project planning, and delegation were all higher among participants who had undergone assessment center measurement at the start of the program. However, this analysis does not consider participation in the development program. To assess the impact of assessment center activity participation on skill development, independent of program participation, we conducted an ANOVA on only the posttest scores. This analysis showed that five skills were affected. Again, this is a different picture from the one based on the *t*-test scores.

Discussion

A comprehensive, multimethod management skills development program for undergraduate business students was developed, administered over three semesters, and evaluated with assessment center measures within a Solomon four-group experimental design. The effects of the program were minimal; however, the assessment center experience itself proved to be an effective source of skill improvement.

Assessment Center Pretest Effects

The most significant finding of this study is the existence of pretesting effects on 5 of the 16 management skills: oral communication, oral presentation, formal written communication, managing the job, and project planning. We provided all who went through the assessment center at T1 with a minimal amount of management development guidance. Reactivity to the pretesting was actually encouraged, and, in a sense, pretesting was an alternative form of management development. This finding plus the results of earlier research (Mullin et al., 1991) suggests that assessment with feedback can lead to significant management skill improvement among undergraduates. The assessment center experience itself can serve as an important learning experience for undergraduates.

TABLE 4

Comparative Analysis of Pretest (Assessment Center) Effects

Skill	*t* tests		Analysis of variance
	Group B, T1 vs. T2: within-subject design	Groups A and B vs. Groups C and D (T2)	
Meeting leadership	Negative effect ($p < .05$)		
Individual leadership			
Impact			
Oral communication		Positive effect ($p < .01$)	Positive effect ($p < .05$)
Oral presentation	Negative effect ($p < .01$)	Positive effect ($p < .05$)	Positive effect ($p < .05$)
General written communication			
Formal written communication		Negative effect ($p < .10$)	Negative effect ($p < .01$)
Managing the job	Positive effect ($p < .05$)	Positive effect ($p < .05$)	Positive effect ($p < .05$)
Project planning		Positive effect ($p < .10$)	Positive effect ($p < .10$)
Operational analysis			
Quantitative analysis	Positive effect ($p < .10$)		
Judgment			
Decisiveness			
Delegation		Positive effect ($p < .10$)	
Follow-up			
Information monitoring			

Program Effects

The program had a significant impact on three managerial skills. Judgment was associated with a significant positive main effect due to the program and positive gains associated with the training. Judgment was also associated with a significant interaction between the program and the pretesting effects. The program appears to have had a negative effect on one skill—general written communication. On the decisiveness dimension, there were no posttest measure main effects, although there was a significant interaction and a greater gain for the control group than for the program group.

Improvement in judgment is consistent with the approach taken in this program. We did not attempt to train particular skills, nor did we teach to the test (i.e., here's how to do better in a leaderless group discussion). Instead, we exposed the students to a variety of potential role models, involved them in role playing and experiential exercises, and required their participation in realistic application activities. Thus, program participants' experiences and course work on problem solving and critical thinking assisted in developing their ability to analyze relevant information and make decisions.

The other significant change associated with the experimental program—a decrease in general written communication skills—may be simply explained by considering factors in the general curriculum. For those who completed the development program, the traditional required business communication course was waived. Students who completed the traditional course had much more practice in writing memos and letters.

There are several possible explanations for the overall lack of behavioral improvement (ordered from most pessimistic to most optimistic):

1. Developing management skills through formal programs is not possible. Positive effects are often found with reaction and learning measures but rarely with performance measures (Russell, Wexley, & Hunter, 1984).

2. Significantly improving management skills among undergraduates is not possible. Undergraduates have limited organizational experience, which results in low motivation to improve managerial skills.
3. It is possible to improve management skills among undergraduates but only with more intense, more narrowly focused programs such as those used by DDI (1985) and Mullin et al. (1991). This might entail behavioral modeling, in which specific desirable behaviors are identified and repeatedly rehearsed and reinforced.
4. Program participants may, in fact, have improved in their ability to perform these skills, but postassessment performance may have been hindered in many cases by a lack of motivation. We have anecdotal evidence that, at the time of the posttesting (T2), some students were concerned only with completing the requirements of the program and were not at all concerned with how well they performed.
5. The type of evaluation and its timing were not appropriate. Our program was a general management development experience. We spent relatively little time on specific skill training, even though, in the interest of rigorous evaluation, we chose to measure specific short-term behavioral responses. Perhaps a long-term evaluation of the effects of this experience on the career success of the participants will show more positive results.

Conclusion

The results of this study provide limited support for management skill development programs with undergraduate business students. Although actual skill development was limited, our experience presents two interesting issues that demand future research. First, we had not anticipated the relatively strong impact of the assessment center experience. For learning to occur, the evaluation and feedback components provided by an assessment center seem crucial. Although anecdotal evidence suggests that assessment centers themselves may serve as development activities (Boehm, 1982; Rea, Rea, & Moomaw, 1990), no studies to date have rigorously evaluated this phenomenon. Perhaps, with students who are naive to business practices and management activities, this first experience with real-life management activity increases awareness and knowledge of acceptable business practices. Our findings, and those of Mullin et al. (1991), suggest that feedback on performance is a critical component necessary for skill improvement. Research is needed to distinguish the skill development capabilities of participation in assessment center exercises from the development potential of feedback on performance in an assessment center.

Second, our review of skills development programs in university settings demonstrates that a range of programs exists. At one end of the continuum, skills training programs concentrate on a specific, limited set of skills. Skills are typically developed one at a time. Often, the skills training consists of practicing activities that are almost identical to those used to measure skill levels, both before and after the training program. These types of programs show very positive effects (DDI, 1985; Mullin et al., 1991). At the other end of the range are management development programs. These programs use several methods to cultivate skills, often developing more than one skill in a single activity. The emphasis is on educating the person, focusing on development at all levels of Bloom's (1956) taxonomy. Few, if any, of the activities resemble the measurement activities. As we have shown, this second type of program is much more difficult to evaluate. More research is needed in this area.

The demand for improved managerial skills and for an increase in the relevance of

managerial programs will continue. However, these results suggest that schools of business need to proceed much more slowly with major curriculum changes than the apparent current pace. Faculty and administration must decide what type of program best meets their goals. Given resources, faculty expertise, the institution's mission, and the clientele, choices along the continuum between skills training and skills development must be made. Quinn, Faerman, Thompson, and McGrath (1990) suggested that a combination approach may also be useful. Future studies need to focus on the fit between program type and program goals.

APPENDIX. Definitions of Behavior Dimensions of Management Skills

Dimension	Definition
Meeting leadership	Using appropriate interpersonal styles and methods in guiding meetings toward their objectives; keeping meetings on course
Individual leadership	Using appropriate interpersonal styles and methods in guiding individuals toward goal achievement
Impact	Creating a good first impression; commanding attention and respect; showing an air of confidence
Oral communication	Expressing ideas effectively (includes nonverbal communication)
Oral presentation	Expressing ideas effectively when given time to prepare (includes organization and nonverbal communication)
General written communication	Expressing ideas clearly in memoranda and letters that have appropriate organization and structure, that are grammatically correct, and that adjust language or terminology to the audience
Formal written communication	Expressing ideas clearly in reports or other documents that have appropriate organization or structure, that are grammatically correct, and that adjust language or terminology to the audience
Managing the job	Effectively scheduling one's own time and activities
Project planning	Establishing a course of action to accomplish a specific project or goal; planning proper assignments of personnel and appropriate allocation of resources; communicating expectation about tasks and deadlines; developing contingency plans
Operational analysis	Relating and comparing data from different sources; securing relevant information and identifying key issues and relationships from a base of information; identifying cause-effect relationships
Quantitative analysis	Understanding and evaluating numerical data, tables, charts, or graphs; performing calculations, making comparisons, and combining quantitative information (an understanding of basic trends is more important than performing a large number of calculations)
Judgment	Developing alternative courses of action that are based on logical assumptions and factual information and that take organizational resources into consideration

Decisiveness	Making timely decisions, rendering judgments, taking action, or committing oneself
Delegation	Allocating decisionmaking and other responsibilities to appropriate subordinates; using subordinates' time, skills, and potential effectively
Follow-up	Establishing procedures to monitor the results of delegations
Monitoring information	Establishing procedures for collecting and reviewing information needed to manage a project or an organization

Note. All definitions were supplied by Development Dimensions International, Pittsburgh, Pennsylvania.

References

American Assembly of Collegiate Schools of Business (AACSB). (1987, May). *Outcome measurement project: Phase III report.* St. Louis, MO: Author.

Bass, B. M. (1981). *The handbook of leadership: A survey of theory and research.* New York: Free Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook 1. Cognitive domain.* New York: David McKay.

Boehm, V. A. (1982). Assessment centers and managerial development. In K. M. Rowland & G. G. Ferris (Eds.), *Personnel management* (pp. 327–362). Boston: Allyn & Bacon.

Boyatzis, R. E. (1982). *The competent manager.* New York: Wiley.

Boyatzis, R. E., Cowen, S. S., Kolb, D. A., & Associates. (1995). *Innovation in professional education: Steps on a journey from teaching to learning.* San Francisco: Jossey-Bass.

Byrne, J. A., & Bongiorno, L. (1994, October 24). The best b-schools. *Business Week,* 62–70.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cascio, W. F. (1991). *Applied psychology in personnel management.* Englewood Cliffs, NJ: Prentice-Hall.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Development Dimensions International. (1985). *Final report: Phase III, report to the AACSB.* St. Louis, MO: American Assembly of Collegiate Schools of Business.

Georgensen, D. (1982). The problem of transfer calls for partnership. *Training and Development Journal, 36,* 75–78.

Goldstein, I. L. (1980). Training in work organizations. *Annual Review of Psychology, 31,* 229–272.

Hotch, R. (1992). This is not your father's MBA. *Nation's Business, 80*(2), 51–52.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64,* 349–371.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart & Winston.

Keys, B., & Wolfe, J. (1990). The role of management games and simulations in education and research. *Journal of Management, 16,* 307–336.

Lombardo, M. M., McCall, M. W., & De Vries, D. L. (1989). *University edition: Looking Glass, Inc.* [Instructor's manual]. Greensboro, NC: Center for Creative Leadership.

Massingale, C., & Dewhirst, D. (1992). Delivery of the new MBA curriculum. *Survey of Business, 28*(1), 10–12.

McConnell, R. V., & Seybolt, J. W. (1991). Assessment center technology: One approach for integrating and assessing management skills in the business school curriculum. In J. D. Bigelow (Ed.), *Managerial Skills: Explorations in practical knowledge* (pp. 105–115). Newbury Park, CA: Sage.

Mintzberg, H. (1975). The manager's job: Folklore or fact. *Harvard Business Review, 53,* 49–61.

Mullin, R. F., Shaffer, P. L., & Grelle, M. J. (1991). A study of the assessment center method of teaching basic management skills. In J. D. Bigelow (Ed.), *Managerial skills: Explorations in practical knowledge* (pp. 116–139). Newbury Park, CA: Sage.

Porter, L. W., & McKibbin, L. E. (1988). *Future of management education and development: Drift or thrust into the 21st century?* New York: McGraw-Hill.

Quinn, R. E., Faerman, S. R., Thompson, M. P., & McGrath, M. R. (1990). *Becoming a master manager: A competency framework.* New York: Wiley.

Rea, P., Rea, J., & Moomaw, C. (1990). Use assessment centers in skill development. *Personnel Journal,* April 126–131.

Russell, J. S., Wexley, K. N., & Hunter, J. E. (1984). Questioning the effectiveness of behavior modeling training in an industrial setting. *Personnel Psychology, 37,* 465–481.

Stone, D. L., Gueutal, H. G., & McIntosh, B. (1984). The effects of feedback sequence and expertise of the rater on perceived feedback accuracy. *Personnel Psychology, 37,* 487–506.

Training rebounded in 1992. (1993). *HR Focus, 70*(2), 14.

Wexley, K. N. (1984). Personnel training. *Annual Review of Psychology, 35,* 51–55.

Whetten, D. A., & Cameron, K. S. (1991). *Developing managerial skills* (2nd ed.). New York: Harper Collins.

Explanation and Critique

This evaluation by Marian Extejt and her colleagues may not grab headlines but it is a reasonable piece of work. John Carroll University, outside of Cleveland, responded to dissatisfaction with business education that universities and colleges provide, as voiced by business leaders. Some of these criticisms of business school curricula included inadequate emphasis on generating "vision" in students and insufficient attention to communications skills and to managing people.

In response, the School of Business at John Carroll developed and implemented a three-semester program designed to develop management skills among a group of undergraduate business majors. Multiple development techniques were used in the program. The authors tell us:

> Forty-two hours were devoted to in-class activities, primarily lectures by faculty and visiting corporate executives, case studies, role playing, and films. Other activities included participation in two management simulations and individual mentoring by a midlevel corporate executive. Time commitments for activities outside the classroom varied, ranging between 50 and 100 hours.
>
> Different skills were emphasized and different techniques used during each semester. The first semester emphasized modeling and motivation and featured guest lectures by and videotapes of successful business leaders. In the second semester, more emphasis was placed on specific skill learning from role playing, experiential exercises, and related readings. Application and integration of managerial skills began with strategic planning in a small business simulation toward the end of the first semester and continued through the rest of the program. . . . Finally, interaction through shadowing and role playing with mentors allowed students to observe application of skills and to practice them in a relevant setting.

A key component of the overall program was an initial skills assessment center experience in which baseline data were obtained to measure program outcomes. Sixteen management skills were behaviorally measured.

Extejt and colleagues reacted to criticisms of evaluations of similar programs found in the literature by designing and implementing a modified Solomon Four-Group Design to evaluate the program. The four-group design was chosen not only because it would allow for a careful evaluation of the skills development program itself, but because it would afford the investigators the opportunity to

determine the impact of the initial assessment center experience—the pretest—on skills development, and to look for any interaction between the program and pretesting.

Only seventy-five undergraduate business majors were selected for the experiment, and only sixty-four of these completed it. Recognizing a potential for "unhappy randomization" with such a small number of students, Extejt and colleagues did not use random assignment to construct the four groups. Instead they used matched pairs (since assignment to groups was not random, this cannot be considered a true experimental design). Then, to ensure that the four groups were equivalent in all relevant respects, they compared the groups in terms of age, sex, race, academic major, grade point average, and work experience. No statistically significant differences ($p < .05$) were found among the four groups. And, in addition, membership in each group reflected the school's population on these same variables.

Table 6.2 outlines the design used for the evaluation. Group A underwent an initial assessment center experience and skills testing both before and after the program. Group B underwent the same experiences except for participating in the program. Students in these groups received both verbal and written assessment center feedback approximately six weeks after participation.

Strangely, Group C was exposed to a form of pretest. Group C participated in a skills self-assessment exercise at the pretest point. Participants then reviewed their self-assessments with a faculty counselor. Suggestions for improving skills were discussed, but no written feedback was provided. Participants in Group D underwent no assessment of skills levels at the pretest point.

One curricular difference existed between the groups. Comparison group students were required to take a junior-level business communications course—but the course was waived for students successfully completing the three-semester development program (the experimental groups).

TABLE 6.2
Solomon Four-Group Design for Evaluation of the Skills Development Program

	Pretest	Program	Posttest
Group A	O	X	O
Group B	O		O
Group C	O[1]	X	O
Group D			O

1. Experienced a limited form of pretest.

Students were evaluated by numeric scores on each of sixteen dimensions defined in the Appendix to the article. When one looks only at a comparison of the pretest and posttest scores of the students in Group A, nine of the sixteen skills seemed to have been significantly affected with seven being positively affected. But when the posttest scores between groups A and C (experimental groups) were compared with Groups B and D (comparison groups), only three skills were significantly affected with only two having positive effects.

So this adaptation of the Solomon Four-Group Design illustrates that sometimes there are pretest impacts that can be stronger than the impacts of the program. This was clearly the case here.

This chapter on the Solomon Four-Group Design is placed in the section of the book on experimental designs because it is normally an experimental design with subjects randomly assigned to the four experimental/control groups. In this case, however, the authors modified the design, and it became a comparison group design using matched pairs because the authors recognized at the outset that with their small number of participants, they were unlikely to have a successful randomization. This kind of flexibility in evaluations is to be encouraged. De-

signs should always be modified to fit the specific situation of the evaluation.

One other strength of the evaluation is in terms of measurement. The instruments had been developed by DDI, a consulting firm specializing in training and development. Data on both the reliability and validity of the instruments had been previously reported (AACSB 1987).

The major methodological disappointment to the evaluation was the participation of Group C in a modified skills self-assessment process. Group C, like Group D, should have received no pretest treatment. It may be that the program designers thought that some form of skills assessment was necessary for program participation—but the evaluators did not elaborate.

References

American Assembly of Collegiate Schools of Business (AACSB). 1987. *Outcome Measurement Project: Phase III Report.* St. Louis, Mo.: AACSB, May.

Bausell, R. Barker. 1986. *A Practical Guide to Conducting Empirical Research.* New York: Harper and Row.

Extejt, Marian M., J. Benjamin Forbes, and Jonathan E. Smith. 1996. "Evaluation of a Multimethod Undergraduate Management Skills Development Program." *Journal of Education for Business* 71 (March/April): 223–31.

Lana, Robert E. 1969. "Pretest Sensitization." In *Artifacts in Behavioral Research,* edited by Robert Rosenthal and Ralph L. Roshow. New York: Academic Press.

Lana, Robert E., and David J. King. 1960. "Learning Factors as Determinants of Pretest Sensitization." *Journal of Applied Psychology* 44, no. 3: 189–91.

Nachmias, David. 1979. *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's.

Solomon, Richard L. 1949. "An Extension of Control Group Design." *Psychological Bulletin* 46 (March): 137–50.

CHAPTER 7

Posttest-Only Control Group Design

The posttest-only control group design is a variation on the pretest-posttest design and the Solomon design. The major difference is that it omits the pretested groups altogether. The posttest-only control group design is illustrated in table 7.1.

Under this design, individuals are randomly assigned to groups E (experimental group) or C (control group). The first group is subjected to the treatment, and progress is measured during or after the program.

Nachmias aptly describes the advantages:

> The . . . design controls for all intrinsic sources of invalidity with the omission of the pretest, testing and instrument decay become irrelevant sources of invalidity. It can also be assumed that the remaining intrinsic factors are controlled, since both groups are exposed to the same external events and undergo the same maturational processes. In addition, the extrinsic factor of selection is controlled by the random assignment of individuals, which removes an initial bias in either group. (1979, 32)

What all this means is that, with random assignment, a pretest may be unnecessary. This provides some distinct advantages. For one thing, it is sometimes difficult to convene subjects for a pretest before a study. For another, repeated measurement can sometimes be expensive in terms of time and resources. In addition, some evaluations are quite transparent, and the researcher may wish to disguise the purpose of the experiment or even hide the fact that the study is in progress. Finally, it is possible that the pretest may cause the subjects to react differently to the treatment—the pretest sensitization problem. To the extent that this problem exists, it is obviously eliminated if there is no pretest.

Why then have a pretest at all? If the posttest-only control group design is so good, why ever use the pretest-posttest control group design? One reason is that a pretest allows the evaluator to reduce the size of the sample (and thus reduce costs). The pretest itself is used statistically as a controlling variable in a repeated measures analysis of either variance or covariance. When subjects are in limited supply, a pretest is typically recommended (Bausell 1986, 93).

TABLE 7.1
Posttest-Only Control Group Design

	Pretest	Program	Posttest
Group E		X	O
Group C			O

Another good reason exists for having a pretest. It provides a good check on the randomization process. Without a pretest, one presumes but does not know that random assignment causes the experimental and control groups to start out at the same point, but one can never be absolutely sure. The pretest thus allows us to test for differences in the two groups that could be accounted for by what Lawrence Mohr terms "unhappy randomization" (1988, 46).

Finally, there is always a concern about client preference. Clients frequently feel more comfortable with a pretest. Evaluators can point out the redundancy of a pretest, but if the client is not happy, a pretest may be appropriate. For these reasons, the pretest-posttest design is much more popular than the posttest-only design (even when the posttest-only design is perfectly appropriate). The following article, "Community Post-hospital Follow-Up Services" by Ann Solberg, is an example of the posttest-only control group design.

READING

Community Posthospital Follow-up Services

Ann Solberg
Department of Health, Fresno County, California

Clients ready for discharge from the Fresno County Department of Health Acute Psychiatric Unit were randomly assigned to an experimental group (n = 71) or a control group (n = 72). The individuals in the experimental group received community follow-up services from one of four psychiatric social workers for a period of 30 days after their discharge from the acute psychiatric unit. The control group received no follow-up services from the project staff. The evaluation of the project was based on a period of 60 days following each client's discharge from the hospital. The results showed that posthospital community follow-up, as provided in this demonstration project, is effective in preventing or at least delaying rehospitalization, without increasing the cost of mental health care.

AUTHOR'S NOTE: *This demonstration project was supported by the California State Department of Health. I wish to acknowledge the project staff: Shirley Carlson, LCSW; Joan Poss, MSW; Donna Ward, MSW; Jane Van Dis, MSW; Alice Sutton, MSW; and Daniel Hart, MSW. Others who facilitated the project were Jeanne Book, Ph.D.; Keith Goble, LCSW; Chris Christenson, LCSW; Denise Gobel, B.A.; and Detlev Lindae, Ph.D. Correspondence may be directed to Ms. Solberg at the Fresno County Department of Health, P.O. Box 11867, Fresno, CA 93775.*

IN COMMUNITY MENTAL HEALTH CENTERS the psychiatric unit is only one phase in the continuum of mental health care. The role of the psychiatric unit is to stabilize acutely ill clients and refer them to other mental health services for continued care. These programs are referred to as posthospital or aftercare programs when they are included in the discharge plan of a client who has been hospitalized.

There is a general consensus among mental health professionals that attending posthospital programs is necessary to the client's continued progress and avoidance of rehospitalization. However, many discharged

patients referred to such programs do not complete the referrals. Bass (1972) pointed out that the operational principles for maintaining the continuity of treatment were specified as requirements for federally funded centers. Therefore, it is the responsibility of each mental health center to promote continuity of care within its own system. Wolkon and associates (1978) concur that it is the responsibility of the community mental center to help discharged patients take advantage of needed services.

Studies in which the problem of rehospitalization has been examined from the viewpoint of aftercare program attendance have yielded inconsistent results. While most studies have shown that rehospitalization rates are lower for individuals who receive aftercare services compared to those who do not (Free and Dodd, 1961; Beard et al., 1963; Hornstra and McPartland, 1963; Mendel and Rappaport, 1963; Greenblatt et al., 1963; Purvis and Miskimins, 1970; Anthony and Buell, 1973; Smith et al., 1974; Winston et al., 1977), other studies show no differences between aftercare and no-aftercare groups (Brown et al., 1966; Michaux et al., 1969; Mayer et al., 1973; Franklin et al., 1975). Kirk (1976) attributes these diverse findings to differences in methodology, outcome measures, length of follow-up, type and size of patient samples, and treatment settings. Winston and associates (1977) identified the lack of systematic comparison of study groups as another possible source of conflicting results. Studies in which individuals who have attended aftercare programs were compared to those who have not, may have been confounded by subject variables stemming from the process of self-selection.

The purpose of this article is to present the results of a 3-month demonstration project that ended in July 1980. The purpose of the project was to evaluate the effectiveness of community follow-up services in reducing both rehospitalization and the overall cost of mental health care of individuals hospitalized at the Fresno County Department of Health (FCDH) Acute Psychiatric Unit (APU). Clients discharged from the APU were randomly assigned to either the experimental group or the control group. The control group received only those aftercare services for which they completed referrals or services that they sought on their own. The experimental group received additional posthospital follow-up services from a team of four psychiatric social workers for a period of up to 30 days following their discharge from the APU. The follow-up team worked with clients, their families, and other treatment staff with the goals of improving clients' personal and community adjustments and increasing their involvement with the aftercare programs (outpatient, partial day, or residential) to which they were referred by APU staff. The follow-up services were treated as an additional aftercare service. They were intended to supplement rather than substitute for aftercare services specified in a client's discharge plan.

The effectiveness of the project was evaluated by comparing the two study groups on selected outcome measures. The evaluation period lasted for the 60 days following each client's discharge from the APU. The decision to extend the evaluation period was based, in part, on a statistical consideration. The total number of clients rehospitalized from both study groups was larger in the 60-day period. The larger sample improved the discriminating power of the test used to compare the survival rates of the study groups. In addition, it was believed that any beneficial effects produced by the follow-up services should extend beyond the service period. In this study no attempt was made to associate client characteristics with differences in predisposition to hospitalization or in relative success in the follow-up project.

Two hypotheses concerning the effectiveness of the follow-up services were tested.

(a) The follow-up services will prevent or at least delay the rehospitalization of clients.
(b) The follow-up services will reduce the cost of mental health services.

Method

Subjects

The population studied was defined to include all persons who were hospitalized at the Fresno County Department of Health (FCDH) Acute Psychiatric Unit (APU), with the exception of individuals meeting one or more of the following disqualifying criteria:

(1) residence outside of Fresno County,
(2) supervision by the law enforcement system,
(3) referral to alcohol or drug abuse programs,
(4) referral to FCDH Advocate program,
(5) referral to subacute locked facilities,
(6) transfer to a psychiatric unit in another hospital.

These criteria helped ensure that the subjects (1) would be accessible to the follow-up workers, (2) were not in need of specialized treatment programs, (3) did not have another mental health worker supervising their treatment plan, and (4) would return to the APU in the event of future need for hospital services.

The sample consisted of 143 persons discharged from the APU during a three-month period ending in July 1980. The subjects were randomly assigned to two groups—experimental and control—upon their discharge from their first hospital episode during the study period. (The first hospital episode was not necessarily the first hospital admission in the client's psychiatric history.) There were 71 subjects in the experimental group and 72 subjects in the control group. The experimental group received the posthospital follow-up services for 30 days, and the control group did not.

The characteristics of the experimental group were as follows: females (45%); males (55%); age ($X = 29.4$; $SD = 10.8$); psychoses (51%); neuroses (49%); global impairment rating ($X = 3.6$; $SD = 1.2$). Similarly, the characteristics of the control group were as follows: females (43%); males (57%); age ($X = 32.8$; $SD = 12.1$); psychoses (51%); neuroses (49%); global impairment rating ($X = 3.6$; $SD = 1.3$).

Measures and Analysis

Information for testing the hypotheses was based on a 60-day evaluation period following each client's discharge from his or her first hospital episode during the study period. Hospital data were limited to hospitalizations taking place at the APU. Following clients closely for 60 days to record possible admissions to other hospitals was neither practical nor necessary. The likelihood that rehospitalization would take place at the APU was increased considerably by including in the sample only those persons who completed their hospital episode at the APU, rather than being transferred to another facility. The exceptions (hospitalizations at other facilities) were treated as a random variable, assumed to affect both study groups equally.

Survival Time

Survival time was defined as the length of time between a client's discharge from the first hospital episode and the onset of their first rehospitalization at the APU during the 60-day evaluation period. The 60-day fixed evaluation period resulted in incomplete information (censored observations) on the actual length of survival for many of the clients. All that is known about these clients is that they did not return to the hospital for at least 60 days after their first hospital episode. The test developed by Gehan (1965), which com-

pares two groups with respect to the length of survival when one or both samples contains censored observations, was used in this study.

Cost Analysis

There were three sources of mental health care costs included in the cost analysis: (1) the cost of rehospitalization, including treatment and intake evaluations at outpatient clinics; (2) the cost of aftercare program participation; and (3) the cost of the follow-up services, which were applicable to the experimental group only. Overhead expenses were included in all cost categories.

Procedure

Clients were assigned to groups at the time of discharge from their first hospital episode during the study period. Clients were not considered for reassignment upon their discharge from subsequent hospitalizations. A clerk at the APU notified the research assistant as soon as the decision to discharge a client had been reached. The research assistant determined, based on the aftercare referrals made for the client by APU staff, if the client being discharged met the target group criteria. A client who met the target group criteria was randomly assigned to the experimental group or the control group by the research assistant using a group assignment sheet. The assignment sheet was prepared by the project evaluator prior to the study using the one-digit columns of a table of random numbers. The numbers were assigned to subjects according to their order of appearance, beginning at the top of the column and progressing downward. Even numbers designated assignment to the experimental group and odd numbers designated assignment to the control group. The last subject was assigned 30 days prior to the last day of the study period to allow for a full 30 days of follow-up. Precautions were taken to ensure that the assignment of clients to groups was not biased by special interests. The APU staff were not informed about the target-group criteria. The project social workers and evaluator were not involved with subject assignment beyond defining the target group and preparing the initial assignment list based on the table of random numbers.

The research assistant contacted one of the social workers when a client was assigned to the experimental group. The social workers then decided among themselves who would accept the case, considering immediate availability and size of caseloads. After reaching a decision, a social worker met the client at the APU. Within 24 hours, the social worker made the first home visit. Follow-up services were provided for a period of 30 days after a client's initial discharge from the APU during the study period. Clients were given only one follow-up period, regardless of subsequent hospitalizations. At the end of the 30-day period, the social worker discharged the client from the follow-up project and made referrals to aftercare programs according to the client's need for continued care.

The project evaluator collected the data for the outcome measures and made periodic visits to the project site. Information regarding hospitalizations was obtained from records at the APU. A 24-hour report showed the date and time of admission and discharge. Aftercare program participation was obtained from the FCDH management information system (MIS). The type and frequency of follow-up contacts were recorded by the project social workers. Cost information was obtained from the MIS and the accounting office.

Results

Description of Follow-up Services

Of the group of 71 clients who received follow-up services, 58 (82%) were seen by a social worker at the APU prior to their dis-

charge. Forty-six clients (65%) received a home visit within 24 hours of their discharge. The clients, as a group, received a total of 546 client contacts (face-to-face and telephone contacts) during the 30-day follow-up period. The median number of contacts was five per client. Collateral contacts—a second category of follow-up contacts—included face-to-face and telephone contacts with mental health staff and family members. There were 551 collateral contacts in the 30-day period, with a median of four contacts per client. Figure 1 shows the percentage of total client contacts (546) and total collateral contacts (551) that were provided each day of the 30-day follow-up period. Not shown on this figure are 19 client and 42 collateral contacts that took place after the 30-day follow-up period ended. It was not in the best interest of 26 (37%) clients to discharge them on precisely the 30th day.

There were 54 clients (76%) who are identified as "service completers," because they cooperated with the social workers throughout the 30 day period. The other 17 clients (24%) were designated as drop-outs because they refused services during the first few contracts or were resistive to the extent that the social workers stopped initiating contact with them before the follow-up period ended.

Hospital Contacts

During the 60-day evaluation period, 8 experimental subjects (11.3%) and 20 control subjects (27.8%) were hospitalized at least once. Two experimental subjects and four control subjects were rehospitalized twice. Clients in the experimental group were in the hospital a total of 83.9 days, compared to 206.6 days accumulated by the control group. The total hospital days for the control group was estimated for 71 clients by multiplying the group average, based on 72 clients, by 71.

The survival-time distributions for the study groups are presented in Figure 2. Comparing the groups on survival time, the results of Gehan's test support the hypothesis that the follow-up services are effective in preventing or at least delaying the recurrence of hospitalization ($V = 2.77$; $p < .05$). The median survival time for the experimental

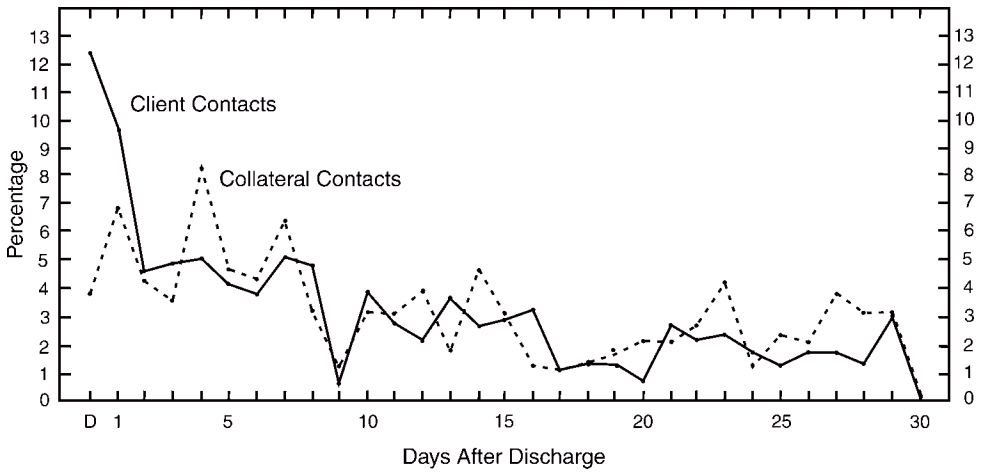


FIGURE 1

Percentage of Client and Collateral Follow-Up Contacts Received by Experimental Group Subjects Each Day of the 30-day Follow-Up Period

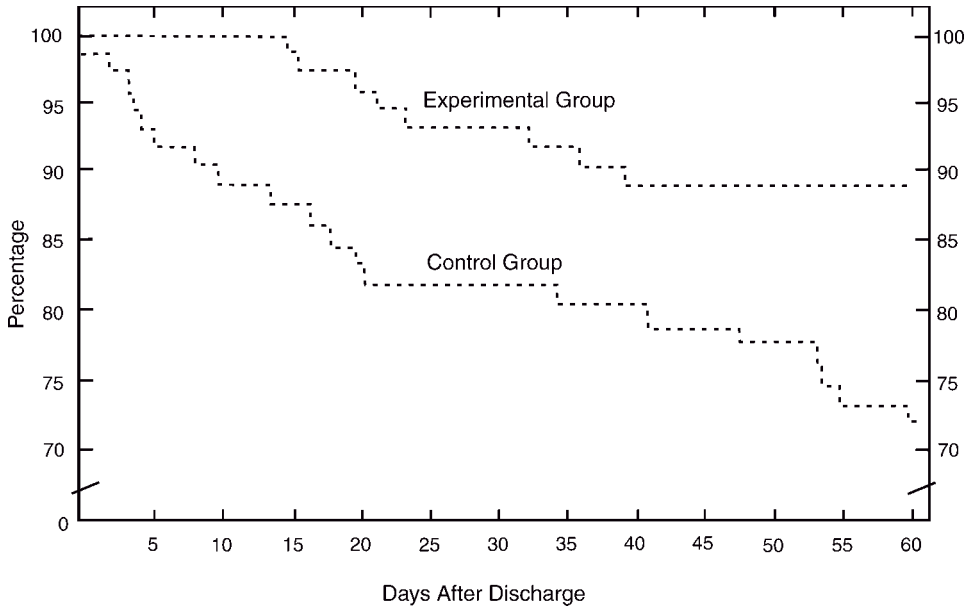


FIGURE 2
Survival Rates for Study Groups

group was 22.8 days, compared to 17.4 days for the control group. Also, none of the experimental subjects was rehospitalized within 14 days of discharge from the APU, compared to nine control subjects (12.7%) who were rehospitalized during the same period. The two-week period after discharge has been identified as a period of high risk for rehospitalization.

Cost of Mental Health Care

Table 1 shows the cost of mental health care for the study groups during the 60-day evaluation period. The total cost for the experimental group ($60,165) was $10,298 less than the total cost for the control group ($70,463). The t-test for these outcomes showed no significant differences between the two groups regarding total program costs ($t = 1.21 < 1.96$; $p > .05$). The total hospital cost (Acute Psychiatric Unit) was $13,407 for the experimental group and $49,703 for the control group. A significant difference was found between the two groups, indicating a reduction in the cost of hospitalization for the experimental group ($t = 8.50 = 1.96$, $p = .05$). The cost of aftercare services was observably higher for the experimental group ($25,093) compared to the control group ($20,760), but the *t*-test showed that the costs were not statistically different from one another ($t = 1.18 < 1.96$, $p > .05$). The cost figures for aftercare services were based on 439 aftercare contacts for 40 experimental subjects and 355 aftercare contacts for 37 control subjects.

Discussion

In this study, the event of readmission was used as a complex index of the individual's overall adjustment and acceptance by the family and community. Solomon and Doll (1979) have indicated that there are many factors that contribute to a decision to rehospitalize a person, other than the individual's psychiatric status. These include family and community acceptance of the individual,

TABLE 1
A Comparison of Study Groups on the Cost of Mental Health Care During a 60-Day Period After Subjects' Discharge from Acute Psychiatric Unit

	Experimental Group		*Control Group*		
Cost Components	*Total*	*SD*	*Total*	*SD*	*t*
Acute Psychiatric Unit	$13,407	$15,882	$49,703	$32,583	8.50*
Aftercare Services	$25,093	$24,072	$20,760	$19,733	1.18
Follow-up Services	$21,665	$15,219	—	—	
Total	$60,165	$40,650	$70,463	$52,313	1.32

Note: The cost of follow-up services includes the salaries of 3.5 full-time equivalent psychiatric social workers, a part-time psychiatric social worker supervisor, a full-time research assistant, supplies, travel expenses, and all overhead expenses.
*$p \leq .05$.

characteristics and perceptions of admitting personnel, factors in the mental health delivery system (e.g., hospital policies, census, and availability of community alternatives), and the individual's perception of the hospital as being a solution to his or her problems. The follow-up team dealt with all factors that had the potential for increasing a client's community tenure. Rehospitalization was considered only when no other solution could be found.

The results showed that community follow-up services, as provided in this demonstration project, provide an effective means of reducing the rehospitalization of individuals who have been hospitalized in a short-term acute psychiatric facility at the Fresno County Department of Health. The results of the survival-time analysis showed that the follow-up services were effective, at least in delaying further hospitalizations. Many clients who received follow-up services and were not hospitalized during the 60-day evaluation period may have been hospitalized afterward. While it is doubtful that clients with a history of relapse will be prevented from ever returning to the hospital, persons whose psychiatric condition is not chronic are more likely to have future hospitalizations prevented altogether. Because not all rehospitalizations can be prevented, the follow-up period should be renewed each time a person is hospitalized. The limit of one follow-up period in this study was a requirement of the design used for the evaluation.

The cost analysis showed that the savings in hospital costs were substantial ($36,296). The follow-up staff stated that they could have served an additional ten cases (total 81) without reducing the effectiveness of their follow-up efforts, and therefore, the potential for savings may be somewhat underestimated. In addition, the cost analysis showed that there were no significant increases in aftercare program costs nor in total mental health care costs as a result of adding the follow-up component to the service delivery system. In other words, the savings in hospital care paid for the additional cost of providing follow-up services.

Factors in the follow-up model that are believed to be effective in reducing hospitalization include (1) an assertive approach to follow-up, (2) contact with the client and staff prior to the client's discharge, (3) more frequent contact in the beginning of the follow-up period, (4) services provided to the client in his or her natural environment, and (5) the role of follow-up workers as case managers. The follow-up workers were strongly encouraged to contact clients more often at the beginning of the follow-up period. Consequently, most clients (82%) received an initial visit by a follow-up worker on the psychiatric unit, and a greater percent-

age of contacts took place early in the follow-up period. Previous readmission data for the psychiatric unit in this study have consistently shown that a greater percentage of clients return within the first two weeks after discharge than during any other time period up to 90 days after discharge. Therefore, follow-up services were especially emphasized during the early days of the follow-up period—identified as a high risk period. The effectiveness of this strategy was demonstrated by the result that none of the experimental subjects returned to the hospital within two weeks after their discharge from the psychiatric unit, whereas nine control subjects were rehospitalized during the same period.

Because this study was carefully conducted, following principles of controlled research, it is believed that these results can be replicated using different samples from the same population. The decision to continue the project was based on this rationale. However, there will be two changes in the follow-up procedure. First, the follow-up period will be renewed each time a client is discharged from the hospital. A second change will allow the follow-up period to vary between 30 and 40 days. The social workers judged that for about one-third of the clients, a strict adherence to a 30-day period may result in abrupt discharges, which are potentially detrimental to their progress.

It is believed that posthospital community follow-up services have a likelihood for simlar success in other locations. Characteristics that are unique to clients in other settings may need to be considered in developing an effective follow-up strategy.

Summary

The purpose of this project was to evaluate the effectiveness of community follow-up services in the context of a true experimental design. One hundred forty-three clients ready for discharge from the Fresno County Department of Health Acute Psychiatric Unit were randomly assigned to either an experimental group or a control group. There were 71 subjects in the experimental group and 72 subjects in the control group. The individuals in the experimental group received community follow-up services from one of four psychiatric social workers for a period of 30 days after their discharge from the acute psychiatric unit. The control group received no follow-up services from the project staff.

The evaluation period lasted for 60 days following each client's discharge from the hospital. The results showed that posthospital community follow-up, as provided in this demonstration project, is effective in increasing the survival time of clients and in reducing the cost of mental health care. The survival time—length of time outside of the hospital—was significantly greater for the experimental group compared to the control group ($p \leq .05$). The median survival time for the experimental group was 22.8 days and for the control group, 17.4 days. The percentage of experimental subjects who were rehospitalized (28%), and the number of hospital days for the experimental group (84) was less than one-half the number of hospital days for the control group (207). The savings in hospital costs ($35,896) more than paid for the cost of the follow-up services ($21,665), and the total cost for the experimental group ($60,165) was $10,298 less than the total cost for the control group. It was concluded, based on the results of the cost analysis, that hospital expenditures were substantially reduced, and this savings paid for the cost of the follow-up program.

Note

1. One of the four social workers left the project during the fifth week. This social worker was not replaced, because the project was found to be overstaffed relative to the number of clients being included as experimental subjects.

References

Anthony, W. A. and G. J. Buell (1973) "Psychiatric aftercare clinic effectiveness as a function of patient demographic characteristics." *J. of Consulting and Clinical Psychology* 41, 1: 116–119.

Bass, R. (1972) *A Method for Measuring Continuity of Care in a Community Mental Health Center.* Report 4, Series on Mental Health Statistics. Washington, DC: National Institute of Mental Health.

Beard, J. N., R. B. Pitt, S. H. Fisher, and V. Goertzel (1963) "Evaluating the effectiveness of a psychiatric rehabilitation program." *Amer. J. of Orthopsychiatry* 33: 701–712.

Brown, G., et al. (1966) *Schizophrenia and Social Care.* London: Oxford Univ. Press.

Franklin, J., L. D. Kiittredge, and J. H. Thrasher (1975) "A survey of factors related to mental hospital readmissions." *Hospital and Community Psychiatry* 26, 11: 749–751.

Free, S. and D. Dodd (1961) "Aftercare for discharged mental patients." Presented at the Five-State Conference on Mental Health, Richmond, Virginia.

Gehan, E. A. (1965) "A generalized Wilcoxon test for comparing arbitrarily singly censored samples." *Biometrika* 52: 203–223.

Greenblatt, M., R. F. Moore, R. S. Albert, and M. H. Soloman (1963) *The Prevention of Hospitalization.* New York: Grune & Stratton.

Hornstra, R. and T. McPartland (1963) "Aspects of psychiatric aftercare." *Int. J. of Social Psychiatry* 9: 135–142.

Kirk, S. A. (1976) "Effectiveness of community services for discharged mental hospital patients." *Amer. J. of Orthopsychiatry* 46, 4: 646–659.

Mayer, J., M. Motz, and A. Rosenblatt (1973) "The readmission patterns of patients referred to aftercare clinics." *J. of Bronx State Hospital* 1, 4.

Mendel, W. M. and S. Rappaport (1963) "Outpatient treatment for chronic schizophrenic patients." *Archives of General Psychiatry* 8: 190–196.

Michaux, W., et al. (1969) *The First Year Out: Mental Patients After Hospitalization.* Baltimore: Johns Hopkins Press.

Purvis, S. A. and R. W. Miskimins (1970) "Effects of community follow-up on post-hospital adjustment of psychiatric patients." *Community Mental Health J.* 6, 5: 374–382.

Smith, W., J. Kaplan, and D. Siker (1974) "Community mental health and the seriously disturbed patient." *Archives of General Psychiatry* 30: 693–696.

Solomon, P. and W. Doll (1979) "The varieties of readmission: the case against the use of recidivism rates as a measure of program effectiveness." *Amer J. of Orthopsychiatry* 49, 2: 230–239.

Winston, A., H. Pardes, D. S. Papernik, and L. Breslin (1977) "Aftercare of psychiatric patients and its relation to rehospitalization." *Hospital and Community Psychiatry* 28, 92: 118–121.

Wolkon, G., C. Peterson, and A. Rogawski (1978) "A program for continuing care: implementation and outcome." *Hospital and Community Psychiatry* 29, 4: 254–256.

Ann Solberg is a research psychologist affiliated with the Fresno County Department of Health in Fresno, California. In addition, she is an instructor at the California School of Professional Psychology, Fresno Campus. Her primary interest is evaluation research as applied to mental health programs.

Explanation and Critique

One of the most interesting facets of program or policy evaluation is that real-world projects do not always exactly fit the model. In fact, they seldom do. The article by Ann Solberg is a perfect example. As we have seen, Solberg's research reported on a project to evaluate the effectiveness of community follow-up services in reducing both rehospitalization and the overall cost of mental health care of individuals hospitalized at the Fresno County Department of Health Acute Psychiatric Unit (APU). Solberg details:

> Clients discharged from the APU were randomly assigned to either the experimental group or the control group. The control group received only those aftercare services for which they completed

> referrals or services that they sought on their own. The experimental group received additional posthospital follow-up services from a team of four psychiatric social workers for a period of up to 30 days following their discharge from the APU. The follow-up team worked with the clients, their families, and other treatment staff with the goals of improving clients' personal and community adjustments and increasing their involvement with aftercare programs (outpatient, partial day, or residential) to which they were referred by APU staff. The follow-up services were treated as an additional aftercare service. They were intended to supplement rather than substitute for aftercare services specified in a client's discharge plan.

When one thinks of the posttest-only control group design, eliminating the pretest comes to mind. For example, one might envision an experimental reading program in which students are randomly assigned to experimental and control groups and are given reading tests before and after the program. In the posttest-only design, the pretest is simply eliminated—knowing that random assignment has, in effect, made the two groups equal and thus the pretest unnecessary.

In Solberg's evaluation, however, a pretest was not possible. What could be pretested? Nothing. Thus, in evaluations such as this, the pretest-posttest design and the Solomon Four-Group Design are not used because they cannot be used. Also, as we know, the design is a valid experimental design in its own right and even has advantages over other forms of experimental designs (e.g., the threat of the testing effect).

In examining Solberg's research, consider her application of the posttest-only control group design. The randomized assignment was obviously the key to the validity of the design, but this was not the only strength of the paper. For one thing, Solberg provided data on the characteristics of the experimental group and the control group (gender, age, psychosis, neurosis, global impairment rating). Given the fact that randomized assignment of the patients was used, presentations of such comparisons was not really necessary. In the introduction to Part II of this volume, "Experimental Designs," it was stated that as long as the number of subjects is sufficiently large, random assignment guarantees that the characteristics of the subjects in the experimental and control groups are statistically equivalent.

With only seventy-one subjects in the experimental group and seventy-two in the control group, Solberg apparently wanted to assure the reader that the groups were, in fact, statistically equivalent. A nice touch.

This discussion has emphasized the fact that the key to experimental designs is random assignment. Random assignment, again, means that the treatment or program to which the subject is assigned bears no relation to any characteristic of the subject. It is important, however, that random assignment be distinguished from random selection.

To illustrate, look again at the Solberg article. Solberg first selected the subjects of her study from the Fresno APU. To achieve true external validity, she should have selected subjects at random from the population of persons discharged from psychiatric hospitals. For Solberg to do this, however, the costs would have been prohibitive. Instead, she randomly assigned all subjects discharged from the Fresno County Department of Health Acute Psychiatric Unit who met the inclusion criteria to experimental and control groups. This procedure reduced the generality of findings, but it also reduced costs and random errors. Random selection is not a requisite for a true experimental design (Langbein 1980, 67–68). Solberg abandoned random selection entirely (but not random assignment) but still has a valid experimental design. However, she lost generalizability.

This is an example of one of the trade-offs discussed in chapter 1. To her credit, Solberg was very careful about this. She reported the following:

> The results showed that community follow-up services, as provided in this demonstration project, provide an effective means of reducing the rehospitalization of individuals who have been hospitalized in the short-term acute psychiatric facility *at the Fresno County Department of Health*. [emphasis added]

Solberg is certainly correct in not generalizing beyond this unit.

Overall, although we may have questions about some segments of Solberg's article (e.g., How were hospitalization costs determined?), the research reported provides a fine example of the posttest-only control group design.

References

Nachmias, David. 1979. *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's.

Bausell, R. Barker. 1986. *A Practical Guide to Conducting Empirical Research.* New York: Harper and Row.

Mohr, Lawrence B. *1988. Impact Analysis for Program Evaluation.* Chicago: Dorsey.

Langbein, Laura Irwin. 1980. *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation.* Glenview, Ill.: Scott, Foresman.

PART III

Quasi-Experimental Designs

It is the absence of random assignments that distinguishes quasi-experiments from true experiments. Quasi-experiments employ comparison groups just as do experiments. In the quasi-experimental design, the researcher strives to create a comparison group that is as close as possible to the experimental group in all relevant respects. And, just as there are a variety of experimental designs, there are a similar variety of basic designs that fall into the quasi-experimental category. These include the pretest-posttest comparison group design, the regression-discontinuity design, the interrupted time-series comparison group design, and the posttest-only comparison group design.

Laura Langbein identifies a number of ways in which quasi-experiments are distinct from randomized experiments in practice. First, quasi-experiments tend to be retrospective—that is, to occur after a program is in place. When the decision to evaluate comes after the program has been implemented, the option of conducting a true experiment is usually foreclosed.

Second, the internal validity of most quasi-experiments is usually more questionable than that of experiments. The researcher must identify and measure all the relevant characteristics of the experimental group and attempt to construct a similar comparison group. Successfully executing this task is frequently problematic (Langbein 1980, 87–89). Although quasi-experiments are not true experiments, they are far from useless. Peter Rossi and Howard Freeman are actually quite positive about them:

> The term "quasi-experiment" does not imply that the procedures described are necessarily inferior to the randomized controlled experiment in terms of reaching plausible estimates of net effects. It is true that, without randomization, equivalence . . . cannot be established with as much certainty. The possibility always remains that the outcome of a program is really due to a variable or process that has not been considered explicitly in the design or analysis. However, quasi-experiments, properly constructed, can provide information on impact that is free of most, if not all, of the confounding processes (threats to validity). . . . Indeed, the findings from a properly executed quasi-experimental design can be more valid than those from a poorly executed randomized experiment. (1985, 267)

References

Langbein, Laura Irwin. 1980. *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation.* Glenview, Ill.: Scott, Foresman.

Rossi, Peter H. and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach.* 3d ed. Beverly Hills, Calif.: Sage.

CHAPTER 8

Pretest-Posttest Comparison Group Design

The pretest-posttest comparison group design, sometimes referred to as the nonequivalent group design, is similar in every respect to the pretest-posttest control group design except that the individuals (cities, organizations) in the program being evaluated and in the comparison group are not assigned randomly—the comparison group is nonequivalent. The researcher attempts to identify a group for comparison comparable in essential respects to those in the program. The validity of the quasi-experiment depends in large part on how closely the comparison group resembles the experimental group in all essential respects. Table 8.1 illustrates the design.

Obviously a multitude of uses exist for this design, and, in fact, it is one of the designs most frequently found in the literature. If random assignment is not possible, researchers attempt to find a group similar to the experimental group to use as a "pseudo-control." In many cases this is extremely difficult. In fact, it is so difficult that one research scholar, Barker Bausell, recommends against its use, as follows:

> Although this is probably a minority opinion, I recommend against the use of nonequivalent control groups, especially for beginning researchers. In many ways, quasi-experimental research (as these models are sometimes called) is more difficult to perform well than research involving the random assignment of subjects. Such research is also seldom conclusive, since it is subject to so many alternative explanations. (1986, 138)

One way in which evaluators seek to establish a comparison group as similar as possible to the experimental group is through the matched-pair design. This approach assigns subjects to experimental and "control" conditions on the basis of some common characteristics the evaluator wishes to measure. For example, in education research, students in an experimental group might be matched with other students on the basis of grade and reading level to form a comparison group. Only these two factors are controlled. Other factors that could affect the outcome (home, environment, intelligence) are not controlled.

When is this design most often used? It is used in retrospective analysis when it is too

TABLE 8.1
The Pretest-Posttest Comparison Group Design

	Pretest	Program	Posttest
Group E	O	X	O
Group C	O		O

late to assign a control group. It is also frequently used in evaluating programs in which the participants are volunteers and thus a true control group cannot be found.

Finding an adequate comparison group for a volunteer program is precisely the problem that was facing Tim Newcomb in the article to follow: "Conservation Program Evaluations: The Control of Self-Selection Bias."

Supplementary Readings

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* Cambridge, England: Cambridge University Press, chap. 6.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 4.

Rossi, Peter H., Howard E. Freeman, and Mark W. Lipsey. 1999. *Evaluation: A Systematic Approach.* 6th ed. Thousand Oaks, Calif.: Sage, chap. 9.

READING

Conservation Program Evaluations
The Control of Self-Selection Bias

Tim M. Newcomb
Seattle City Light

The Seattle City Light Department evaluated its weatherization program for low-income homeowners to determine how much electricity was conserved, and whether the program was cost-effective. The study employed a control group of future program participants in order to avoid the biased estimate that could result because participants are a voluntary, nonrandom group of utility customers. Surveys, energy audits, and electricity meter data confirmed the similarity of the experimental and control groups prior to home weatherization. Following the installation of conservation measures, the consumption of the experimental group declined by an average of 3400 Kwh per year, when compared with the control group. This experimental design is believed to have nullified the potential biases due to selection, testing, regression, history, and instrumentation. The bias due to mortality, defined in this study as customer turnover, was seen to be very minor. This design can be applied to evaluate programs that operate continuously, provided that the guidelines for the program do not change and that clearly defined measurements of program success are available.

PUBLIC AND PRIVATE electric utilities across the United States have developed a variety of conservation programs for their residential customers during the past five years. This effort has been spurred in part by the federal Residential Conservation Service, which began in 1979 and requires the larger utilities to offer a home energy audit service to their customers. From the point of view of the utilities, these programs are beneficial to the extent that they reduce the residential load in a cost effective way.

Precise analyses of the cost-effectiveness of these programs are difficult to make because relatively little is known about the amount of electricity conserved by the programs. Several recent reviews have noted the lack of accurate research on residential electricity conservation and have called for better evaluation strategies to measure conservation (Berry, 1981; Pease, 1982). In the early stages of conservation program planning, estimates of potential electricity savings came from engineering studies. Recent evaluation studies by Seattle City Light

(Weiss and Newcomb, 1982; Newcomb, 1982) and by Oak Ridge National Laboratory (Hirst et al., 1983a) found that conservation estimates developed prior to program implementation were considerably larger than the observed amount of conservation.

Inaccuracies in engineering estimates may be due, in part, to the tremendous variation in electricity consumption patterns among residential customers. Olsen and Cluett (1979) and Hirst et al. (1983a) report that some of this variation can be explained by differences in home size, income level, and number of occupants. More than half of the variation among households remains after these factors are accounted for. Engineering studies, which are conducted in a few homes, cannot take into account the complex relationships among unique home dwellers, unique homes, and varying combinations of energy conservation measures. For these reasons, accurate estimation of the conservation achieved by a residential program must be based upon actual field measurements of electricity use by the participating population.

The purpose of this article is to describe a quasi-experimental design for estimating average household electricity conservation that controls for the major threats to internal validity and does not require complex statistical methods for analysis. The design uses a nonequivalent control group as described by Campbell and Stanley (1966) consisting of residential customers who participated in the same conservation program at a later date. This research strategy is useful in situations where a program operates over a period of several years and maintains constant and restrictive guidelines for accepting participants. Under these conditions, early participants will resemble later ones with respect to important preprogram characteristics.

This approach to estimating electricity conservation has been applied to three residential conservation program evaluations at Seattle City Light. The evaluation of electricity savings for the Low Income Electric Program (LIEP) will be presented here as an example of the approach.

Seattle City Light serves approximately 270,000 residential customers and plans to weatherize 20,000 homes by 1989 through the LIEP. This is expected to yield between 60 and 70 million kilowatt-hours of conserved electricity each year. The predicted cost of the weatherization over 8 years is 25 million in 1980 dollars. The LIEP offers a free home energy audit and free home weatherization to applicants whose total annual household income is less than 90% of the median income for the Seattle-Everett Standard Metropolitan Statistical Area. All applicants must accept ceiling and underfloor insulation as needed, as well as duct insulation, hot-water tank insulation, and repairs that are judged necessary by the utility auditors. Wall insulation, pipe insulation, and weatherstripping and caulking are optional. The participants included in this analysis all had electric heat as their primary source of space heat. The average 1981 LIEP participant received approximately $1400 in weatherization materials and labor at no cost.

The average cost to Seattle City Light for a weatherization job was approximately $2500 in 1981, including administration, home energy audits, and inspection of weatherization. An accurate assessment of the electricity savings is essential to evaluate and justify such large expenditures.

Reasons for Selecting LIEP Participants as a Control Group

A simple pre-post experimental design could not provide accurate program conservation estimates for several important reasons. Weather fluctuations from one year to the next have a strong impact upon electricity use in homes that use electric space heat, and could easily mask or magnify the effects of home weatherization. Seattle City Light increased its residential rates approximately 65% in August

of 1980, causing its customers to decrease their consumption of electricity. This effect coincided with the LIEP weatherization and, while the rate hike cannot be separated cleanly from the effects of the program, it probably produced reductions in electricity use beyond the results of weatherization. There were no time series data available of sufficient quality to allow the estimation of weather and rate hike effects using multiple regression.

All of these reasons demonstrate the need for the inclusion of a control group in this study that is as similar as possible to the LIEP participants. No group of residential customers fitting this description had been selected for analysis prior to the program, and current household income data were not available for any group except participants. Further, since the program is voluntary, participants almost certainly differ from the average low-income customer in attitudes and education. Recent studies by Olsen and Cluett (1979) and by Berry (1981) report that participants in voluntary residential conservation programs tend to have more education and use more electricity than the average customer. Without knowing beforehand exactly what the attributes of the program participants were, the best choice for a control group was the group of people who would later sign up for the program. This choice was confirmed through a comparison of survey data and electricity consumption figures for the LIEP weatherized homes and the control group to be presented in a later section.

A description of the experimental and the control groups, and the general approach to calculating electricity conservation estimates from the program are outlined as follows:

- The program participants, or "experimental group," received home conservation measures in 1981. The electricity use for this group prior to the program (September 1980 to May 1981) will be referred to as E_1. The electricity use for this group after the program weatherized the homes (September 1981 to May 1982) is referred to as E_2.
- The "control group" of homeowners received home conservation measures in 1982. The electricity consumption during the period from September 1980 to May 1981 will be called C_1. The electricity in use from September 1981 to May 1982 for this group is C_2. In this case, C_2 represents electricity use for the control group before the homes were weatherized.
- The computation of the estimates of conservation due to the program involves computing the pre-to-post change in consumption for the experimental group, $E_2 - E_1$, and the corresponding change for the control group, $C_2 - C_1$. The latter difference is subtracted from the former to find the estimate of conserved electricity, which is $(E_2 - E_1) - (C_2 - C_1)$.

Estimating the Impact of Weatherization on Electricity Use

Comparison of the Preprogram Attributes of the Weatherized Homes and the Control Homes

Three sources of information about the experimental and control groups of homes were available. Data collected at the time of the home energy audit describe the age and size of the home, the number of occupants, and the type of space heating. A mail survey of a random sample of each of the groups gathered information on the conservation actions taken in addition to the LIEP, and on the reasons for the homeowner's participation. Finally, the utility's customer metering system provided electricity consumption figures for each customer in periods of two months. Each of these sources can be used to test the assumption that the two groups were similar before the weatherization occurred.

Table 1 compares the groups with respect to average heated floor area, percentage of

homes having each type of electric space heat, and average number of nighttime occupants. The reports by Hirst et al. (1983a) and by Olsen and Cluett (1979) showed that home size and number of occupants were positively correlated with level of consumption. The fact that there is little difference between the experimental and control groups for these two measures is strong support for the argument that the groups are similar. The close similarity in types of space heating is also significant because different types of space heating respond differently to fluctuations in temperature.

The survey responses to the mailed questionnaire are shown in table 2. The response rates to these surveys were 81% and 76% for the experimental group and the control group, respectively, indicating that the responses are representative of the respective survey populations sampled. There is little difference with regard to the respondents' stated reasons for participating in LIEP, except for a small increase in those who mentioned saving on electricity bills in 1982. The percentage of respondents who took each of the four conservation actions in addition to the LIEP is also very similar, which shows that any measured effects of the program weatherization are not affected by differences in extraprogram weatherization. The two groups appear to resemble each other with respect to their motivation and their reliance on the program for changes in home weatherization.

The single most critical test of the similarity of the two groups must be their respective levels of electricity use during the same time period before the experimental group received weatherization. Any differences in either the amount of electricity used or in the response curves of electricity use to weather would require adjustments before the impact of the program could be measured. Table 3 presents the average electricity consumption figures and the standard errors for these averages, for five two-month billing periods prior to LIEP participation. The differences between the groups are not significant at the 5% level for any of the five two-month periods. Since electricity use is

TABLE 1

LIEP Participants, 1981 and 1982: Selected Demographic and Dwelling Characteristics

	1981 Participants Experimental Group (N = 326)	*1982 Participants Control Group (N = 227)*	*t-Tests of Differences Between Groups Two-Tailed*
Floor area of home			
Mean square feet	1258	1323	t = –1.26
Standard Error	35	38	p > .10
Number of occupants			
Mean number	2.6	2.5	t = 0.71
Standard error	0.1	0.1	p > .10
Type of electric heat			
% Furnaces	20.4	21.6	t = –0.34
% Baseboard	70.6	73.6	p > .50
% Other	3.5	3.0	t = –0.77
			p > .10
			t = 0.33
			p > .50

the criterion variable in this analysis, the similarity offers strong support for the choice of later participants as the control group.

The evidence for similarity was judged strong enough to support the estimation of electricity conservation using the calculations described earlier. A further discussion of the advantages and disadvantages of this design follows the estimation of the annual program conservation.

Calculation of the Electricity Conservation Estimates

The description of the quasi-experimental design that has been developed up to this point is oversimplified. This is so because homes were weatherized under LIEP continuously during 1981 and 1982, rather than in a short period. Figure 1 shows that as homes were weatherized in 1981, they were dropped from the preprogram consumption averages, and added to the postprogram averages. As homes in the control group were weatherized in 1982, they were dropped from the control group consumption averages.

All of the preprogram bimonthly billing records for the experimental group were grouped by two-month periods and an average was computed for each period, although the number of cases in each preprogram period changed (decreased) as homes received weatherization and were removed from the preprogram group. In a similar fashion, all of the electricity billing data for postprogram periods for the experimental group were

TABLE 2

LIEP participants, 1981 and 1982: Survey Responses – Reasons for Participating in LIEP and Conservation Actions Taken Outside LIEP

	1981 Participants Experimental Group (N = 197) %	*1982 Participants Control Group (N = 110) %*	*t-Tests of Differences Between Groups Two-Tailed*
Reasons for participating			
Program is free	70	70	t = 0
Save on bills	89	97	t = –2.85 p < .01
Believe in conservation	75	77	t = –0.40 p > .50
Increase value of home	55	58	t = –0.51 p > .50
Increase comfort of home	80	77	t = 0.61 p > .50
Help solve energy crisis	61	57	t = 0.68 p > .10
Conservation actions taken outside LIEP within 12 months of weatherization			
Installed thermal windows	10	12	t = –0.54 p > .50
Installed storm windows	16	21	t = –1.06 p > .10
Installed energy-efficient water heater	9	7	t = 0.63 p > .50
Installed solar water heater	1	1	t = 0

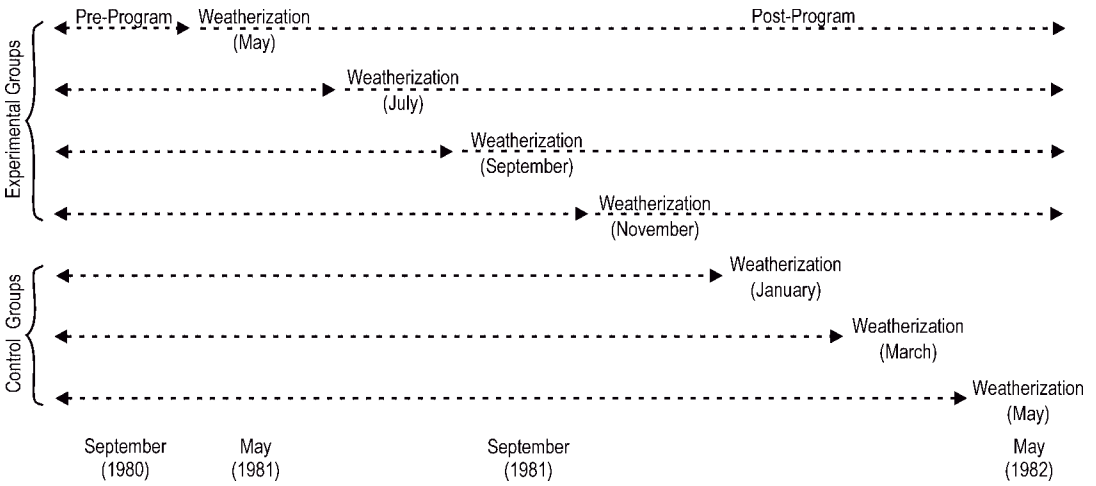


FIGURE 1

grouped by period and averaged. The number in each billing period changed (increased) over time as more homes were weatherized.

The control group remained stable throughout the preprogram period. However, as these homes entered the program in early 1982, they were dropped from the control group, and the number of cases available for control purposes declined during successive billing periods in 1982 through May of 1982, which was the last period used in the analysis.

The varying numbers of homes in the two groups is a problem that will occur in a design of this type unless the treatment is applied to the entire group of participants in a brief period. The evaluation of programs operating in a natural setting is more likely to face the continuous mode of program operation described here. An accurate system for tracking and assigning participants to one or the other of the two groups as the treatment occurs is helpful for a design of this kind.

Figure 2 is a graph of each two-month electricity consumption average for the experimental group and the control group using the method described above. While the two groups are similar during the period up to September 1980, they diverge after that time. The small difference during the September 1980 period is due to the relatively minor impact of weatherization during summer months when consumption for space heating is very low. The number of heating degree days during the period from November 1980 to April 1981, was only 87% of the number of heating degree days from November 1981 to April 1982. The colder winter of 1981–1982 had an obvious impact upon the consumption of the control group homes in Figure 2. This increased consumption was not shared by the experimental group; however, a simple pre-post design would have reduced the estimate of electricity conservation over this period for the experimental group because the colder winter raised the consumption of electricity by comparison with the previous year.

Table 3 displays the mean values for the experimental and control groups, before and after the experimental group received weatherization. The t-test values for each of the ten differences between mean values are listed in the third column with the probability that such a value would occur by chance. One-tailed t-tests were used for both the before and after periods, because the hypothesis to be tested is that the experimental group does not have a lower bi-

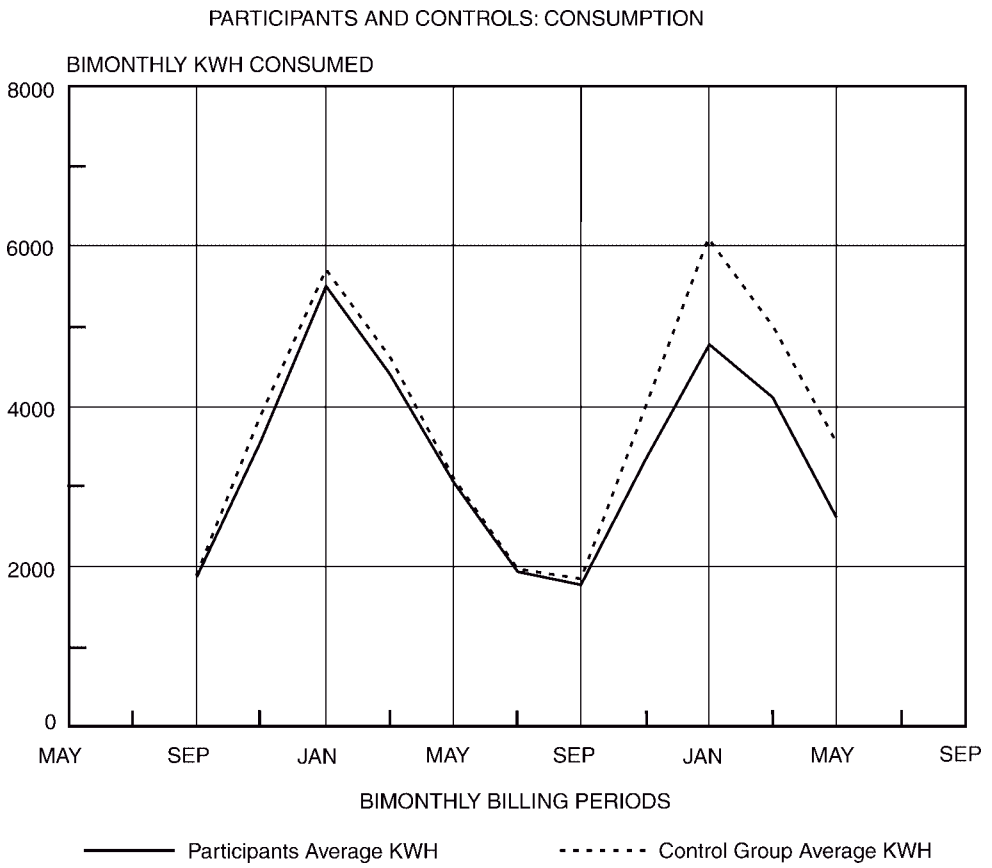


FIGURE 2
LIEP Electricity Consumption (KWH) 1980–1982

monthly electricity level at each period than the control group. While all of the preperiod differences between means are insignificantly different from zero, four of the five postperiod differences are highly significant.

The actual computations of electricity conservation are presented in Table 4. The first column contains the pre-to-post changes by billing period for the experimental group, referred to earlier as E_2 - E_1. The corresponding differences for the control group are given in column 2, and referred to as C_2 - C_1. The NET CHANGE, (E_2 - E_1) - (C_2 - C_1), is computed in column three. The total net change for the ten-month period under study is -3083Kwh. To compute an estimated annual amount of electricity conservation, the ratio of electricity use by the control group for one full year (July 1980 to May 1981) to that group's use for ten months (September 1980 to May 1981), equal to 1.11, was multiplied by -3083 Kwh/10 months. The result is -3422 Kwh/year of electricity conservation. The value of the conserved electricity to the average LIEP participant is $485 (1981 dollars) over the 30-year lifetime of the weatherization measures (Weiss, 1983; this is also the net present value to the customer since the weatherization was free).

TABLE 3

LIEP Participants, 1981 and 1982: Electricity Consumption Before and After Weatherization of the Experimental Group

Bimonthly Billing Period	*1981 Participants Experimental Group*	*1982 Participants Control Group*	*t-Tests of Differences Between Groups One-Tailed*
Preperiod			
September 1980			
Mean Kwh/2 months	1893	1902	t = –0.08
Standard Error	98	55	p > .05
N	68	201	
November 1980			
Mean Kwh/2 months	3686	3829	t = –0.70
Standard Error	160	128	p > .05
N	113	208	
January 1981			
Mean Kwh/2 months	5195	5434	t = –1.13
Standard Error	155	144	p > .05
N	185	210	
March 1981			
Mean Kwh/2 months	4389	4567	t = –1.11
Standard Error	105	121	p > .05
N	293	209	
May 1981			
Mean Kwh/2 months	3010	3114	t = –0.91
Standard Error	76	86	p > .05
N	293	209	
Postperiod			
September 1981			
Mean Kwh/2 months	1780	1871	t = –0.85
Standard Error	91	57	p > .05
N	68	209	
November 1981			
Mean Kwh/2 months	3232	3960	t = –3.95
Standard Error	137	123	p < .05
N	113	209	
January 1982			
Mean Kwh/2 months	5031	6169	t = –5.03
Standard Error	156	164	p < .01
N	185	208	
March 1982			
Mean Kwh/2 months	4116	5027	t = –4.92
Standard Error	98	157	p < .01
N	293	160	
May 1982			
Mean Kwh/2 months	2631	3519	t = –4.15
Standard Error	70	202	p < .01
N	293	39	

TABLE 4

LIEP Participants, 1981 and 1983: Computation of Electricity Conservation from Bimonthly Changes in Consumption of the Experimental and Control Groups

Bimonthly Comparison Periods	*Changes for the 1981 Participants Experimental Group* $E_2 - E_1$	*Changes for the 1982 Participants Control Group* $C_2 - C_1$	*Net Change Due to LIEP Program Weatherization* $(E_2 - E_1) - (C_2 - C_1)$
September 1981 minus September 1980	–113 Kwh/2 mo.	–31 Kwh/2 mo.	–82 Kwh/2 mo.
November 1981 minus November 1980	–454	131	–585
January 1982 minus January 1981	–164	735	–899
March 1982 minus March 1981	–273	460	–733
May 1982 minus May 1981	–379	405	–784

Note: Net change over ten-month period: –3083 Kwh.

Discussion

From the point of view of quasi-experimental design theory, the design described here has some significant strengths. The assumption that the control and experimental groups are alike can be examined carefully. The somewhat unique group of customers that participated in the LIEP can be compared to another very similar group without performing a multiple-regression analysis of the relationship between income levels, electricity consumption, and weatherization levels.

Most of the threats to internal validity described by Campbell and Stanley (1966) can be discounted in this study. The problem of selection in the LIEP study is a recurrent one in residential energy conservation research (Hirst et al., 1983b). Since voluntary participants may be unlike a random sample of customers, this difference must either be controlled through the proper design, or counteracted through statistical techniques. The nonequivalent control group design used here effectively dispenses with this threat through its comparison of two groups of customers who have volunteered for the same program during adjacent time periods.

The problem of mortality in the experimental group has its counterpart here in those LIEP participants who had to be excluded from analysis either because they moved into homes shortly before the weatherization, or moved out shortly after the weatherization. For these cases there were not enough two-month billing periods under one owner to permit pre-post comparison. Approximately 10% of the controls and 13% of the experi-

mentals were discarded for this reason. An examination of the impact of change of ownership on consumption shows that in 50% of the cases the level of electricity consumption changed by more than 25% with the ownership change. Given this large impact, a comparison of preprogram and postprogram levels regardless of ownership is not possible. These large changes underscore the high level of variability in electricity use among the single-family customers of Seattle City Light.

Table 5 presents preprogram consumption data for the experimental group and for a sample of homes excluded due to ownership changes. The comparison suggests that the excluded cases may have used slightly less electricity than the cases that were included in the analysis. The result of this exclusion may have been to overestimate the average conservation due to the program.

The use of future program participants has limitations for program evaluation that stem from the method's dependence upon constant program guidelines for participation. As the requirements for entry into a program change, the possibility of finding a similar group of future participants disappears.

The design employed in this analysis can only be used when a program operates continuously under relatively unchanging guidelines for a sufficient period to permit the accumulation of large groups for statistical analysis. Even under these conditions, such a design will be difficult to defend if the dependent variables are qualitative, or less well-defined than bimonthly electricity consumption. In such cases, it must be defended on theoretical grounds rather than by comparing preprogram characteristics of the experimental and control groups.

TABLE 5

The LIEP Experimental Group and Excluded LIEP Cases Average Preprogram Electricity Use for Four Billing Periods

Bimonthly Billing Period	*1981 Participants Experimental Group*	*1981 Participants Excluded*	*t-Tests of Differences Between Groups Two-Tailed*
September 1980			
Mean Kwh/2 months	1893	1840	t = 0.29
Standard Error	98	155	p > .50
N	68	20	
November 1980			
Mean Kwh/2 months	3686	3837	t = –0.41
Standard Error	160	336	p > .50
N	113	24	
January 1981			
Mean Kwh/2 months	5195	4870	t = 0.90
Standard Error	155	328	p > .10
N	185	28	
March 1981			
Mean Kwh/2 months	4389	3928	t = 1.55
Standard Error	105	278	p > .10
N	293	25	

References

Berry, L. (1981) *Review of Evaluations of Utility Home Energy Audit Programs*, ORNL/CON-58. Tennessee: Oak Ridge National Laboratory.

Campbell, D. T. and J. C. Stanley (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Hirst, E., E. Broneman, R. Goeltz, J. Trimble, and D. Lerman (1983a) *Evaluation of the BPA Residential Weatherization Pilot Program*, ORNL/CON-124. Tennessee: Oak Ridge National Laboratory.

____J. Trimble, R. Goeltz, and S. Cardell (1983b) *Methods to Deal with Self-Selection in Estimating Conservation Program Energy Savings: Use of Synthetic Data to Test Alternative Approaches*, ORNL/CON-120. Tennessee: Oak Ridge National Laboratory.

Newcomb, T. M. (1982) *Electricity Conservation Estimates for the Low Income Electric Program.* Washington: Seattle City Light Department.

Olsen, M. and C. Cluett (1979) *Evaluation of the Seattle City Light Neighborhood Energy Conservation Program.* Seattle, WA: Battelle Human Affairs Research Center.

Pease, S. (1982) *An Appraisal of Evaluations of Utility-Sponsored Programs for Residential Energy Conservation*, N-1925-DOE. Washington, DC: U.S. Department of Energy.

Weiss, C. (1983) *Low Income Electric Program Cost-Effectiveness Analysis.* Washington: Seattle City Light.

____and T. M. Newcomb (1982) *Home Energy Loan Program Energy Savings Analysis.* Washington: Seattle City Light.

Tim M. Newcomb is a Program Evaluator in the Conservation Division of Seattle City Light, Seattle, Washington. His current interests are in using Arima Time Series models to estimate electricity conservation, and in the use of surveys to market conservation programs.

Explanation and Critique

Developing an evaluation is an exercise of the dramatic imagination. The art of evaluation involves using the science of evaluation to create a design and gather information that is appropriate for a specific situation and a particular policymaking context. The quasi-experiment with constructed controls (comparison group) provides the researcher with the opportunity to exercise artistic creativity. Tim Newcomb rose to this challenge. "Conservation Program Evaluations: The Control of Self-Selection Bias" is absolutely elegant.

Seattle City Light operated a Low Income Electric Program (LIEP) that offered a free energy audit and free home weatherization to applicants whose total annual household income was less than 90 percent of the median income for the Seattle-Everett Standard Metropolitan Statistical Area. The average 1981 LIEP participant received approximately $1,400 in weatherization materials at no cost. The average cost to Seattle Light for each job was $2,500, including administration and energy audits. The company planned to weatherize 20,000 homes by 1989 through LIEP at a cost of $25 million in 1980 dollars. Given such cost, it was imperative to develop an approach to accurately estimate electricity conservation that could be attributed to the program. How could this be done?

Because the program was voluntary, Newcomb could not use an experimental design and randomly assign low-income utility customers into program and control groups. Furthermore, a comparison of low-income utility participants with nonparticipants would be inappropriate because simply by volunteering for the program a select group of low-income utility customers showed their concern with utility costs. They thus might be assumed to conserve more in response to the 1980 rate increase than nonvolunteers. (This rate in-

crease of some 65 percent was an effect of history which could have had an impact on the evaluation without the comparison group.) The simple pretest-posttest design could not be used because of year-to-year fluctuations in the weather. Nor were time-series data available to allow the estimation of weather and rate hike effects using multiple regression.

Thus, Newcomb was faced with the need to find a comparison group for the study as similar as possible to LIEP participants. Obviously, no true control group was available. He described his predicament:

> No group of residential customers fitting this description had been selected for analysis prior to the program, and current household income data were not available for any group except participants. Further, since the program is voluntary, participants almost certainly differ from the average low-income customer in attitudes and educations.

Newcomb found his comparison group and overcame the self-selection problem by using as "constructed controls" the group of people who later signed up for the program. The characteristics of the comparison group in terms of housing, income, preweatherization electricity use, and so on were proven to be comparable. Most importantly, the self-selection problem was overcome through Newcomb's unique design. The pretest-posttest comparison group design used in this study effectively dispensed with the selection threat to validity through its comparison of two groups of customers who volunteered for the same program during adjacent time periods.

Substantively, what did this evaluation show? Newcomb compared the energy consumption of both groups for the before-and-after periods by using the one-tailed *t*-test. The *t*-test was used to determine whether the average energy consumption of weatherized residents was different from the average consumption by those not weatherized. Thus Newcomb used a difference of means test with the *t*-statistic to determine whether there was a statistically significant difference between the means. In general, the *t*-test is an inferential statistic used to determine whether the null hypothesis, the test or alternate hypothesis (that there is a difference) is accepted. The one-tailed *t*-test was appropriate in this case because Newcomb's hypothesis was that the experimental group would have lower bimonthly electricity consumption rates than the control group. When a direction (lower) is implied in a hypothesis (rather than just "different"), a one-tailed test is dictated.

Newcomb found that there was no significant difference between the energy consumption of the two groups in all the pretest periods. After the homes of the experimental group had been weatherized, however, there was a significant difference in energy consumption between the two groups for four of the five periods. It was only in the warmest months that the differences between the two groups was not significant.

By using this carefully selected comparison group, Newcomb was able to show that the program resulted in a savings of 3,422 kilowatt hours per year for the average participant. The value of the conserved electricity (1981 dollars) amounted to $485 per participant over the 30-year lifetime of the weatherization measures. Given variations in the temperature from winter to winter, Newcomb never could have made these calculations without having a credible comparison group. More simply, because winters are different, the energy consumption of the families participating in the program will vary from year to year; therefore, it is not possible to simply compare this year's energy use with last year's.

Reference

Bausell, R. Barker. 1986. *A Practical Guide to Conducting Empirical Research.* New York: Harper and Row.

CHAPTER 9

Interrupted Time-Series Comparison Group Design

THE INTERRUPTED TIME-SERIES comparison group design is a very strong quasi-experimental design. This design examines whether and how an interruption (of treatment, program, etc.) affects a social process and whether the observed effect is different from the process observed in an untreated group or among different types of treatments. The design assumes that data are collected at multiple points before the treatment and at multiple points after the treatment and that the same data are collected for units not treated. The design is schematically shown in table 9.1, in which "O" is the observation, "t_1" and "t_2" are treatment groups 1 and 2, respectively; "c" is the untreated comparison group; "$t-1$" and "$t-2$" are observation times at two points before the treatment, whereas "$t+1$" and "$t+2$" are observation times following the treatment; and "X_1" and "X_2" denote two different treatments.

TABLE 9.1

Interrupted Time-Series Comparison Group Design

Before			After	
O_{t_1t-2}	O_{t_1t-1}	X_1	O_{t_1t+1}	O_{t_1t+2}
O_{t_2t-2}	O_{t_2t-1}	X_2	O_{t_2t+1}	O_{t_2t+2}
O_{ct-2}	O_{ct-1}		O_{ct+1}	O_{ct+2}

For a design to be an interrupted time-series with comparison group, the comparison group can be an untreated group similar with respect to the treated group on as many variables are possible, or the comparison can be made to similar units receiving a different form of the treatment, or both. This is a strong design because the nature of the time-series design eliminates the bias that results when one makes only one observation of a phenomenon. In other words, when data are collected at only one point, the researcher must determine whether the observation is reflective of the normal trend in the data. Without this assurance, the researcher cannot be sure that the collected data are not reflective of an abnormal high or low fluctuation. The model is strengthened further by the number of observations over time. Models with data collected at four points before and after the treatment are superior to models with two points but inferior to those with ten observation points before and after the treatment. This is reasonable because the more points of data, the better one can specify the trend and the more certain the research is that the obser-

vations are reflective of the social process. Consequently, researchers try to assemble as many data points before and after the treatment as possible or as resources will allow.

The type of intervention dictates the type of impact the researcher expects. In other words, the hypothesized impact is driven by theory—that is, knowledge of the form the impact will take. Laura Langbein provides a graphic presentation of the different forms the impacts can take in her book, *Discovering Whether Programs Work* (1980, 91); see figure 9.1. The following is a paraphrased description from her original work.

The patterns shown in figure 9.1(a) and 9.1(b) indicate that the program has had an immediate and abrupt impact. The treated group in figure 9.1(a) rises immediately after the treatment, but the untreated group remains basically unchanged. In figure 9.1(b), both groups continue an upward trend. However, the rapid change in slope for the treated group immediately after the treatment displays the program impact above and beyond that which one would expect through history or maturation. Moreover, the effect is permanent; the treated group's scores remain high in the period following the treatment. Langbein warns that investigators should not expect patterns like these when program impact is hypothesized to be gradual or incremental. Figures 9.1(c) and 9.1(d) suggest a delayed result. When this is the case, researchers should attempt to rule out any changes that occurred subsequent to the treatment that could explain the delayed change. Figure 9.1(e) could indicate a gradual intervention that was effective. Figure 9.1(f) shows a program that only increased the instability of the outcome measure. Thus, the researcher must have in mind the kind of impact he or she is expecting in order to interpret the results of the time-series analysis.

When using time-series analysis, the researcher can plot the outcome measures discretely (as demonstrated in figure 9.1) or smooth the curve by using a moving average to plot points. For example, the plotted point at "$t - 3$" would be the average of the outcome measures at "$t - 2$," "$t - 3$," and "$t - 4$"; the point at "$t - 2$" would be the average of "$t - 1$," "$t - 2$," "$t - 3$," and so on. Physically this removes some abrupt or drastic changes (noise) caused by spurious events. It captures the trend in the data. The fact that time-series analysis takes account of trends in the data and noise make it superior to simple before-and-after designs. ("Noise" refers to departures from the underlying trend — fluctuations above or below the trend line.) One can assess the impact of an interruption or treatment on a preexisting trend in a number of different ways. Perhaps the easiest is a simple, visual assessment of the data arrayed by year. According to Lawrence Mohr, this technique is most effective in "well-behaved" series—those without a great deal of noise and with visually consistent slopes (1988, 155–56). Simply, one tries to get an impression of the trend before the treatment and compares that with the posttreatment trend (sometimes referred to as the *transfer function*). Dramatic changes either in the slope or the intercept during the posttest time is indicative of the treatment's effects. It always is a good idea for evaluators to visually assess the relationships in their data to get a feeling for the form of the transfer function.

When there is a great deal of noise, it becomes difficult to assess trends. Unfortunately, social processes seldom exhibit purely linear trends. This problem is confounded by the fact that not all noise is "white noise"—that is, nonrandom disturbances or correlated error terms which result in what is called autocorrelation. Autocorrelation violates a fundamental assumption of ordinary least squares regression, that assumption being independent, random errors—among other things. Although violations of this assumption do not bias the regression coefficient (b) over infinite trials, the vari-

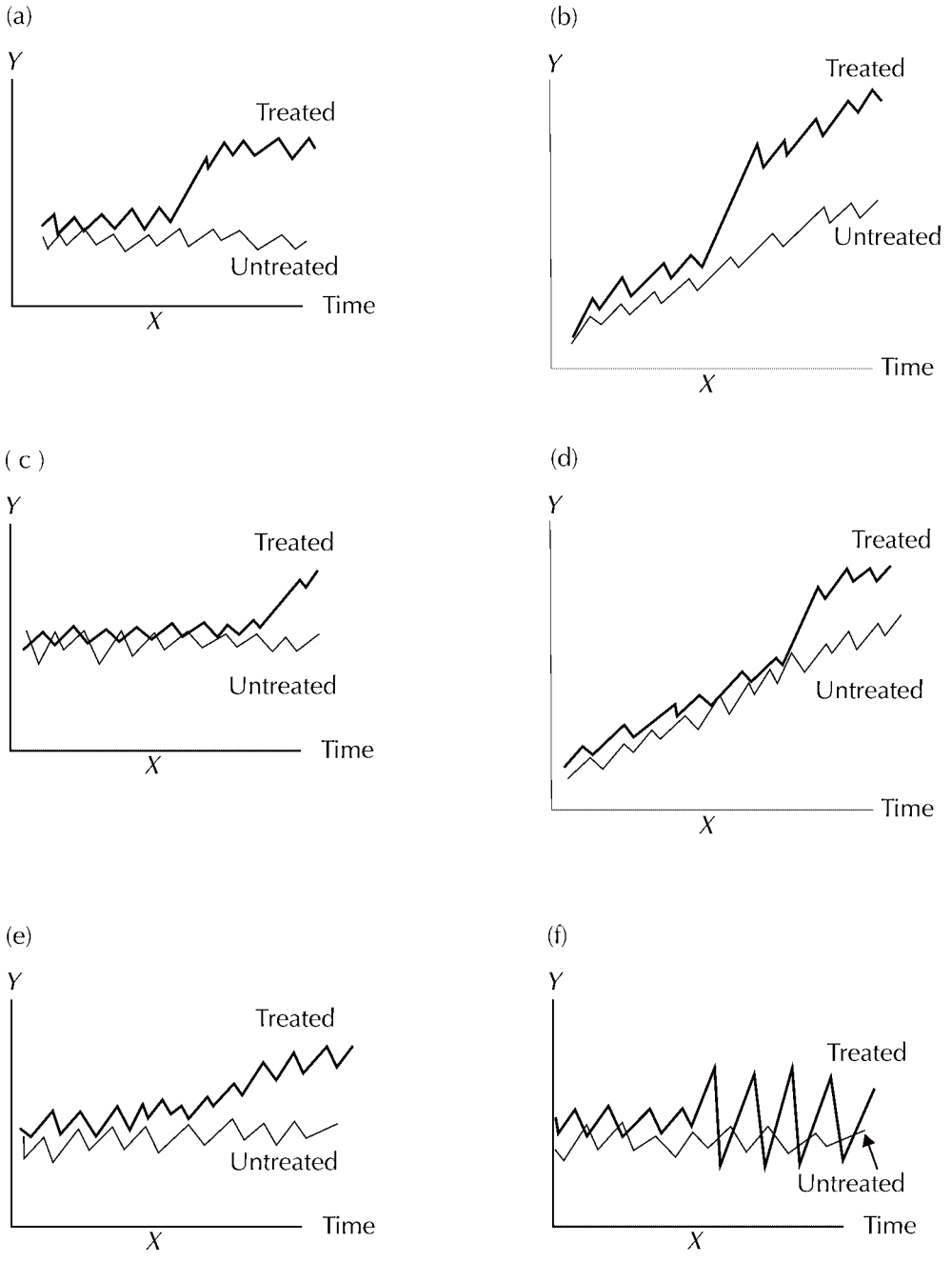


FIGURE 9.1
Possible Results from ITSCG Design

Source: Langbein 1980, 91.

ance estimates are understated. This not only makes it easier to get a statistically significant result but reduces the confidence that the particular *b* in a specific equation is accurate. Fortunately, there are ways to control for autocorrelation statistically. For example, the ARIMA procedure accomplishes that.

The availability of computer programs that easily compute time-series statistics has increased the number of time-series evaluations. Systematic collection of data on a myriad of topical areas has also assisted in the growth of use. Many of the assumptions and mathematics involved are complex—surely beyond the boundaries of this book.

The following article, "Regulatory Strategies for Workplace Injury Reduction" by Garrett Moran, is an example of the interrupted time-series comparison group design. It also illustrates how an evaluator deals with regression artifact as a threat to internal validity.

Supplementary Reading

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 9.

READING

Regulatory Strategies for Workplace Injury Reduction
A Program Evaluation

Garrett E. Moran
The American University

Methods for the effective regulation of workplace safety have been the subject of continuing controversy. One possible method is the targeting of high injury rate establishments, labeled THIRE. This article evaluates the PAR program of the Mine Safety and Health Administration, which employs a THIRE strategy to regulate safety in stone, sand, and gravel mines. The results suggest that the program may have been effective in reducing the number of injuries in this sector of the mining industry.

AUTHOR'S NOTE: *The author wishes to express his sincere appreciation to Laura Irwin Langbein, Associate Professor, School of Government and Public Administration, The American University, for her support and guidance throughout the preparation of this manuscript. Thanks also to Michael Greene, Associate Professor, Center for Technology and Administration, The American University, for assistance with methodological issues and computer programming; to Lorraine Blank who provided helpful comments and assisted with the editing of the document; and to Mr. Harrison Combs, Jr., of the Mine Safety and Health Administration who supplied the data. No approval of the analytic methods or conclusions should be attributed either to Mr. Combs or to MSHA.*

DURING THE LAST FEW YEARS great attention has been focused on the issue of safety in the workplace. Economists and policy analysts have vigorously debated the merits of government intervention in the "safety marketplace." This debate has focused both on whether such involvement is appropriate and, if so, what the optimal form for intervention

should be. There seems to be widespread agreement that market imperfections exist, yet there is little confidence that a governmental remedy would improve the situation. As Wolf (1979) has so clearly shown, the character of the public sector often makes it unlikely that the legislature will succeed where the marketplace has failed.

Several economic theorists have argued that the social costs incurred as the result of the standards enforcement approach to safety enhancement, the approach almost universally employed in federal programs, far exceed the resulting benefits. A determination of this sort requires consideration of such issues as the valuation of human life and health, which are beyond the scope of this discussion (for example, see Viscusi, 1979, 1980). Other authors, notably Mendeloff (1980), argue that considerations other than that of economic efficiency must be brought to bear. This position admits a loss of efficiency is a geniune possibility,[1] but suggests that society as a whole may be willing to make such a trade-off in order to lessen the probability that identifiable subgroups may be unduly exposed to possible injury (Okun, 1975).

Regardless of which position one endorses on this question, an effective and inexpensive method for reducing injuries is called for. In the evaluation literature, much attention has been given to each aspect of this issue. Smith (1980) and Mendeloff (1980) both conclude that the regulation and inspection efforts of OSHA (Occupational Safety and Health Administration) probably, though not certainly, have a small, but statistically significant effect. Viscusi (1979, 1980) and others come to the opposite conclusion, claiming the injury rates are not changed by OSHA's programs. Differences in methodology and approach can certainly account for the seemingly contradictory results, but perhaps the most significant conclusion is that the determination of the effectiveness of such programs is a difficult and complex matter.

Mendeloff (1980: 103–105) cites data suggesting that workforce demographics alone account for a very substantial proportion of the variance in injury rates.[2] In developing a model of the possible impact of OSHA he points out that an inspection program of the typical sort can only be expected to reduce the probability of injuries that result from violation of an existing safety standard and are detectable in the context of periodic site visits. He concludes that this class of injuries, which he labels injuries caused by detectable violations (ICDV's), represents a rather small proportion of the total. Because OSHA has historically targeted industry groups with high overall injury rates, he suggests that interindustry variation in the percentage of ICDV's may, in an aggregate analysis, obscure any effects the inspection program might have.

> It is rather easy to construct numerical examples that illustrate how such a situation could easily lead to underestimates of the OSHA inspection effect. For example, suppose that the ICDV category contains 10% of all injuries in the nontarget industries, but only 5% in the target industries. Suppose further that OSHA really has been effective, that ICDV injuries are going down by 5% in other industries and—because of the intense inspection effort—going down by 10% in the target industries. However, suppose that non-ICDV injuries are going up by 5% in all industries. Because of the injury mix, the calculations show that the overall target industry rate would go up by 4.25%, while the nontarget rate would go up by only 4.00%. Thus the lack of impact found by this study could be consistent with an actual program impact [Mendeloff, 1980: 94–95].

Economists, in particular, have suggested that a more effective approach than the combination of standards and inspections might

be developed around the idea of an injury tax. Such a system would, they argue, completely avoid the issue of whether a cause of injury was detectable by inspection, might be less expensive to administer, and would more consistently provide economic incentives to proprietors to lessen the possible causes of injuries. The question of whether a tax system could perform in the intended manner will in all likelihood remain an academic one. Mendeloff (1980) cogently argues that the political environment is unlikely to permit passage of legislation that may be described as making it acceptable to harm workers as long as one could pay for the privilege.

In place of the taxation approach, Mendeloff (1980) calls for an approach that he labels THIRE, Targeting High Injury Rate Establishments. Such a program, by concentrating on the demonstrated "bad apples," would be politically appealing and because it would select firms with high injury rates compared to others in the same industry, it would avoid the problem of differential ICDV's that made it difficult to determine program effectiveness. He suggests that the time, penalty, and compliance costs associated with high inspection frequencies would serve as a form of injury tax on these dangerous establishments. The employer would thus have an incentive to reduce perils outside of the ICDV class, just so he could get the inspectors and the costs they impose "off his back."

> The crucial operating principle of the targeting program is that it imposes incremental costs on employers for each additional injury or injury loss. . . . Linking the frequency of inspection to relative safety performance may have an analytic rationale as well. Poorly performing establishments may plausibly carry out injury prevention more cheaply at the margin than high-rate establishments. Under a system based on changes in rates, a high-rate firm would not be inspected at all unless its rate worsened. Under a THIRE system, in contrast, it would be inspected frequently unless its rate improved [Mendeloff, 1980: 134].

Mendeloff admits that such a program might be less effective in practice than it would appear on paper. He acknowledges the possibility of worsened performance in industries with relatively low injury rates,[3] but he hypothesizes that the reductions obtained by what would effectively be a tax on any injury in the high-rate group would outweigh the increases in injuries caused by ICDV's in the lower-rate group. Also, postaccident investigations would continue in the entire population of firms, and the threat of future inclusion in the targeted population would theoretically provide motivation to prevent a deteriorating safety record. Given the possibility of such shifts in performance in the different groups of firms, Mendeloff suggests that THIRE be implemented on a regional basis so that control districts would be available for comparison. As an alternative or supplement, he suggests time-series data be collected on such a program so that it would be possible to separate, at least partially, the impact of the program from that of ongoing trends.

The Par Program

A program that demonstrates many of the features of THIRE has been in operation within the Mine Safety and Health Administration (MSHA) since 1975. This accident reduction program, identified by the acronym PAR, selects for special attention those metal/nonmetal mines which, on the basis of a weighted injury index, are shown to have significantly greater numbers of injuries than other comparable operations. The index numbers that serve as the selection criterion take into account the size of the operation, the rate of lost workday injuries and their seriousness, policy differences regarding the definition of lost

workdays, the difference in rates compared to the national norm, and the upward or downward trend compared to the norm.

The analysis presented below focuses on mines from a single sector of the metal/nonmetal mining field—sand, gravel, and stone mines—and thus can be expected to have comparable proportions of ICDV's. As Table 1 illustrates, the targeting strategy of the PAR program clearly focuses intensive attention on the program mines. The mean numbers of inspections, orders, citations, and inspector hours on site are all from two to ten times higher for the PAR mines than for the general population of sand, gravel, and stone mines (here labeled as "All Mines"). A comparison group of mines that also had recorded poor performance in the area of safety but had not been included in the PAR program (labeled Non-PAR Hi Mines) received even less attention than did the general population of mines.[4] The costs associated with such inspection-related activity should, according to Mendeloff's thesis, provide a substantial economic incentive to reduce dangerous conditions.

TABLE 1
Comparison of Quarterly Group Means

Variable	*All Mines*[1] *N = 3855*	*PAR Mines N = 24*	*Non-PAR Hi N = 286*
No. of miners	30.34	119.20	10.85
Hours worked/ employee	495.87	536.42	439.84
No. of inspections	.80	1.61	.70
No. of orders	.08	.79	.05
No. of citations	2.53	8.92	1.45
No. of hours on site	6.51	22.09	4.12
Nonfatal day-lost injuries	.29	2.55	.25
No. of fatalities[2]	.003	.020	.004
Total number of injuries	.30	2.57	.26
Injury index	2.87	12.26	19.06
Fatality rate[3,4]	.22	.43	1.05
Injury rate[4]	14.55	61.23	95.40

Note: All means differ significantly (Alpha = .05), except for those variables referenced in notes (2) and (3) below.
1. The mines included in the PAR and Non-PAR Hi groups have been excluded from the All Mines group for purposes of the difference of means tests.
2. PAR mines differ significantly from All Mines (Alpha = .01) and from Non-PAR Hi Mines (Alpha = .02), but there is no significant difference between All Mines and Non-PAR Hi Mines.
3. There is a significant difference between All Mines and Non-PAR Hi Mines (Alpha = .03), but there are no significant differences between either All Mines and PAR Mines or Non-PAR Hi Mines and PAR Mines.
4. Fatality and injury rates are defined respectively as the number of fatalities or injuries per million hours worked.

The mines in the program are larger than those excluded from PAR, having nearly four times as many miners as the general population. Examination of this table demonstrates that the program has certainly focused on the "bad apples," as defined in terms of absolute numbers of injuries. The mean number of total injuries (including fatalities and nonfatal, day-lost injuries) is more than eight times as high in the PAR mines, whereas the weighted injury index numbers, used as the selection criterion for the program, are more than three times as high. Thus, even when the injury data is weighted for the size of the operation and other relevant factors by the injury index numbers, the PAR mines are still substantially more hazardous places to work than the average.

Table 1 also shows clearly that there is another group of mines that may be even more dangerous workplaces than the PAR mines. The injury index numbers for the Non-PAR Hi group are, on average, 55% higher than those of the PAR mines and nearly seven times as high as the comparable figures for the general mine population. Although these mines were excluded from the PAR program due to their small size, they are of interest because they provide some basis for determining what patterns may emerge in the injury rate trends over time when no governmental intervention is involved other than the typical standards enforcement approach. This notion will be addressed at greater length below.

It appears that PAR does in fact meet the conditions for defining the THIRE program

that Mendeloff has outlined. Let us turn our attention now to the issue of its effectiveness in reducing injuries.

Methodology

The overriding difficulty in evaluating a program of this sort is that of regression to the mean, a fact clearly recognized by Mendeloff (1980: 143–144) in his discussion of the THIRE approach. When mines are selected for participation based on their extreme scores, it is highly probable that regression artifacts will result in those scores decreasing, regardless of the presence of any intervention. The problem then becomes one of attempting to separate the effects, if any, resulting from the program from those that would have occurred due to statistical artifacts.

Campbell and Stanley (1963) recommend the use of the multiple time-series design as a means of addressing the problem. The difficulty persists, however, unless a sufficiently similar comparison group can be located. Although there is no wholly satisfactory way around this problem, it was determined that an approximate solution may be found by comparing PAR mines to others that had high-injury index numbers, yet were excluded from the program for other reasons. As was apparent in Table 1, the group of mines designated as Non-PAR Hi mines were significantly more injury prone than were the PAR mines. Unfortunately, they also differ on most other variables examined. Notably, these mines are quite small, averaging only about 10 miners compared to 30 in the general population and nearly 120 in the PAR group. The absence of a viable alternative dictates that this group serve as the comparison group. Regression techniques will be used to hold size differences between the PAR and Non-PAR mines constant.

A second difficulty is that mines were inducted into PAR at different points in time and were retained only until the index numbers suggested they were again in line with industry norms. This problem was dealt with by constructing a counter that had an initial value of zero, became one when the mine entered the program, and was incremented by one each quarter thereafter. This counter was intended to represent both the effect of the PAR program intervention and because mines would have been inducted into the program about the time of their peak injury index numbers, the regression artifact as well.[5] For the regression to the mean simulation comparison group (the Non-PAR Hi group) the counter assumed the value one when the mine had its peak injury index number and was incremented identically, thus representing the regression artifact alone.

In order to control for other influences on the mines that may be associated with the passage of time, a separate counter is included in the model. This variable had the initial value of one and was increased by one throughout the 22 quarters for which data were available.

Because reduction of the absolute number of injuries is the more important policy consideration, the dependent variable selected was the total number of injuries per mine per quarter.[6] This necessitated inclusion of an independent variable to measure exposure because, as was discussed above, the PAR mines tended to have more miners who worked longer hours than the Non-PAR Hi mines. The total number of hours worked per mine per quarter was consequently included in the model.

Quarterly data for all sand, gravel, and stone mines were obtained from MSHA for the period from the beginning of 1975 until the middle of 1980. The resulting pooled cross-sectional time-series data set was analyzed by means of weighted least squares (Kmenta, 1971; Berk et al., 1979: 385–410). This allowed for the appropriate corrections for the problems of heteroscedasticity and serial correlation that are likely to occur in data

of this character.[7] From the full data set two subgroups were selected, consisting of the following: (1) all PAR mines; and (2) all mines not included in the PAR program that had, at one or more points during the time series, injury index numbers more than two standard deviations above the mean. The first group was of course the target of the analysis. The second group (the Non-PAR Hi mines) served as the regression to the mean simulation comparison group.

The first question is whether, controlling for the total exposure of miners to the workplace and any long-term trend factors, the number of injuries decreases as the result of exposure to the PAR program. Even if this occurs, it is necessary to determine whether a similar outcome is observed in the comparison group, that is, whether a regression to the mean effect produces a result similar to that found in the PAR mines. If PAR is effective, then any reduction in numbers of injuries seen in the PAR mines must be significantly greater than that found in the comparison mines.

This calls for a test of the research hypothesis that the PAR mines and the Non-PAR regression to the mean simulation comparison group constitute different populations. That is, any downward trend following the peak injury index quarter in the Non-PAR Hi mines must be significantly less steep than any downward trend resulting from the intervention of the PAR program. Inability to reject the null hypothesis that they are from the same population would suggest that any decline in numbers of injuries observed after their inclusion in the program could not confidently be attributed to the effects of the program but may instead be attributable to regression effects.

Results

To test the above hypothesis, observations from the two groups of mines, those in the PAR program and those in the Non-PAR Hi comparison group, were pooled. The dependent variable was the absolute number of injuries per mine per quarter. The independent variables of primary interest were those counters indicative of the trend following induction into the PAR program or, in the case of the comparison group, the trend following the peak injury index number. While the two counters were treated as a single variable, the inclusion of a slope dummy variable that had the value of zero for comparison mines and the value of one if the mine had participated in the PAR program permitted distinctions to be made between the respective slopes. In recognition of the clear differences in the absolute numbers of injuries between the larger PAR mines and the smaller mines of the comparison group, a dummy intercept was included in the model as well.

Two additional independent variables were included. The first was simply the time counter variable, which had the effect of controlling for any ongoing trends over the full time series. The final variable was the number of hours of exposure to the work setting. This last variable served the crucial role of controlling for the size of the mines, a function made particularly important by the clear disparity in size between PAR and comparison mines. The results of the estimated model are presented in Table 2.

The time trend counter indicates that, controlling for all other independent variables, there was an upward trend over the full time series such that the passage of one additional quarter was associated with an average increase of .0258 injuries. All else being equal, each additional ten thousand hours of exposure to the work setting was associated with a small (0.16) but statistically significant increase in the number of injuries. Finally we note that both the PAR/Post Peak counter and the slope dummy-PAR/Post Peak counter were, controlling for other variables, associ-

TABLE 2
Results of the Regression Model

Variable	*Regression Coefficient*	*t Value*	*t Significance*
Intercept	–0.0069	–0.147	0.8832
PAR/Post peak counter	–0.0281	–3.000	0.0027
Dummy intercept	1.1049	15.868	0.0001
Dummy x PAR/Post peak	–0.1540	–9.261	0.0001
Time counter	0.0258	4.280	0.0001
Hours of exposure	0.000016	15.938	0.0001

F Value Full Equation (F Prob.)	*F Value-Test of Equality of Two PAR/Post Counters (F Prob.)*	*N (310 mines x an average of 15.6 quarters)*	*R Square*	*Durbin Watson Statistic*
294.94 (0.0001)	33.15 (0.0001)	4844	.2336	2.101

Note: As noted above, this model includes weighted least squares corrections for the problems of autocorrelation and heteroscedasticity. These corrections make the R-Square statistic a misleading measure of goodness of fit.

ated with decreases in the mean number of injuries per quarter. Inspection of the table reveals that the magnitude of decrease was more than six times as large in the PAR mines than in the comparison mines.

The associated Chow test of the null hypothesis that the two groups constitute a single population resulted in an F-statistic of 33.15 that is statistically significant (Alpha = 0.0001). We are thus able to conclude that the PAR program produced a substantially greater reduction in numbers of injuries than that resulting from the regression artifact alone.[8]

Conclusions

The above discussion pointed to the potential difficulty in drawing definitive conclusions regarding the effectiveness of programs designed to improve workplace safety. The current effort provides an additional illustration of the point. On the positive side, the (negative) partial slope coefficient associated with participation in the PAR program was six times as large as the comparable counter for the Non-PAR Hi group. Although regression artifact in the Non-PAR Hi mines led to a change when other variables are controlled, of –0.0281 injuries per quarter from the peak injury index number period, the comparable figure for PAR mines was –0.1540. If we assume that the PAR and comparison groups are otherwise comparable, we must conclude that the program was highly effective.

As examination of Table 1 indicated, the PAR and Non-PAR mines are quite different from each other. Reexamination of that table shows that PAR mines have twelve times as many miners as the Non-PAR Hi group, and that those miners work longer hours. Such differences between the two groups of mines raise questions about what otherwise appears as clear evidence of the effectiveness of the PAR program. The inclusion of hours of exposure to the work setting should, however, provide an effective statistical control for the size differences between the PAR and Non-PAR Hi mines. Moreover, given the post hoc character of this evaluation and the limitations of the available data base, it is difficult to conceive of an alternate methodology that could more effectively control for the obvious threat to validity.

The importance of this study may ultimately lie less in its findings than in what it attempted. The literature is full of efforts to answer such questions as whether safety regulations taken generically have been effective, or whether economic theory would suggest that an approach that could never be implemented given the political realities may be superior to what is currently in place. Although such issues may present academically interesting questions, they seem to offer little hope of making a genuine contribution to the improvement of safety enhancement methods.

By focusing instead on the evaluation of an approach that is both potentially effective

and politically viable, one may hope to obtain knowledge that could be of greater practical value. It is unlikely that Congress will suddenly change its mind about an injury taxation scheme, but, should it be shown to be effective, an alternate approach to targeting of inspection effort could be readily adopted, perhaps without even the need for congressional approval. Conversely, a determination that such approaches, whatever their conceptual merit, are ineffective in application could result in theorists turning their attention to alternate and perhaps more fruitful methods.

The PAR program, with its targeting of high injury rate establishments, must be considered as an example of one such potentially promising approach. The results of this evaluation are indicative that the program may have been quite effective, although the lack of comparability between comparison and treatment groups on important dimensions calls for some caution in attempting to reach definitive conclusions. Given the promising yet indefinite findings of the current study, further research on this program is clearly warranted.

Notes

1. It should be pointed out that the current work attempts only to examine the PAR program as an example of Mendeloff's (1980) targeting of a high injury rate establishment (THIRE) strategy. Attention is given only to whether the program is effective. No consideration is given to the relative effectiveness of such alternate approaches as workmen's compensation or conventional injury taxation proposals, nor are efficiency considerations brought to bear.
2. See Mendeloff (1980:103–105). His model using only demographic factors and pay rate data explained 83% of the variation in annual injury rate changes.
3. Such a decline in performance in the previously better performing industries would be predicted both by standard deterrence theory and by such formal models as that developed by Langbein and Kerwin (1984). Langbein and Kerwin, paying particular attention to the effects of time delays and negotiated compliance, suggest that any simultaneous relaxation in agency policy on these factors may both exacerbate the problems of lessened compliance in the nontargeted firms and also serve to induce noncompliance in the targeted firms. For the sake of simplicity, it will be assumed here that no relaxation in these policies accompanied the development of the THIRE targeting strategy. It may also be noted that the evaluation strategy proposed by Mendeloff facilitates evaluation of such effects, though the current effort admittedly does not.
4. This comparison group, the Non-PAR Hi Mines, will be discussed in greater detail below. They were selected from the general population of sand, stone, and gravel mines exclusively on the basis of the injury index numbers. Mines were included in the group if they had, at any point during the time series, an injury index number that was more than two standard deviations above the mean for all mines.
5. As a test of the assumption that induction into the PAR program would roughly coincide with the peak injury index number, a separate model was developed using the same counter assignment strategy for PAR and Non-PAR Hi groups. The results of this alternate model proved to be essentially identical to those obtained when the PAR mine counter was started upon entry into the program.
6. This included both fatal injuries and nonfatal, day-lost injuries. The same models were run using nonfatal day-lost injuries alone as the dependent variable, producing essentially identical results. Models were attempted using the number of fatalities alone as the dependent variable, but the results proved insignificant.
7. Autocorrelation was corrected by setting each variable equal to the value of that variable minus the product of rho and the one-period lag value of the variable. Heteroscedasticity was corrected by weighting the observations by the number of hours of exposure to the work setting through use of the SAS (1982) Proc Reg WEIGHT statement.
8. It should be mentioned that another model was developed that included an independent variable summing the various types of inspection and enforcement activities. The model was run for PAR Mines only with the same dependent variable. This produced a statistically significant partial slope coefficient, but the sign was positive, suggesting that inspection and enforcement activities were associated with increased numbers of injuries. The result is no doubt attributable to postaccident investigations occurring in the same quarter, yet experimentation with lagged enforcement activity failed to indicate any significant relationships, though the sign did become negative as was predicted.

References

Ashford, N. (1976) *Crisis in the Workplace: Occupational Disease and Injury.* Cambridge, MA: MIT Press.

Berk, R. A., D. M. Hoffman, J. E. Maki, D. RaumaA, and H. Wong (1979) "Estimation procedures for pooled cross-sectional and time series data." *Evaluation Q.* 3 (August): 385–410.

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

Cook, T. J. and F. P. Scioli, Jr. (1975) "Impact analysis in public policy research," pp. 95–117 in K. M. Dolbear (ed.) *Public Policy Evaluation.* Beverly Hills, CA: Sage.

Gordon, J. B., A. Akman, and M. Brooks (1971) I*ndustrial Safety Statistics: A Re-examination.* New York: Praeger.

Kmenta, J. (1971) *Elements of Econometrics.* New York: Macmillan.

Lagather, R. B. (1979) "The federal mine safety and health act." *Labor Law J.* 30 (December): 723–731.

Langbein, L. and C. M. Kerwin (1984) "Implementation, negotiation and compliance in environmental and safety regulation." Unpublished manuscript.

Lewis-Beck, M. S. and J. R. Alford (1980) "Can government regulate safety? The coal mine example." *Amer. Pol. Sci. Rev.* 74 (September): 745–756.

McDowall, D., R. McCleary, E. E. Meidinger, and R. A. Hay, Jr. (1980) *Interrupted Time Series Analysis.* Beverly Hills, CA: Sage.

Mendeloff, J. (1980) *Regulating Safety: An Economic and Political Analysis of Occupational Safety and Health Policy.* Cambridge, MA: MIT Press.

Nichols, A. L. and R. Zeckhauser (1977) "Government comes to the workplace: an assessment of OSHA." *The Public Interest* 49: 36–69.

Okun, A. M. (1975) *Equality and Efficiency: The Big Tradeoff.* Washington, DC: The Brookings Institution.

Sands, P. (1968) "How effective is safety legislation?" *J. of Law and Economics* II (April): 165–174.

SAS Institute Inc. (1982) *SAS User's Guide: Statistics*, 1982 Edition. Cary, NC: SAS Institute.

Smith, R. S. (1980) "The impact of OSHA inspections on manufacturing injury rates." *Evaluation Studies R. Annual* 5: 575–602.

Stokey, E. and R. Zeckhauser (1978) *A Primer for Policy Analysis.* New York: Norton.

Viscusi, W. K. (1980) "Labor market valuations of life and limb: empirical evidence and policy implications." *Evaluation Studies R. Annual* 5: 545–574.

____(1979) "The impact of occupational safety and health regulation." *Bell J. of Economics* 10 (Spring): 117–140.

Wolf, C., Jr. (1979) "A theory of nonmarket failure: framework for implementation analysis." *J. of Law and Economics* 22 (April): 107–139.

Zeckhauser, R. (1975) "Procedures for valuing lives." *Public Policy* 23: 419–464.

Garrett E. Moran is a clinical psychologist and a doctoral candidate in Public Administration and Policy Analysis in the School of Government and Public Administration of The American University. His research interests are in the areas of workplace safety and community support systems for the chronically mentally ill.

Explanation and Critique

Garrett Moran meets his problem with regression artifact straightforwardly using a pooled interrupted time-series design with a comparison group. In this analysis, his goal is to disentangle those effects from the efforts of the Mine Safety and Health Administration. As a clinical psychologist interested in workplace safety, Moran is concerned whether intensified investigations aimed at the worst offenders decreases injuries. What is interesting for us in this chapter is the sensible way Moran tests for regression artifacts.

The accident reduction program, PAR, selected for special attention those mines with significantly greater numbers of injuries than comparable mines. The mines were chosen for treatment on the basis of their high injury rates, controlling for size and other rel-

evant characteristics. Donald Campbell and Julian Stanley (1963, 11) recommend the use of time-series designs with comparison groups to address the problem of regression artifact; Moran, therefore, searched for a group of untreated mines to use as a comparison group. Because his evaluation was conducted after the fact, Moran did not have the luxury of randomly assigning mines to treated and untreated groups. He found a set of mines that also had high injury rates but that admittedly were different from the treated group on a number of important variables. This group Campbell and Stanley refer to as a nonequivalent group—one chosen for a match on extreme scores but lacking a match on other relevant variables. Moran uses regression techniques to control for these differences. The most notable difference is size, which is measured by the total number of hours worked in the mine during the quarter.

Using a pooled time-series design, Moran got the results in table 2 of the article. The treatment for the experimental group is the dummy variable representing the quarter the mine went on the PAR program, while the treatment for the comparison group is the dummy variable indicating the highest quarter's accident rate. Moran literally interprets the regression coefficients (*b*s) as shown in chapter 3 by associating the passage of time (b = 0.0258) and hours of exposure in the work setting (b = 0.000016) with increases in mean number of injuries per quarter, while PAR participation (b = –0.1540) and the regression artifact (b = –0.0281) accounted for decreases in average injuries. By means of a significance test, he finds that the PAR program had an impact above and beyond that attributed to the regression artifact.

The most innovative aspect of this article is the modeling of the regression artifact. Mines were included in the PAR intervention when their injury and accident rates soared above comparable mines. Presumably, mines entered the program when they were at their worst. The program of inspection and intervention, if effectively operational, should act to reduce these rates at a pace above and beyond that anticipated by the regression effect. Therefore, Moran constructed a counter variable for each treated mine with an initial value of zero when it entered PAR. The counter was incremented by one for each additional quarter in the program. The comparable counter for the comparison group was initiated when each comparable mine reached its peak injury rate, with increments being added each subsequent quarter. For a treatment effect to be present, the improvement for treated mines must have been statistically significantly better than that for the untreated mines, which would be improving via regression to the mean. Note that Moran also controls for history with a similar counter and a dummy variable to account for PAR participation.

The policy implications of these findings are obvious: PAR decreases injuries. With this information, policymakers can decide whether this impact justifies further funding or whether a cost-benefit or cost-effectiveness analysis (see chapters 13 and 14) should be conducted in tandem with the present analysis. This second step, however, could not be taken without establishing independent impact. Moran provides the evaluator with a prescription for coping with regression artifacts when programs are targeted to extreme cases.

Moran could have tested his hypotheses another way. In fact, the logic of this approach is similar to the regression-discontinuity analysis covered in chapter 2. The cut point would still be the quarter the mine entered PAR (experimental group) and the quarter with the highest number of injuries (comparison group). Rather than pooling the data from both sets of mines and running one regression, however, four separate regres-

sions could be estimated. The first regression would include the pre-PAR time-series of the experimental group while the second would include the pre-post peak time series of the comparison group. We suspect that if the regression lines were drawn graphically, the regression lines of the experimental and comparison groups would be similar and approach the cut point at similar levels. Then two regressions would be estimated similarly with the post-PAR data and the post-post peak data. An examination of the intercepts of these data should reveal that both intercepts fall below the point of the line predicted by the first two regressions (regression artifact) and that the intercept of the experimental group falls much farther (statistically significant distance) than that of the comparison group (program impact). Either estimation procedure would yield the same results and interpretations.

References

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Boston: Houghton Mifflin.

Langbein, Laura Irvin. 1980. *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation.* Glenview, Ill.: Scott, Foresman.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey.

CHAPTER 10

Posttest-Only Comparison Group Design

The posttest-only comparison group design is similar to the posttest-only control group design. The only difference between the two designs is that the latter uses a randomly assigned control group for comparison and the former uses a nonrandomly assigned comparison group. The posttest-only comparison group design is shown in table 10.1. In this design neither group is pretested, and both groups are posttested after the experimental group receives the treatment.

This form of evaluation is used primarily when an evaluation is proposed after the fact and is one of the most commonly used designs. For instance, if a city is experiencing fiscal constraints, the council may ask staff or a consultant to prepare evaluations of existing programs and provide the results before budget deliberations the following month. Within these time constraints (and certainly budget constraints), full-fledged experimental designs are out of the question. However, the decision makers will need to make budgetary decisions on the basis of something more than hearsay. Consequently, the posttest-only comparison group design is appropriate.

TABLE 10.1
The Posttest-Only Comparison Group Design

	Pretest	Program	Posttest
Group E		X	O
Group C			O

One of the weaknesses of the design is that the data are collected at only one point. The lack of time-series data do not allow the researcher to tell if the observation is really "true" or if it is a result of random fluctuations. In addition, the evaluator needs to be sure of the form the outcome should take and collect the data to reflect it. In other words, if the treatment is supposed to have an immediate effect, then the data should be collected as soon after the treatment as possible. If the impact is incremental, however, a longer time period between treatment and measurement is necessary. The consequences of poor timing are critical; the problems of committing Type I and Type II errors should be apparent. In the example of a city that is facing fiscal problems, the future of programs is at stake. The classic example is the results of the early evaluations of the Head Start programs, which showed little short-term impact. On the basis of those data, it was proposed that

the program be terminated. However, subsequent tracking of Head Start graduates demonstrated a substantial long-term impact. Fortunately for the children who entered the program later, the directives of the early results were not followed.

Choosing a comparison group is also a major consideration. Because the posttest-only comparison group design is ex post facto, random assignment is impossible. If a good comparison group can be identified, however, valid comparison across the groups is possible. Members of the comparison group should be similar to the treated group on as many relevant variables as possible save the treatment. To enhance internal validity, the measurement of the characteristics of the treated and comparison groups should be taken at the same time.

In the following article, Martha Bronson, Donald Pierson, and Terrence Tivnan used the posttest-only comparison group design to estimate the long-term impact of the Brookline Early Education Program (BEEP) on children's in-classroom behavior.

Supplementary Reading

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 4.

READING

The Effects of Early Education of Children's Competence in Elementary School

Martha B. Bronson Donald E. Pierson Terrence Tivnan
Brookline Early Education Project

Programs of early education have focused primarily on low-income populations and on outcomes available from traditional assessments of children's intelligence or achievement. This evaluation used observations of children's behavior in elementary school classrooms several years after the program services had been delivered. Children from a wide range of family backgrounds were included. The results indicated that program participants benefited particularly in the area of mastery skills or academic learning behaviors. Children with highly educated mothers showed advantages regardless of program service level, but children whose mothers were less highly educated showed advantages only with relatively intensive service level.

AUTHORS' NOTE: This article is based in part on a presentation prepared for the Biennial Meeting of the Society for Research in Child Development, Detroit, Michigan, April, 1983. The work was made possible by funds granted by Carnegie Corporation of New York and The Robert Wood Johnson Foundation. The statements made and views expressed are solely the responsibility of the authors.

OVER THE PAST TWENTY YEARS there have been a large number of early education programs designed to provide services to preschool children and their families. These programs have provided major challenges for evaluators because of the diversity of program goals, settings, and types of families for whom these programs are intended. Partly as a result of this diversity it has been difficult to come to any conclusions about the overall ef-

ficacy of such early interventions. It has only been relatively recently that any consistent evidence of long-term effects has emerged. The Consortium for Longitudinal Studies (Lazar and Darlington, 1982) has documented that early education programs for children from low-income families have important and lasting effects. The effects—evident several years after conclusion of the intervention—include fewer assignments to special education services and fewer retentions in grade. At the present time, however, neither the Consortium data nor other available studies (e.g., Gray and Wandersman, 1980; Schweinhardt and Weikart, 1980) have determined adequately whether the effects of early education can be identified in elementary school classroom behaviors and, if so, whether such effects extend to children other than those from low-income backgrounds.

In developing measures to evaluate early education programs, many psychologists and educators have criticized the traditional reliance on intelligence tests (Zigler and Butterfield, 1968; McClelland, 1973; Seitz et al., 1975; Zigler et al., 1973). Others have stressed the importance of focusing on social or functional competence, including cognitive, social, and motivational components (Anderson and Messick, 1974; Zigler and Tricket, 1978; Gotts, 1979; Zigler and Seitz, 1980; Scarr, 1981). Despite these persuasive appeals to measure how children function within the classroom domain, the prekindergarten and elementary education evaluation literature has shown little response. The difficulties encountered in developing and implementing innovative methods for assessing effectiveness have limited most evaluations to more traditional measures.

In only very few instances have early education program evaluations been carried out on heterogeneous groups of participants. Fewer still have focused on children's behaviors in classrooms. Pierson et al. (1983) found significant behavioral advantages during fall and spring of the kindergarten year for a heterogeneous group of early education participants. The outcome measures were designed to assess the child's functional competence in social and learning situations in the classroom setting. These measures corresponded closely to the goals of the intervention and to the notions of competence held by the school system in which the children were enrolled. The study reported in this article extends the earlier work by following up on children when they were enrolled in second grade, using the same classroom observation instrument. Children who had been enrolled in an early education program were observed in their second grade classrooms. The observations focused on assessing their functional or educational competence.

Thus this study provides an approach to evaluation of early education programs that addresses several of the limitations of the available literature. First, the participants come from a wide range of family backgrounds, not just low-income families. Second, this study represents an opportunity for follow-up evaluation of a program that ended at the children's entry into kindergarten. Finally, the evaluation involves the use of classroom observations, focusing on behaviors considered important by the project as well as the local school system.

Method

Program Description

The Brookline Early Education Project (BEEP) involved a combination of services: parent education, periodic monitoring of the children's health and development, and early childhood programs. The parent education services were administered at three levels to assess the cost effectiveness of different service packages. Families were assigned at random to one of three cost levels: Level A included home visits and meetings at least once per month as well as unlimited contacts with staff and some

child care in the project Center; Level B offered home visits and meetings about once every six weeks as well as the Center options; Level C was restricted primarily to services available at the Center. BEEP staff were available to families at the Center but no home visits, meetings or child care services were scheduled. In working with parents, teachers followed curriculum goals that were oriented toward supporting the parents in the role of primary educators of their child.

The monitoring of health and development was intended to prevent or alleviate conditions that might interfere with learning during the first five years of life. A team of psychologists, pediatricians, and a nurse conducted the examinations of hearing, vision, neurologic status, and language, perceptual, and motor skills. Follow-up conferences and advocacy work were pursued by teachers and social workers as needed. These diagnostic services were available equally to all participants in the project.

The early childhood programs involved a weekly play group at age two years and a daily prekindergarten program at ages three and four years. These were held in the Center or at elementary schools and were focused on individually tailored educational goals (Hauser-Cram and Pierson, 1981). As with the exams, the children's educational programs were available to all families regardless of the parent education cost-level assignment.

The project's primary goal was to reduce or eliminate school problems. For this reason, the evaluation focuses on whether the program decreased the proportion of children who fell below certain minimal competencies defined as necessary for effective functioning in second grade.

Subjects

The experimental subjects were 169 second grade children who participated from birth to kindergarten in BEEP. Enrollment was open to all residents of Brookline and to ethnic minority families from the adjacent city of Boston. Participants were recruited through the schools, health, and social service agencies, and neighborhood networks. Recruitment strategies were designed to locate and inform families who were expecting to have babies in the coming months and who ordinarily would not seek or hear about such a program. All participants in BEEP had the option of attending the public schools of Brookline either as town residents or through METCO, a state-funded desegregation program.

At second grade, 104 of the participants were enrolled in the Brookline public schools; another 65 children had moved to other schools, but were within commuting distance of Brookline and available to the study.

A comparison group of 169 children was selected from the same classrooms as the BEEP children. Each comparison child was matched with a BEEP participant by selecting a child at random from within the same classroom and sex group as the participant group child. The children were spread out over 82 classrooms in over 50 different elementary schools. Table 1 shows that demographic characteristics of the two groups are quite similar and that the children involved in this study represent a heterogeneous collection of families. For example, the education levels of the families range from those with less than a high school education to those with college and postcollege degrees. Over 10% of the families spoke first languages other than English.

Measure

The outcome measures in this study are taken from a classroom observation instrument. The instrument was developed in response to the well-documented evidence that a dearth of adequate measures had hindered previous efforts to evaluate Head Start and other early education programs (Smith and Bissell, 1970;

Walker et al., 1973; Butler, 1974; Raizen and Bobrow, 1974). Trained observers recorded the frequency of specific classroom behaviors. The instrument thus avoided the traditional reliance of educational evaluations on tests individually administered outside the classroom milieu. The instrument also focused on specific behaviors that reflected the project's position on what constitutes competent functioning in young school-age children.

The observation instrument, entitled the Executive Skill Profile (Bronson, 1975, 1978, 1981), provided a way of recording a child's performance in planning and organizing work, interacting with others, and carrying out social interactions and learning tasks. The concept of "executive skill" is applied to both social and learning activities in the sense of using effective strategies for choosing and reaching goals.

TABLE 1

Distribution of Background Characteristics for BEEP and Comparison Group

		Percentage Number of Children	
Characteristics		*BEEP (N = 169)*	*Comparison (N = 169)*
Mother's education	College graduate	59	54
	High school graduate	33	37
	Not high school graduate	8	9
Father's education	College graduate	63	65
	High school graduate	29	29
	Not high school graduate	8	6
Number of parents in home	Two	83	81
	One	17	19
First language	English	87	82
	Spanish	9	4
	Chinese	4	2
	Russian	0	4
	Hebrew	0	2
	Other	0	6
Birth order	First	44	50
	Second	38	27
	Third	11	12
	Fourth or later	7	11
Gender	Female	48	48
	Male	52	52

The trained observers followed and recorded the behavior of each child for six 10-minute periods in the spring of the second grade year. The modified time-sampling procedure required that three of the observations begin at the start of a social interaction and three at the start of academic work. The observers were instructed to make certain that at least one of the academic tasks be a language arts (reading/writing) task and at least one be a math activity. Each 10-minute observation was scheduled for a different day, spread out over no less than three and no more than six weeks. Sometimes children's absences from school or scheduling difficulties extended or compressed the observation period.

Behaviors were recorded on a sheet that allowed both the frequency and duration of specific behaviors to be collected. The duration of behaviors was measured either with a small timing device that clicked at 15-second intervals into an earphone worn by the observer (Leifer and Leifer, 1971) or with a stop watch (some observers found this easier). During a two-week training period observers studied a manual (Bronson, 1983) and were trained in classrooms to record each behavior variable to a criterion of 90% interobserver agreement with the observation supervisor. Observers were not informed about the specific purposes of the study and safeguards were employed to avoid divulging children's membership in the participant versus comparison group.

The observations covered eleven variables divided into three categories of behavior: mastery skills, social skills, and use of time. The mastery skills category included three variables designed to measure a child's success in planning and carrying out school learning tasks: resistance to distraction, use of appropriate task attack strategies, and successful completion of tasks.

The social skills category included four variables pertaining to interpersonal rela-

tions: cooperative interaction with peers, successfully influencing others, use of effective cooperative strategies and use of language rather than physical force to persuade and gain the attention of others.

The use of time category consisted of four variables related to a child's degree of involvement in activities within the classroom: proportion of time spent in social activities, proportion of time in mastery tasks (not necessarily mutually exclusive with social time), proportion of time spent without any focused activity (mutually exclusive with both preceeding variables) and rate of social acts. Each observation variable yielded a rate or percent score based on a full set of six 10-minute observations.

Taken as a whole, the observation procedures used in this study provide several important advantages for this evaluation. Information is collected on a wide range of important behaviors. Observation is directed at the behavior of individual children rather than on groups or classrooms or teacher behavior. The focus is on in-classroom behavior, rather than individualized, out-of-classroom assessment. In addition, the Executive Skill Profile has been successfully used in other settings (Bronson, 1975, 1978, 1981; Pierson et al., 1983), and it attempts to measure behaviors that are consistent with the goals of the local school system that is the setting for the evaluation. Thus it addresses some of the important limitations of more traditional evaluation instruments used in elementary schools.

A major goal of the Brookline Early Education Project was to prevent school problems—or, conversely, to increase the proportion of children attaining minimal competencies. Thus it was considered important to analyze the data by looking at the proportions of children who were having difficulty in school as well as by looking at average performance. Criteria for determining adequate performance versus "concerns" or "problems" were derived from clinical impressions of effective behavior and from analyses of data on non-BEEP children. The pivotal point for determining whether a given score should be regarded as "adequate" or "concern" always fell at least one standard deviation below the group mean score for that category. Concerns were considered to indicate a "problem" if scores for two or more variables in a given area (i.e., mastery skills, social skills, or use of time) fell below the criteria for "concern." The results of the comparison of mean scores and the comparisons of percentages of children showing problems will be presented separately.

Results

Table 2 shows the differences between the category means of the BEEP participants and the randomly selected comparison group. The BEEP participants show significant advantages over the comparison group in both mastery and social skills, with the strongest effects in the mastery skills area. There is no difference between the two groups in the use of time area.

Another way of viewing the data is by focusing on the proportions who have problems. Table 3 shows the percentage of children with "concerns" (low scores in categories) or "problems" (two or more low scores in an area) in the BEEP and comparison groups. Again the BEEP group shows a significant advantage over the comparison group in both mastery and social skills with the strongest effect in the mastery skills area. The BEEP group has fewer overall problems and many fewer children with problems in more than one area.

The reduction in severity of difficulties for BEEP participants can also be seen in the numbers of children with low scores across several of the eleven categories. Figure 1 presents the frequency distributions, showing

TABLE 2

Difference Between the Means of BEEP and Comparison Children in Second Grade Spring Classroom Observations

	BEEP (N = 169)		*Comparison (N = 169)*		
	Mean	*SD*	*Mean*	*SD*	*Significance*[a]
Mastery Skills Area					
Percentage tasks completed successfully	84	18	76	23	<.001
Rate task attack strategies	2.13	0.86	1.82	0.83	<.001
Percentage time attending not distracted	94	6	91	12	<.001
Social Skills Area					
Percentage time in cooperative interaction	79	20	77	23	ns
Rate cooperative strategies	0.50	0.41	0.41	0.36	<.01
Percentage success in influencing others	97	4	96	5	<.05
Percentage use of language to influence	97	5	95	6	ns
Use of Time Area					
Percentage time in mastery tasks	54	13	53	13	ns
Percentage time in social activities	51	12	51	12	ns
Rate of social acts	5.34	1.29	5.34	1.32	ns
Percentage time involved	98	2	98	3	ns

a. Significance is based on *t*-tests for matched pairs.

more children with multiple concerns in the comparison group than in BEEP.

Figure 2 shows the numbers of children in the BEEP and random comparison groups who met the criteria for competence in the observations (by having no problems in any of the three areas assessed) with distinctions for mother's education and program cost level. BEEP children with highly educated mothers (college graduates) show advantages over their counterparts. For these subgroups, even the minimal cost-level program resulted in significant ($p < .01$) advantages. However, for children whose mothers were not so highly educated (less than college graduates) a more substantial investment and outreach, represented by cost level A, was required to attain significant ($p < .01$) advantages over the comparison group.

Focusing on the cost levels within BEEP, we find that no significant overall differences across the three groups emerge. The only trend is for the Level A participants to be

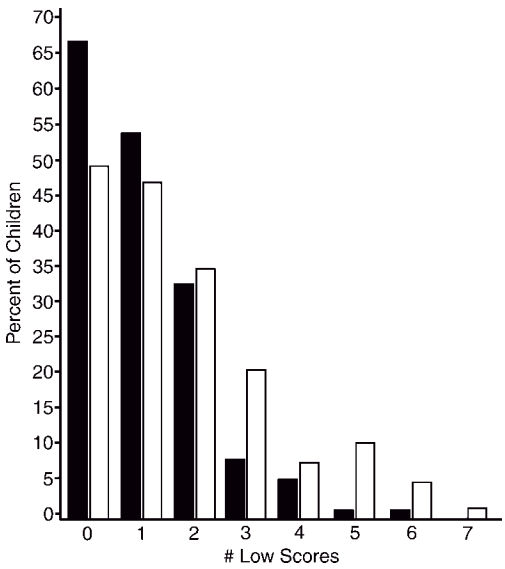


FIGURE 1

Frequency Distribution of the Number of Low Scores Obtained by BEEP (solid bars) and Comparison (clear bars) Children on Eleven Observation Variables

TABLE 3

Percentage Number of Children with Concerns[a] or Problems[b] for BEEP Participants and Randomly Selected Comparison Group in Second Grade Classroom Observations

	Percentage Number of Children Below Competence Criteria		
	BEEP (N = 169)	*Comparison (N = 169)*	*Significance[c]*
Mastery Skills Concerns			
Tasks not completed successfully	14	27	<.01
Inadequate rate of task attack strategies	18	32	.001
Time distracted	18	29	<.05
Social Skills Concerns			
Inadequate time in cooperative interaction	6	8	ns
Inadequate rate of cooperative strategies	18	21	ns
Unsuccessful in influencing others	2	7	<.05
Ineffective use of language to influence	3	7	ns
Use of Time Concerns			
Inadequate time in mastery tasks	11	11	ns
Inadequate time in social activities	17	18	ns
Inadequate rate of social acts	2	2	ns
Time not involved	1	2	ns
Problem in Mastery Skills	12	25	<.01
Problem in Social Skills	2	8	<.05
Problem in Use of Time	2	1	ns
Overall Difficulty: Problems in One or More Areas	14	29	<.01

a. A "concern" is a score below the established criterion in any single category.
b. "Problems" are two or more scores below the criteria in an area.
c. McNemar's matched-pairs test.

ahead of Levels B and C among the less educated families. These results are consistent with other analyses of the within-program differences.

Conclusions and Discussion

Children who participated in the Brookline Early Education Project showed several advantages in second-grade classroom behavior indices over comparison children. Advantages for the BEEP participants were apparent both in mean differences in the behavior categories and in the relative numbers of children performing below competence criteria. The BEEP advantage was most pronounced in the mastery, or academic learning area. BEEP children with highly educated mothers showed significant advantages over comparison children regardless of program level, but children whose mothers were less highly educated showed significant advantages only with the relatively intensive level A program.

Observations of BEEP and comparison children in kindergarten (Pierson et al., 1983) with the same observation instrument showed significant advantages for the program group with the greatest differences appearing in the social and use of time areas. In the second grade observations the BEEP advantage is greatest in the mastery area. This shift in the pattern of BEEP advantages over

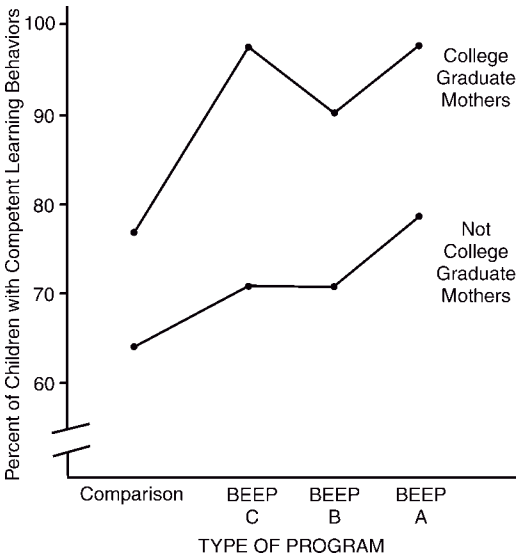


FIGURE 2
Overall Competencies Observed

NOTE: Percentage of children with no difficulties in mastery skills, social skills, or use of time behaviors, analyzed by level of mother's education and type of program, is shown.

comparison children is interesting. In the fall of kindergarten year classroom behavior differences favoring BEEP children were strongest in the social and use of time areas. In the spring of kindergarten year there was a shift in the pattern of BEEP advantages, with use of time behaviors being less important, mastery behaviors becoming more important, and social behaviors continuing to be strong discriminators between the two groups. By second grade, mastery behaviors are the strongest discriminators. The social behavior categories show a consistent but less strong BEEP advantage, and the use of time categories reveal no differences between the BEEP and comparison groups.

This shift in the pattern of the relative advantages of BEEP over comparison children seems to be related to the changing patterns of classroom demands at these three time periods. In the fall of kindergarten year, academic demands are few and the emphasis in classroom is on school adjustment and learning school routines. The behaviors in the social and use of time areas are those most likely to pick up differences in school-related competence under these circumstances. In the spring of kindergarten year some academic demands are being introduced—numbers, letters, printing, and so on—and most children have adapted to school routines. This shift in emphasis in the classroom away from routines and toward academic demands is reflected in the changing pattern of advantages of BEEP children, away from use of time categories and toward differences in mastery categories. By the second grade, the primary demands in the classroom are academic. There is less time for social interaction and less room for differences in use of time categories since children's involvement in various activities is much more controlled. The pattern of BEEP advantages reflects these shifts in classroom demands and constraints.

An additional finding that deserves some attention was the lack of strong main-effect differences across the three program service levels. For the highly educated families the least intensive level of service was as effective as the most intensive level. For the less highly educated, there were no differences between the moderate and lowest levels of service, and only the most intensive level showed a significant advantage over the comparison group. Of the three major service components that were part of BEEP—parent education, diagnostic monitoring of children's status, early childhood programs for children—only the parent education services were offered differentially; diagnostic and education programs for children were equally available to all. So it should not be surprising that in this study, focusing on outcomes for children some years after the provision of services, few differences among the service levels should be observed. Nevertheless, the search for interaction ef-

fects, or differential impact on different types of families, is important in evaluating and planning early education programs. The evidence here, although limited, suggests that service levels need to be more intensive when parents are less educated. More educated parents appear to benefit from even minimal services. This finding indicates the potential usefulness of early education for all children, not only the less educated or economically needy groups. The benefits may well vary across different types of families, and the types of services that will be helpful may also vary considerably. This issue deserves more attention from evaluators.

From the evaluation perspective, the results suggest the value of classroom observations as an evaluation technique and the importance of tailoring outcome measure to intervention goals. It is noteworthy, in this regard, that there were few differences between BEEP and comparison groups on a traditional measure of IQ or ability obtained at entry into kindergarten (see Pierson et al., 1983). So without the use of observations in the classrooms the impact of the early intervention would be very difficult to detect, and criticisms similar to those cited earlier concerning reliance on traditional tests (e.g., Zigler and Seitz, 1980; Scarr, 1981) would be relevant.

The reduction in classroom behavior problems also has significant practical implications for elementary schools. Even if behavior problems are not so severe as to require expensive special services, behavior problems in the classroom require teacher attention and reduce the amount of productive teacher time and energy available to all children. Fewer behavior problems in a classroom result in a more positive classroom atmosphere and more "learning time" for all children.

In summary, the results of this study suggest the importance of a carefully planned program like the Brookline Early Education Project for all school systems. Education and support services to parents of young children coupled with early education programs for the children should be recognized as an essential part of a high quality elementary school curriculum. Early detection and prevention of learning difficulties is more effective, and less expensive in the long run, than remediation.

References

Anderson, S. and S. Messick (1974) "Social competence in young children." *Developmental Psychology* 10: 282–293.

Bronson, M. B. (1975) "Executive competence in preschool children." Presented to the annual meeting of the American Educational Research Association, Washington, D.C., April. ERIC Document Reproduction Service ED 107 378.

Bronson, M. B. (1978) "The development and pilot testing of an observational measure of school-related social and mastery skills for preschool and kindergarten children." Doctoral dissertation, Harvard Graduate School of Education.

Bronson, M. B. (1981) "Naturalistic observation as a method of assessing problems at entry to school." Presented to the annual meeting of the Society for Research in Child Development, Boston, April.

Butler, J. A. (1974) *Toward a New Cognitive Effects Battery for Project Head Start.* Santa Monica, CA: Rand Corp.

Gotts, E. E. (1979) "Early childhood assessment." In D. A. Sabatino and T. L. Miller (Eds.) *Describing Learner Characteristics of Handicapped Children and Youth.* New York: Grune & Stratton.

Gray, S. W. & L. P. Wandersman (1980) "The methodology of home-based intervention studies: Problems and promising strategies." *Child Development* 51: 993–1009.

Hauser-Cram, P. and D. E. Pierson (1981) The BEEP Prekindergarten Curriculum: A Working Paper. Brookline, MA: Brookline Early Education Project.

Lazar, I. and R. Darlington (1982) "Lasting effects of early education: A report from the Consortium for Longitudinal Studies." Monographs of the Society for Research in Child Development 47 (2–3, Whole No. 195).

Leifer, A. D. and L. J. Leifer (1971) "An auditory prompting device for behavior observation." *J. of Experimental Child Psychology* 2: 376–378.

McClelland, D. C. (1973) "Testing for competence rather than for 'intelligence'." *Amer. Psychologist* 28: 1–14.

Pierson, D. E., M. B. Bronson, E. Dromey, J. P. Swartz, T. Tivnan, and D. K. Walker (1983) "The impact of early education as measured by classroom observations and teacher ratings of children in kindergarten." *Evaluation Rev.* 7: 191–216.

Raizen, S. and S. B. Borrow (1974) *Design for a National Evaluation of Social Competence in Head Start Children.* Santa Monica, CA: Rand Corp.

Scarr, S. (1981) "Testing for children: Assessment and the many determinants of intellectual competence." *Amer. Psychologist* 36:10 1159–1166.

Schweinhart, L. and D. P. Weikart (1980) *Young Children Grow Up.* Ypsilanti: Monographs of the High/Scope Educational Research Foundation Seven.

Seitz, V., W. D. Abelson, E. Levine, and E. F. Zigler (1975) "Effects of place of testing on the Peabody Picture Vocabulary Test scores of disadvantaged Head Start and non-Head Start children." *Child Development* 45: 481–486.

Smith, M. S. and J. S. Bissell (1970) "Report analysis: the impact of Head Start." *Harvard Educ. Rev.* 40: 51–104.

Walker, D. K., M. J. Bane, and A. S. Bryk (1973) *The Quality of the Head Start Planned Variation Data* (2 vols.) Cambridge, MA: Huron Institute.

Yurchak, N.J.H. (1975) *Infant Toddler Curriculum of the Brookline Early Education Project.* Brookline, MA: Brookline Early Education Project.

Zigler, E. F. and E. C. Butterfield (1968) "Motivational aspects of changers in IQ test performance of culturally deprived nursery school children." *Child Development* 39: 1–14.

Zigler, E. F. and V. Seitz (1980) "Early childhood intervention programs: A reanalysis." *School Psychology Rev.* 9(4): 354–368.

Zigler, E. F. and P. K. Trickett (1978) "I.Q., social competence, and evaluation of early childhood intervention programs." *Amer. Psychologist* 33: 789–798.

Zigler, E. F., W. D. Abelson, and V. Sietz (1973) "Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test." *Child Development* 44: 294–303.

Martha B. Bronson is an educational psychologist who specializes in observational research methods for studying young children's behavior in preschool and elementary school classrooms.

Donald E. Pierson was the director of the Brookline Early Education Project since the inception of the longitudinal study in 1972. He is currently Professor of Education at the University of Lowell in Massachusetts.

Terrence Tivnan is currently Assistant Professor at the Harvard Graduate School of Education, where he teaches courses in research methods and data analysis.

Explanation and Critique

This evaluation is a reexamination of children observed earlier, while in kindergarten. At the time of the evaluation, they were in the second grade. Martha Bronson, Donald Pierson, and Terrence Tivnan provided a posttest comparison of children who received early education programming with similar children who did not. One goal was to determine whether the programming had a long-term impact on children's competence as defined by the consumers of the evaluation. Competence was assessed on the basis of the children's mastery of task skills, social skills, and the use of time.

Parents of children who could potentially participate in the BEEP program were recruited at the time the children were born. The recruitment included births in Brookline, Massachusetts, a suburb of Boston, and low-income minority family births in Boston. Participation was not mandatory; families volunteered for the program. The problems

of self-selection are severe and will be discussed later.

The researchers concluded that BEEP had a positive impact on children's in-classroom behavior and sought to see if that impact was carried over into the second grade and whether the form of the impact changed over the two-year period. The subjects were the 169 second graders who had participated in the BEEP from birth to kindergarten. The researchers observed the children who had left the school district as well as those still in the district.

The children in the comparison group were chosen randomly from the classrooms of the BEEP students, matching them by sex. This method was used to control for variations in classroom exposure, teacher style, or any other variable in the learning environment. The characteristics of the BEEP children and comparison students are shown in table 1 of the article. The comparison group members were similar to the BEEP students.

The evaluators used trained classroom observers to gather data for the BEEP and non-BEEP students. Notice that they did not divulge the specific purposes of the study or the experimental/comparison status of the students to the observers. This was done to maximize the observers' objectivity. The time-sampling technique and the validity of the instrument are reasonable and well documented.

Tables 2 and 3 of the article summarize the comparisons of the BEEP and non-BEEP students. Notice that the tests of significance were for matched pairs, not the typical *t*-test of differences of group means. This is appropriate because the research question revolves on the differences in otherwise similar matched students. The findings indicate that BEEP students scored significantly higher than the comparison group on all measures of mastery skills and on some measures of social skills. In addition, BEEP students had a lower incidence of problems or concerns in these two areas. The BEEP and non-BEEP students did not differ in their use of time.

Figure 2 of the article provides some measure of cost effectiveness. Bronson, Pierson, and Tivnan controlled for the home educational environment by examining separately the percentage of children with educational competence who had college-educated mothers and those whose mothers were not college-educated. Presumably they used the education of the mother because mothers traditionally spend more time with children than fathers do. The article is not explicit in this regard. The findings indicated that children of college-educated mothers scored higher on competence ratings than those whose mothers were not college-educated. It appears from figure 2 that, even when given the most expensive early education program, children of mothers who were not college-educated barely exceeded in competence those nontreated comparison students whose mothers were college-educated. Level A training affords children the best chance of increased competence, according to figure 2, whereas the program type does not seem to make much difference among children of college-educated mothers. The policy implications indicate that to decrease the competence gap between those whose parents are more educated and others, the early education program at Level A should be targeted to children whose parents do not have a college education.

But, in our opinion, the research was virtually fatally flawed because the authors were unable to rule out self-selection as an explanation for differences in the experimental and comparison groups. To recap, the researchers compared second grade children who had participated in the preschool program—the Brookline Early Education Program (BEEP)—with other second graders who had not. The preschool program was voluntary and had as its main goal the reduc-

tion or elimination of school problems. The research project sought to assess the impact of the program after several years had passed since the children had participated in BEEP. The investigators matched each BEEP participant with a control by selecting a child at random from within the same classroom and sex group. The investigators found that the demographic characteristics of the two groups were quite similar.

BEEP children showed significant advantages over the comparison group in both mastery and social skills with the largest differences being in the mastery skill area. In addition, BEEP children with highly educated mothers showed significant advantages over comparison children regardless of the BEEP program level experienced by the participants (again, see figure 2 in the article). The researchers conclude:

> the results of the study suggest the importance of a carefully planned program like the Brookline Early Education Project *for all school systems* [emphasis added]. Education and support services to parents of young children coupled with early education programs for the children should be recognized as an essential part of a high quality elementary school curriculum.

In our view, the authors have not made their case. The selection of the comparison group (demographically similar children and families) does not control for all elements necessary to overcome the self-selection bias. Is it not possible that, on average, BEEP parents show more concern for their children and thus instill "better" behavior patterns in them through more training and care?

Let us look at the evidence in figure 2. First, take the college graduate mothers. There is a difference between the BEEP children and comparison group for these mothers, *but there is no difference between BEEP children based on the level of BEEP intervention.* This suggests to us that self-selection and all that it means in a child-rearing situation *may* be a relevant factor.

Second, examine the scores for children with non–college graduate mothers. The relationship shown in figure 2 indicates some difference between the comparison group and all BEEP children, no difference between BEEP C and BEEP B treatments, and a significant difference between BEEP A and BEEPs B and C. Is it not plausible that self-selection differentiates BEEPs from non-BEEPs and that only the BEEP A treatment is really effective?

If these interpretations of figure 2 are accurate, then the reader would be forced to conclude that the only effective BEEP program is the BEEP A for children of non–college graduate mothers.

This is a far cry from the authors' conclusions quoted earlier. This is not to say that the conclusions drawn here are more valid than the authors', only that the authors' comparison group does not overcome the self-selection problem. The difference between the BEEP children and the children in the comparison is probably explained by differences in parenting *and* the BEEP program. Not enough information exists, however, to weight the relative importance of either factor.

In conclusion, this article documents a carefully designed posttest-only comparison group design in terms of the collection of outcome data and selection of the comparison group. The major shortcoming of the research lies in the problem posed by the selection effect threat to the internal validity of the evaluation. When searching for other plausible explanations for the outcomes of programs, evaluators should take pains to be candid about both the strengths and weaknesses of their approach.

PART IV

Reflexive Designs

The classifications of experimental designs and quasi-experimental designs, the subjects of the previous two parts of this book, are relatively straightforward. That is, they are easy to define. Not so with the subjects of this part, which is concerned with evaluations that use the target group of the program as its own control. Such designs are termed reflexive controls by Peter Rossi and Howard Freeman (1985, 297). Donald Campbell and Julian Stanley considered them to be pre-experimental designs (1963, 171–247). The two types of reflexive designs covered by this part are simple before-and-after studies that are formally known as the one-group pretest-posttest design and the simple time-series design.

For full-coverage programs (i.e., for programs for which most of the population is eligible), it may be impossible to define randomized or constructed control groups or, in fact, to locate nonparticipants. Many programs, for example, are directed at all targets within a specific geographic area. In these cases, the researcher may have no choice but to use reflexive controls.

The essential justification of the use of reflexive controls is that in the circumstances of the experiment it is reasonable to believe that targets remain identical in relevant ways before and after the program. Without the program, pretest and posttest scores would have remained the same.

References

Campbell, Donald T. and Julian C. Stanley. 1963. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching*, edited by N.L. Gage. Chicago: Rand McNally, 171–247.

Rossi, Peter H., and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach.* 3d. ed. Beverly Hills, Calif.: Sage, 297.

CHAPTER 11

One-Group Pretest-Posttest Design

PROBABLY THE MOST commonly used form of reflexive design is the one-group pretest-posttest design, sometimes called comparisons of "before" and "after" data. Not a very powerful design, it is subject to most of the traditional invalidity problems. It is typically used when nothing better can be done. The one-group pretest-posttest design is shown in table 11.1.

The target group is measured before the implementation of the program (O_1) and again after the program is completed (O_2). The difference scores are then examined and any improvement ($O_2 - O_1$) is usually attributed to the impact of the program. The major drawback to this design is that changes in the target may be produced by other events and not the program. The longer the time lapse between the preprogram and postprogram measurements, the more likely it is that other variables besides the program affected the postprogram measurement. Harry Hatry, Richard Winnie, and Donald Fisk outline the conditions under which this design might be applied:

> This design often is the only type that is practical when time and personnel are limited. It is most appropriate (1) when the period covered by the evaluation is short (this making it less likely that non–program related factors will affect the evaluation criteria); (2) when the link between the program intervention and the outcomes being measured is close and direct so no other major events are likely to have had a significant influence on the values measured with the evaluation criteria; or (3) when the conditions measured have been fairly stable over time (and are not, for example, likely to be distorted by seasonal changes), and there is reason to believe such stability will continue. (1981, 28)

Rossi and Freeman give a "hypothetical" example:

> In evaluating the outcome of a nutritional education program testing participants' knowledge of nutrition before and after participation in a three-week set of lectures, the use of reflexive controls is likely

TABLE 11.1
One-Group Pretest-Posttest Design

	Pretest	Program	Posttest
Group E	O_1	X	O_2

to provide a good measure of the impact of the course because knowledge of nutrition is unlikely to change spontaneously over such a short period of time. (1985, 297–98)

The following article—"Nutrition Behavior Change: Outcomes of an Educational Approach" by Patricia Edwards, Alan Acock, and Robert Johnston—reads much like the Rossi and Freeman example.

READING

Nutrition Behavior Change
Outcomes of an Educational Approach

Patricia K. Edwards
Virginia Polytechnic Institute and State University

Alan C. Acock
Louisiana State University

Robert L. Johnston
Virginia Polytechnic Institute and State University

This study addresses four issues in the evaluation of nutrition education programs: (1) the reliability of knowledge, belief, and behavior scales; (2) the effectiveness of programs targeted to the general public; (3) the longitudinal effects of nutrition education interventions; and (4) the relationship between changes in the cognitive, belief, and behavioral domains. Our findings indicate that reliable knowledge and behavior scales can be developed, but that the internal consistency of belief scales are more problematic. Moreover, improvements in all three domains can be attained with an heterogenous target audience. Although knowledge deterioriates after the course is completed, beliefs remain stable and nutrition behavior continues to improve significantly. Finally, changes in knowledge and beliefs are influential on changes in behavior as a result of the course, but postcourse changes in knowledge and beliefs are not associated with changes in behavior.

AUTHORS' NOTE: *This evaluation was funded by the American Red Cross and the United States Department of Agriculture. We would like to thank the members of the nutrition course development team: Barbara Clarke, Kristen DeMicco, Mary Ann Hankin, Louise Light, Patricia Marsland, Rebecca Mullis, Anne Shaw, and Fred Troutman for their assistance in the development of the outcome measures used in this study.*

A MAJOR RESPONSIBILITY of evaluators is to test the basic theoretical premises underlying the delivery of human service programs (Flay and Best, 1982; Neigher and Schulberg, 1982). With respect to health education programs, numerous studies have examined the assumption that an expansion in pertinent knowledge and positive changes in beliefs is associated with improvements in health-related behavior. Findings from an array of programs—for example, weight reduction (Becker et al., 1977), alcohol use (Goodstadt, 1978), smoking (Thompson, 1978), and breast self-examination (Calnan and Moss,

1984), suggest that the viability of health education in promoting positive health behavior is highly problematic. Moreover, the nature of causal linkages between knowledge, belief, and behavior changes remains in question. Although the consistency model predicts that valid knowledge change is the first stage in a pathway proceeding to changes in beliefs and attitudes—which then culminate in behavior change (Swanson, 1972; Stanfield, 1976; Zeitlan and Formacion, 1981)—some researchers propose alternative causal models. Changes in attitudes may be, in fact, a necessary prior step to the acquisition of health-related knowledge (Mushkin, 1979; Rosander and Sims, 1981), and health-related beliefs can be modified subsequent to changes in behavior to maintain consistency among the affective and behavioral domains (Almond, 1971; McKinlay, 1972).

This body of research is highly relevant to the evaluation of nutrition education programs. However, although a considerable amount of investigation has focused on the magnitude of cognitive, belief, and behavioral outcomes, extant studies suffer from a number of constraints that limit our understanding of the basic premises and potential of nutrition education programs. Firstly, most of the research has examined programs designed for and targeted to specific subpopulations such as low-income mothers (Ramsey and Cloyd, 1979; Rosander and Sims, 1981), primary school children (St. Pierre and Cook, 1981), hospital staff (Looker et al., 1982), and pregnant teenagers (Perkin, 1983). Little is known about the efficacy of programs aimed at a broadly based constituency. Secondly, to our knowledge, there are no studies involving nutrition education programs that report the extent to which positive cognitive, belief, and behavioral changes are sustained over time. Most health education programs are of a relatively short duration, but their objectives—particularly those that are behavior oriented—are not timebound (Hochbaum, 1982; Flay and Best, 1982). Despite long-range expectations, evaluations of the effectiveness of nutrition education programs, and most other health change programs as well, are generally confined to a pretest-posttest design.

A third problem endemic to many nutrition education evaluations deals with the reliability of outcome measures. In some studies the reliability of scales is estimated from tests piloted with samples that are not equivalent to the target audience of the program (i.e., Looker et al., 1982). We cannot assume that because a test is reliable for one population it will be equally reliable for others (Talmage and Rasher, 1981). Other studies, which use internal consistency as a measure of the reliability of their scales, present pretest results only, ignoring the possibility that the scales may have a different level of internal consistency for the posttest stage (i.e., St. Pierre and Cook, 1981; Rosander and Sims, 1981). The difficulty in constructing reliable outcome scales, even when they are administered to fairly homogeneous groups, is demonstrated by the wide variation of reliability coefficients. For example, Sullivan and Schwartz (1981) report a coefficient as low as .00.

Finally most nutrition education program evaluations measure the intended effects of the program in isolation, making assumptions regarding causal linkages between outcome domains, while failing to demonstrate the actual relationships between cognitive, belief, and behavioral changes. Thus, it is impossible to determine if there is, indeed, an empirical association among the knowledge, belief, or behavioral changes that do occur.

This study addresses four issues in the evaluation of nutrition education programs: (1) Can reliable and valid measures of nutrition knowledge, beliefs, and behavior be developed that will enable longitudinal assessment of programs targeted to the general

public? (2) Can a program directed to a broad audience be effective in terms of changing nutrition knowledge, beliefs, and behavior? (3) Are positive changes sustained over time? (4) What is the relationship of changes in the cognitive, belief, and behavioral domains? At a time when federal funds for nutrition education programs have been drastically curtailed, these issues are of critical importance in terms of demonstrating the potential of educational interventions as a strategy for promoting positive dietary behavior.

The American Red Cross Nutrition Course Evaluation

This article draws from the results of an evaluation of a nutrition course, "Better Eating for Better Health," developed jointly by the American Red Cross (ARC) and the United States Department of Agriculture (USDA). The primary goals of the course—presently being offered by many of the 3000 ARC chapters nationwide—concern the promotion of nutrition knowledge, positive beliefs, and improved dietary behavior of the general public. The curriculum consists of six two-hour modules that may be presented over a period ranging from two to six weeks. Participant workbooks and supplementary reading materials are provided to each attendee. Prior to conducting a course, ARC instructors are expected to have completed the Nursing and Health Services core curriculum, as well as the Nutrition Instructor Specialty Course. Evaluation was carried out in five separate stages concurrently with the development of the course and included both formative and summative elements. The selected findings reported here are derived from the final stage of the project, a national field test, conducted at 51 ARC chapter sites.[1]

Data Collection Procedures

Course participants were surveyed at three distinct points in time: prior to the beginning of the nutrition course, immediately after the last session of the course, and approximately 10 weeks following completion of the course. Group-administered questionnaires, provided by course instructors who had been trained to implement the surveys, were used to collect baseline (N = 1461) and posttest (N = 1031) data. All individuals attending the first and last sessions of the nutrition course participated in the first two surveys. Although we were unable to use random selection for our sampling procedures, because participation in ARC classes is voluntary, our baseline sample is differentiated along an array of sociodemographic variables.[2] The third participant survey was conducted by means of telephone interviews with a systematically selected subsample of baseline respondents. A total of 248 interviews were attempted to achieve a quota of 200 telephone survey respondents, accounting for a completion rate of 81%. Respondents to the telephone survey are representative of the initial sample along the range of background characteristics.[3]

In order to assess the possible effects of exogenous factors on changes in the nutrition course participant's knowledge, beliefs, and behavior, we solicited volunteers for a nonequivalent control group from individuals who were simultaneously attending other Red Cross courses during the field test. Again, random selection was not possible without severe disruption of ARC chapter activities. The control group completed both the baseline (N = 212) and posttest questionnaires (N = 133). An earlier analysis indicates that although there are some statistically significant sociodemographic differences between the experimental and control groups, these differences are not influential in interpreting the

effects of the nutrition education intervention (Edwards et al., 1983a: 15–17, 100–107).

Measurement Techniques

Items for the outcome scales were constructed cooperatively by the course development and evaluation teams, composed of personnel from ARC and USDA, as well as outside consultants. Each item reflects a major knowledge, belief, or behavioral objective of the course.[4] The three outcomes scales were initially tested with a prototype sample of participants during the third stage of the evaluation, which constituted of a "best chance" pilot assessment of course materials and teaching strategies at six ARC chapter sites. (The nutrition course was taught by experienced ARC instructors who were assisted and observed by members of the course development and evaluation teams). Subsequent to this stage of the evaluation, the course development team revised course objectives, materials, and teaching strategies to more adequately coincide with participant needs identified in the evaluation. In addition, the evaluation scales were examined, using factor analysis and alpha reliability to assess the accuracy and validity of each measure. Inadequate items were deleted, and new measures reflecting changes in the nutrition course objectives were added.

The evaluation team conducted a second pilot test of the revised instruments at ten ARC sites during the fourth stage of the evaluation, in a more naturalistic setting, without either observation or supervision. Again, changes were made in the instrumentation to improve the validity and reliability of the items, as well as to ensure that the items represented the fourth-stage modifications in the nutrition course.

The final nutrition knowledge scale consisted of 15 multiple-choice items pertaining to facts about nutrients, sodium, vitamins, food additives, weight loss, and the relationship of disease to nutrition. Respones were coded so that a score of one indicates a correct answer and zero an incorrect response. Nutrition beliefs, operationalized as a dimension of the affective domain (see Fishbein and Raven, 1962), were constructed using a five-category Likert-type scale. The items were coded so that a score of one represents the least positive belief, three denotes the respondent was undecided, and five indicates the most desirable response in terms of the course objectives. The scale included 8 items concerning beliefs that related to the content areas addressed in the knowledge items. The nutrition behavior scale measured the frequency of participant's conduct related to the knowledge and belief questions. The 12 items making up this scale were coded so that one represents the least positive behavior pattern and five the most desirable, according to course objectives. Belief and behavior items were constructed in both positive and negative terms to avoid a response set.

Reliability of Nutrition Outcome Scales

Developing adequate scales to measure nutrition outcomes has been a thorny problem confounding the evaluations of nutrition education programs. Even when a single domain of outcomes is divided into subtests dealing with specific content areas of nutrition knowledge, beliefs, or behaviors, the internal consistency of each subscale is often unacceptable (i.e., Sullivan and Schwartz, 1981). Moreover, as St. Pierre and Cook (1981) illustrate, the reliability of scales may vary considerably among subsamples of a target population.

The Cronbach alpha coefficients for the field test surveys shown in Table 1 demonstrates the problematic nature of achieving reliable nutrition outcome measures. For

course participants, the inter-item reliability is only slightly differentiated between the baseline and posttest points on the knowledge and behavior scales. Despite the fact that the scales were administered to a heterogeneous population and include a range of content areas, the reliability coefficients are within an acceptable range. Both scales were examined using principal component factor analysis. All 15 items of the knowledge scale have a positive loading on the first factor, accounting for 21.9% of the variance on the baseline and 25.6% on the posttest. There is a clear first factor on the 11-item behavior scale, explaining 36.0% and 39.1% of the variance on the baseline and posttest, respectively.

Measurement of nutrition beliefs on a single scale had been highly problematic from the start of the project. On our pretest analysis, it was evident that many of the belief objectives of the course were not integrated along a single dimension. Baseline and posttest reliabilities for the eight items that were finally selected are relatively low for course participants. The eight-item scale has a first principal factor explaining 25.8% of the baseline and 29.7% of the posttest variance.

Comparisons between the reliability coefficients for the participant and control group baseline scales show a great deal of consistency for the knowledge and behavior scales. The reliability coefficient for the control group belief scale, however, is considerably lower than that for the treatment group. This difference persists in the comparison of the posttest coefficient. Moreover, the knowledge reliabilities deteriorate for controls on the second test, to a less satisfactory range. The differences between the reliability of scales administered to both groups may, indeed, be a result of the fact that they constitute samples of two separate populations. The participant group, by virtue of electing to take the course, perhaps had a more acute "nutrition awareness," resulting in the higher level of internal consistency of responses. These findings underscore the need to pretest scales with controls, as well as experimentals, when nonequivalent samples are expected.

TABLE 1
Reliability of Nutrition Knowledge, Belief, and Behavior Scales

		Alpha Reliability				
		Course Participants			***Control Group***	
Scale	*Number of Items*	*Baseline*	*Posttest*	*Telephone*	*Baseline*	*Posttest*
Field Test Survey Participants and Controls						
Knowledge	15	.75	.79	—	.72	.62
Beliefs	8	.56	.65	—	.46	.49
Behavior	11	.79	.82	—	.82	.83
Number of respondents		1461	1031	—	212	133
Telephone Survey Participants						
Knowledge	7	.66	.70	.58	—	—
Beliefs	5	.53	.60	.69	—	—
Behavior	7	.78	.77	.71	—	—
Number of respondents		196–199	137–147	196–200	—	—

Note: Number of respondents varies due to missing data. The lower number of respondents included in the posttest results reflects the proportion of participants dropping the class before it was completed, approximately 25%.

In order to keep length of the telephone survey to a minimum, it was necessary to reduce the number of items in each scale for this follow-up survey. Table 1 presents a reexamination of the scales in their reduced form for telephone respondents only. The results show that decreasing the number of items and the pool of respondents has a deleterious effect on scale reliability. Of particular concern here is the deterioration in the interitem coefficients for the knowledge scale used in the telephone survey. Our findings suggest that caution must be taken when developing multiple choice knowledge items for telephone surveys. Although it may be necessary to utilize several different data collection techniques in longitudinal studies, it is evident that each technique must be thoroughly pretested to assure consistent results.

As mentioned previously, prior research has shown that scale reliability can vary among sociodemographic subsets of a sample. When these groups are important in assessing differential outcomes of the program under examination, their responses should be analyzed for scale reliability. Table 2 presents reliabilities for the baseline knowledge, belief, and behavior scales by gender, race, age, education, and income. The findings show that pattern of scale reliability found in the aggregated sample is generally maintained among subsets of respondents. However, none of the scales perform as well for nonwhites and the youngest group of participants. Furthermore, those participants with the lowest level of educational attainment have considerably lower knowledge and belief scale reliabilities than participants who have completed high school or college. Although these findings illustrate the need to test for scale reliability among relevant subsets of a sample, the overall consistency found here does not pose a serious problem in interpreting outcome results for this study.

TABLE 2
Reliability of Baseline Scales by Selected Subsamples of Participants

	Knowledge	*Beliefs*	*Behavior*	*N*
Gender				
Male	.78	.49	.81	177
Female	.74	.56	.78	1269
Race				
White	.72	.57	.80	1105
Nonwhite	.68	.43	.75	298
Age				
<24	.66	.47	.75	219
25–54	.74	.56	.78	855
54+	.76	.60	.78	320
Education				
<HS	.63	.42	.78	143
HS/some college	.69	.51	.77	919
Bachelors +	.69	.56	.82	380
Income				
<$15,000	.67	.50	.77	353
$15,000–29,999	.71	.49	.80	427
$30,000+	.71	.58	.80	468

Note: Variation in the total number for each subsample is due to unreported data.

Effectiveness of Nutrition Education on a Variegated Constituency

The second issue posed in this study concerns the potential of a health education course targeted to a broadly based audience. Table 3 shows that the ARC nutrition course had substantial positive cognitive, belief, and behavioral effects on the overall sample of participants.[5] In contrast, no significant changes are found between the baseline and posttest scores for the control group, despite the fact that their baseline means are not statistically different from those of the treatment group. The key point is that the control group makes no statistically significant improvement and, therefore, the improvement in the treatment group is most reasonably attributed to the course itself.

Our analysis further indicates that positive knowledge, belief, and behavior changes

are consistent among subgroups of course participants. There are no statistically significant differences in improvements on the basis of sex, race, and income in the three outcome measures. Although age is a factor that differentiates the degree of positive effects, of the four age groups examined (under 19, 19 to 24, 25 to 54, and 55 or over), all made significant improvements in nutrition knowledge and behavior, and only those participants under 19 years of age did not gain in terms of positive nutrition beliefs. Marital status and education also affect the level of change. Nonetheless, statistically significant and substantial improvements in nutrition knowledge, beliefs, and behavior are found within each category of these two variables. Our results show that although sociodemographic variables have selected influences on how much participants gain from the course, they do not differentiate to the extent that certain groups fail to improve their nutrition knowledge, belief, or behavior after taking the course.[6]

One reservation must be noted here, however. Participation in the ARC nutrition course is, by design, voluntary. There is no reason to expect that we could find the same dramatic positive results with a group of participants who had not been motivated to participate in the course. Our reservations may, however, be tempered by the fact that the reasons for participating in a health-related Red Cross class are extremely complex. Analysis of open-ended responses illicited from course participants indicates that though many of the respondents decided to attend the course to improve their own dietary behavior, others were motivated to do so because the course was job-related or provided an opportunity for social interaction.

Are Positive Changes Sustained Over Time?

Our third concern in evaluating a nutrition education program relates to the longitudinal effects of the intervention. Table 4 presents the mean changes in nutrition knowledge, beliefs, and behavior for the modified instruments used in the telephone survey undertaken approximately ten weeks after the course was completed. Our analysis only in-

TABLE 3
Changes in Nutrition Knowledge, Beliefs, and Behavior: Immediate Effects

	Participants				*Controls*			
Scale	*Means (N = 883)*	*S.D.*	*Change*	*Probability (2-tailed)*	*Means (N = 104)*	*S.D.*	*Change*	*Probability (2-tailed)*
Knowledge								
Baseline	8.91	3.24	1.96	.000	8.09	3.36	.10	NS
Posttest	10.87	3.24			8.19	3.14		
Beliefs								
Baseline	27.49	4.17	3.29	.000	26.82	3.67	–.10	NS
Posttest	30.78	4.38			26.72	3.70		
Behavior								
Baseline	36.00	7.76	5.05	.000	36.08	8.01	.94	NS
Posttest	41.05	7.10			37.02	7.91		

Note: Possible ranges for each scale are: Nutrition knowledge, 0–15; Nutrition beliefs, 5–40; Nutrition behavior, 5–55. It should be noted that the absolute degree of change cannot be compared across the scales due to differences in the length and coding of the instruments. The probabilities are based on individual T-tests.

TABLE 4
Telephone Survey Comparison: Follow-Up Effects

	Knowledge				*Beliefs*				*Behavior*			
	N	*Means*	*S.D.*	*Change*	*N*	*Means*	*S.D.*	*Change*	*N*	*Means*	*S.D.*	*Change*
Baseline to posttest	147	5.11 5.93	1.80 1.50	.82*	136	16.62 19.15	3.05 2.98	2.53*	132	22.32 25.48	5.50 4.82	3.16*
Posttest to telephone	147	5.93 5.69	1.50 1.39	-.24*	137	19.12 18.71	2.99 3.18	-0.41	132	25.56 26.63	4.79 4.39	1.07*
Baseline to telephone	200	5.21 5.69	1.75 1.42	.48*	197	16.67 18.62	3.06 3.20	1.95*	193	21.84 26.33	5.49 4.84	4.49*

Note: Possible ranges for each scale are: Nutrition knowledge, 0–7; Nutrition beliefs, 5–25; Nutrition behavior, 5–35. Because some of the original items were deleted in the telephone survey instruments, these scores are not comparable to the change scores in Table 2.
*p < .05 based on individual T-tests.

cludes participants who had been involved in the follow-up telephone survey. Because we wanted to acquire further information from individuals who had, for some reason, dropped the course, the attrition rate is reflected in the lower number of cases available for analyses involving posttest responses.

The first two rows in Table 4 provide the change when baseline and posttest scores are compared. These changes are substantial and significant for all three measures in a positive direction. Thus, the outcome results for this subsample of participants are consistent with the full complement of participants. We expected a significant drop-off in the scores between the posttest and the telephone survey. In the weeks that elapsed between the two measurements, a great deal of information and the motivation to change behavior could be lost. Indeed, there is a significant reduction in the score on the nutrition knowledge scale. We can conclude that although participants learn a great deal initially, a substantial proportion of this new knowledge is lost soon after completing the course. In contrast, although there is some loss in terms of positive beliefs, the deterioration is not statistically significant. What is, perhaps, most surprising is the significant improvement in the quality of behavior at the end of the ten week period. Finally, looking at the comparison of the baseline and telephone responses in the last two rows of Table 4, we can see that, despite the drop in knowledge subsequent to completion of the course, knowledge, belief, and behavior improvements persist.[7]

We might speculate that the initial improvement in nutrition knowledge provides a cognitive influence in the beliefs that participants have about good nutrition. Even after the students lose some of their specific knowledge, the improved beliefs are retained. Perhaps behavioral changes become self-reinforcing elements in the participant's lifestyle.

The Relationship of Cognitive, Belief, and Behavioral Domains

Table 5 presents the correlations between changes in the outcome domains for two data sets. The first column of correlations shows the changes between the baseline and the posttest. The second column does the same for the changes between the posttest and subsequent telephone interview. These correlations refer to change scores rather than the actual scores on the scale themselves. For example, the correlation of .21 between knowl-

edge and belief means that the more a person improves their knowledge between the baseline and the posttest, the more they also improved their beliefs.

The changes between the baseline and the posttest are positively correlated across all three domains. These correlations are modest, but all are statistically significant. The more a person improves his or her knowledge, the more beliefs ($r = .21$; $p \leq .5$) and behavior ($r = .15$; $p \leq .5$) also improve. Similarly, the more one's beliefs improve, the more one's behavior improves ($r = .26$; $p \leq .5$).

Analyzing the pattern of these correlations, we can see that changes in beliefs may have a greater influence on improvements in behavior ($r = .26$) than do changes in knowledge alone ($r = .15$). Although both of these correlations are statistically significant by themselves, the difference between them does not achieve statistical significance ($p = .11$, one-tail). Therefore, we hesitate to generalize these results beyond this particular data set. Nonetheless, these results do suggest that educational interventions should address changing beliefs, along with providing relevant knowledge. At the very least, these results emphasize that changing beliefs are an important covariate of changing behavior.

Examining the correlations that appear in the last column of Table 5 provides very different results. These correlations reflect the relationship between changes in knowledge, beliefs, and behavior that occur from the posttest to the subsequent telephone interviews. Earlier we demonstrated that behaviors continue to improve after the completion of the course, whereas knowledge deteriorates significantly and beliefs deteriorate somewhat, but not significantly. The fact that none of these correlations is statistically significant means that changes after the course in one domain are independent of changes in the two other domains. This is especially true for nutrition behavior. Nutrition behavior improves after the course, regardless of what happens to the participant's knowledge and beliefs following the course. Thus, a subject who forgets much of what is learned from the course is nearly as likely to continue improving nutrition behavior as a person who remembers all of the nutrition information and belief material.

TABLE 5
Correlation of Changes in Nutrition Knowledge, Belief, and Behavior Scales

Change	*Baseline to posttest (N = 883)*	*Posttest to telephone (N = 134)*
Knowledge/belief	0.21*	–0.08
Knowledge/behavior	0.15*	–0.10
Belief/behavior	0.26*	–0.05

*$p < .05$.

These results indicate that change in knowledge and especially change in beliefs are important to produce the initial changes in behavior (baseline to posttest). Just as important, however, the maintenance and enhancement of improved nutrition behavior is largely independent of how much of the knowledge and belief information is retained subsequent to the course.

Conclusions

At the beginning of this article we raised four questions about the measurement of nutrition outcome variables; what can be accomplished on a large and heterogeneous population; the ability to sustain changes over time; and the interdependence of changes in cognitive, belief, and behavioral domains.

This study has the advantage of being a formative evaluation. This allowed two systematic pretests of the measurement scales prior to the actual implementation of the evaluation. There is an additional advantage in terms of having the opportunity to share the empirical results of the pretest with the program personnel and combine the empiri-

cal and expert opinions in developing the final scales. Unfortunately, much evaluation research will not have these advantages.

It was possible to develop a highly reliable scale to measure nutrition behavior. This high reliability is consistent using the group-administered long form of the questionnaire (11 items) for the case of both the baseline and the posttest in both the treatment group and in the control group. The long-form result is also consistent with the short form of the scale (7 items) used in the telephone interviews. This consistency is evident, even though fewer items were used and the method of measurement was changed from group-administered to telephone interview.

Beliefs are consistently problematic. The reliability for the control group is poor. Interestingly, for the experimental group there is improvement in the reliability between the baseline and the posttest and even more improvement moving from the posttest to the telephone survey. Perhaps there is a tendency to organize one's beliefs about nutrition as a result of the course. Thus, the respondent's nutrition beliefs not only improve but become more integrated due to the systematic exposure to nutrition information in the course.

Knowledge is reliable, but the reliability appears to drop off on the telephone follow-up. Perhaps this is because knowledge is the most likely of the three domains to deteriorate after the course is completed. Thus, nutrition knowledge not only deteriorates, but the integration of that knowledge in a coherent fashion also drops off.

A general conclusion regarding the reliability of the scales is that it is possible to develop reasonably reliable sales even though we are applying them to a heterogeneous population and using different methods to administer them. Impressively, the telephone interviews do nearly as well as the group administered questionnaire method even though the telephone interviews have fewer items in the scales.

The second major question concerns the ability to produce changes in a broad audience. Our results show that progress is substantial with a diverse population. However, as we have dealt with a volunteer population, we do not know if such impressive improvements in knowledge, belief, and behavior are possible with a less motivated, nonvoluntary population.

The third issue concerns the ability to sustain changes over time. What happens after the course was completed? As our results show, there is a significant deterioration in nutrition knowledge, a slight (nonsignificant) decrease in the quality of beliefs, and an actual improvement in nutrition behavior after the course is completed.

Although we had hoped behavior would be sustained after the course, we were surprised that it actually improved significantly. One possible explanation for this improvement can be gained by examining some of the classic studies conducted by social psychologists on the effect of fear on changes in attitudes and behavior. Leventhal (1980) reports great inconsistency in the results of the use of fear in change programs. He suggests that fear produces two possible effects. One of these is a realistic fear that can be controlled by directed changes in behavior. Thus, when people are scared and are told what specific behavior will mitigate this fear, they are likely to change their behavior dramatically.

The nutrition course did not intend to produce high levels of fear among the participants. Still, the substantial changes in knowledge and beliefs could be expected to make people fearful of practicing poor nutrition behavior. On the other hand, proper nutrition conduct could be expected to give a very positive feeling because it mitigates concerns developed because of the improved knowledge and beliefs.

This argument means that proper dietary behavior may become a self-reinforcing behavior. For example, proper diet makes

people feel good because they do not need to be concerned about their health (at least as far as nutrition is a health issue). Even after the participant forgets some of the details they learned in the course, the good feeling associated with proper nutrition behavior retains its self-reinforcing property and continues to improve.

If this argument is valid, then the findings have far-reaching implications for change programs. The creation of moderate levels of fear based on improved knowledge and beliefs combined with clear guidelines on specific behavior to control such fear can make the behavior self-reinforcing. Once the behavior becomes self-reinforcing, it can continue to improve long after the completion of the intervention and in spite of a deterioration in the level of specific knowledge and beliefs.

The final major issue of this article is the relationship among the three domains (cognitive, belief, and behavioral) of nutrition. Some interventions focus on only one or two of these domains, however, we have demonstrated that all three are interrelated. In particular, positive beliefs appear to be even more important improvement than knowledge as a prescursor of changes in behavior. Although changes in knowledge and beliefs are influential on changes in behavior as a result of the course, postcourse changes (drop-off) in beliefs and knowledge are not associated with changes in behavior. This is further evidence that nutrition behavior can become self-reinforcing after the completion of the course.

Notes

1. This article reports only selected findings from the field test, which included a more comprehensive analysis of the relationship between participant attributes, program delivery variables, (e.g., instructor training and experience) exogenous factors, and a variety of program outcomes, such as attendance, extent of involvement, level of difficulty of the course, and participant satisfaction with course content, materials, and teaching strategies. A description of the evaluation model and results of the analysis have been reported elsewhere (Edwards, 1984; Edwards et al., 1983a; Edwards et al., 1983b). Data from two additional field test surveys directed to course instructors and chapter administrators are not included in this analysis.
2. For example, gender: female (88%); age: <25 years (16%); 25–34 (24%); 35–54 (39%); 55+ (21%); marital status: married (54%), single (25%), divorced/widowed (21%); ethnicity: white (77%), black (17%), Hispanic (4%), other (2%); household income: <$15,000 (28%); $15,000–34,999 (44%); $35+ (28%); employment status; full time (43%), part time (15%), not employed (42%); education <high school (10%), high school graduate (28%), some college (36%), college graduate (26%).
3. There are no statistically significant differences among the baseline sample and telephone respondents in terms of sex, age, educational attainment, income, race, marital status, mean household size, or participant's perception of their personal health status.
4. To illustrate, an objective of session 3 was to clarify knowledge of health problems related to excess sugar consumption and suggest strategies for reducing sugar in the diet. Examples of survey questions used to measure this objective within each outcome domain are shown below.

   Knowledge: The most common disease in the United States related to excessive sugar consumption is:

   (1) diabetes

   (2) tooth decay

   (3) obesity

(4) cirrhosis

(5) gastrointestinal disorders

Beliefs: Honey and molasses are better for you than table sugar (strongly disagree, disagree, undirected, agree, strongly agree).

Behavior: (I) eat fruits canned in their own juice without sugar added rather than fruits canned in syrup (never, rarely, sometimes, usually, always).

The complete sets of scale items can be obtained from the authors.

5. Analysis of the effects of the course showed that sociodemographic factors explained only a small proportion of the variance on those outcome measures (Edwards et al., 1983a).
6. Multivariate analyses of the effect of process and participant perception variables (e.g., extent to which the course design was followed, team versus individual instruction and participant assessments of course content, activities, length, difficulty, and quality of instruction) on the three outcome domains are reported elsewhere (Edwards, 1984; Edwards et al., 1983a).
7. Because 25% of the telephone sample had dropped the course, these results may underestimate the longer range positive effects of the intervention for participants who attended most of the course.

References

Almond, R. (1971) "The therapeutic community." *Scientific Amer.* 224: 34–42.

Becker, M. H., L. A. Maiman, J. P. Kirscht, E. P. Haefner, and R. H. Drachman (1977) "Dietary compliance: a field experiment." *J. of Health and Social Behavior* 18: 348–366.

Calnan, M. W. and S. Moss (1984) "The health belief model and compliance with education given at a class in breast self-examination." *J. of Health and Social Behavior* 25: 198–210.

Edwards, P. K. (1984) "The American Red Cross nutrition course: findings from the field test," pp. 575–586 in *Agricultural Outlook 85*. Washington, DC: U.S. Department of Agriculture.

Edwards, P. K., A. C. Acock, R. L. Johnston, and K. L. Demicco (1983a) *American Red Cross Nutrition Course Field Test: Technical Report.* Washington, DC: American Red Cross.

Edwards, P. K., R. Mullis, and B. Clarke (1983b) "A comprehensive model for evaluating innovative nutrition education programs." Paper presented at the Evaluation Research Society meetings, Baltimore, MD.

Fishbein, M. and B. H. Raven (1962) "The AB scales: an operational definition of beliefs and attitudes." *Human Relations* 15: 35–44.

Flay, B. R. and J. A. Best (1982) "Overcoming design problems in evaluating health behavior programs." *Evaluation and the Health Professions* 5: 43–69.

Goodstadt, M. (1978) "Alcohol and drug education: models and outcomes." *Health Education Monographs* 6: 263–279.

Hochbaum, G. W. (1982) "Certain problems in evaluating health education." *Amer. J. of Public Health* 6: 14–20.

Leventhal, H. (1980) "Findings and theory in the study of fear communications," pp. 120–186 in L. Berkowitz (ed.), *Advances in Experimental Social Psychology,* Vol. 5. New York: Academic.

Looker, A., S. Walker, L. Hamilton, and B. Shannon (1982) "Evaluation of two nutrition education modules for hospital staff members." *J. of the Amer. Dietetic Assn.* 81: 158–163.

McKinlay, J. B. (1972) "Some approaches and problems in the study of the use of services: an overview." *J. of Health and Social Behavior* 14: 115–151.

Mushkin, S. J. (1979) "Educational outcomes and nutrition," pp. 269–302 in R. E. Klein et al., (eds.) *Evaluating the Impact of Nutrition and Health Programs.* New York: Plenum.

Neigher, W. D. and H. C. Schulberg (1982) "Evaluating the outcomes of human service programs: a reassessment." *Evaluation Rev.* 6: 731–752.

Perkin, J. (1983) "Evaluating a nutrition education program for pregnant teenagers: cognitive vs. behavioral outcomes." *J. of School Health* 53: 420–422.

Ramsey, C. E. and M. Cloyd (1979) "Multiple objectives and the success of educational programs." *J. of Nutrition Education* 11: 141–145.

Rosander, K. and L. S. Sims (1981) "Measuring effects of an affective-based nutrition education intervention." *J. of Nutrition Education* 13: 102–105.

St. Pierre, R. G. and T. D. Cook (1981) "An evaluation of the nutrition education and training program." *Evaluation and Program Planning* 4: 335–344.

Stanfield, J. P. (1976) "Nutrition education in the context of early childhood malnutrition in low resource communities." *Proceedings of the Nutrition Society* 35: 131–144.

Sullivan, A. D. and N. E. Schwartz (1981) "Assessment of attitudes and knowledge about diet and heart disease." *J. of Nutrition Education* 13: 106–108.

Swanson, J. E. (1972) "Second thoughts on knowledge and attitude effects upon behavior." *J. of School Health* 42: 362–363.

Talmage, H. and S. P. Rasher (1981) "Validity and reliability issues in measurement instrumentation." *J. of Nutrition Education* 13: 83–85.

Thompson, E. L. (1978) "Smoking education programs 1960–1976." *Amer. J. of Public Health* 68: 250–257.

Zeitlan, M. F. and C. S. Formacion (1981) *Nutrition Education in Developing Countries: Study II, Nutrition Education.* Cambridge, MA: Oelgeschlager, Gunn, & Hain.

Patricia K. Edwards is Associate Professor of Urban Affairs in the College of Architecture and Urban Studies at Virginia Polytechnic Institute and State University. She has had experience evaluating international, national, and local programs dealing with public health, housing, recreation, and community development.

Alan C. Acock is Professor of Sociology, Rural Sociology, and Graduate Director in the Sociology Department at Louisiana State University, Baton Rouge. His current research concerns using LISREL to isolate a family effect from individual effects, reviewing intergenerational relations research, and using personal computers in content analysis.

Robert L. Johnston is completing his doctorate in Sociology at Virginia Polytechnic Institute and State University and is an Adjunct Instructor at Roanoke College. His recent research projects include the analysis of labor process development in the post-World War II era.

Explanation and Critique

The first thing a student is likely to say on reading the Edwards, Acock, and Johnston article is this: "This is not an example of the one-group pretest-posttest design; it had a comparison group. It is really a quasi-experimental design." But is it? Look at the article another way.

Assume that you are taking a course in program evaluation at American University and Professor Felbinger is your instructor. Down the hall from you is a class in astrophysics. Professor Felbinger decides to find out how much you learn about evaluation research in the course, so she gives you a comprehensive 200-question multiple choice test on evaluation research on the first day of class. The class performs miserably. She gives you the same test on the last day of class and you all do quite well. She also gives the same test on the first and last days of class to the physics students. They perform miserably on both the pretest and the posttest. Has the use of a comparison group—the physics students—somehow magically transformed Professor Felbinger's evaluation from a reflexive design to a quasi-experimental design? Of course it has not. There is no reason in the world to expect that the population in general, including physics students, would show a statistically significant increase in their knowledge of program evaluation practices during the course of a semester.

Remember, when identifying a comparison group, to carefully match characteristics of the comparison group with those in the experimental group. If, for example, your experimental group consists of cities, make the matches on the basis of similar size, revenue base, services delivered, and so forth. Select groups of people on normal demographic

variables so that the groups are similar in all possible *relevant* respects except the treatment. In the Edwards, Acock, and Johnston article, there is no reason to expect that a comparison group of individuals attending other Red Cross courses (e.g., life saving) would either be similar in relevant respects to those interested in nutrition training or would show a statistically significant improvement in their knowledge of nutrition during the time frame that the nutrition course was being offered. One might call this the use of a pseudo-control group—at most, it controls for the effects of history. It certainly is not a systematically chosen comparison group.

In reality the Edwards, Acock, and Johnston article presents an example of the one-group pretest-posttest design. Volunteers for a "Better Eating for Better Health" course consisting of six two-hour modules (presented over a period ranging from two to six weeks) were tested at three points: before the beginning of the course, immediately after the last session of the course, and approximately ten weeks later. The results showed a statistically significant change in knowledge, beliefs, and behavior (table 3 of the article). Pretty simple, is it not? There is no reason to expect that exogenous factors caused these changes, nor is there any reason to cloud up a simple (and appropriate) design with a pretense of elegance.

The article can be criticized from the standpoint of the comparison group, but the evaluation has its strengths. In particular, the authors handled the problem of measuring the outcome of their nutritional education program very well. They were careful in the selection of their measurement items and took the time to compute reliability coefficients. They also examined reliabilities for knowledge, belief, and behavior scales by gender, race, age, education, and income.

Their analysis indicated significant changes in participants' knowledge, belief, and behavior as a result of the course. Furthermore, there were no significant improvements based on gender, race, or income. Age, marital status, and education, however, did affect the level of change.

Edwards, Acock, and Johnson were quite thorough in this evaluation in that they were also interested in how well the results held up over time. To test this, they used a telephone survey of participants ten weeks after the course was completed—expecting a significant drop in the scores between the original program posttest and the telephone survey. They found a significant reduction in the score on the nutrition knowledge scale, a smaller reduction in terms of positive beliefs (the deterioration was not statistically significant), and a significant improvement in the quality of nutrition behavior at the end of the ten weeks. Thus, despite a drop in knowledge, changes in belief and behavior brought about by the course persist. The authors conclude:

> The creation of moderate levels of fear based on improved knowledge and beliefs *combined* [emphasis in original] with clear guidelines on specific behavior to control such fear can make the behavior self-reinforcing. Once the behavior becomes self-reinforcing, it can continue to improve long after the completion of the intervention and in spite of a deterioration in the level of specific knowledge and beliefs.

Without the second posttest, such conclusions would never have been possible.

References

Campbell, Donald T., and Julian C. Stanley. 1963. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching*, edited by N.L. Gage. Chicago: Rand McNally, 171–247.

Rossi, Peter H., and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach*. 3d ed. Beverly Hills, Calif.: Sage.

CHAPTER 12

The Simple Time-Series Design

The second type of reflexive design that we discuss is the time-series design. Although this design is a significant improvement over the one-group pretest-posttest design, it is still in the reflexive category, although some, such as David Nachmias, consider the time- series design to be quasi-experimental (Nachmias 1979, 57). This design is often referred to as a single interrupted time-series design, as the implementation of the program acts to interrupt the prevailing time-series. The impact of the interruption is interpreted as programmatic impact. The simple time-series design is shown in table 12.1.

Time-series designs are useful when preprogram and postprogram measures are available on a number of occasions before and after the implementation of a program. The design thus compares actual postprogram data with projections drawn from the preprogram period.

TABLE 12.1
The Simple Time-Series Design

	Before		After
Group E	$O_1O_2O_3$	X	$O_4O_5O_6$

The design is useful when adequate historical data are available (e.g., numerous data points) and when there appears to be an underlying trend in the data—that is, the data are fairly stable and not subject to wild fluctuations. Crime or accident statistics are good examples of data that are adaptable to the use of the time-series design. Finally, a simple time-series design is considered appropriate when no other rival explanations of the effect can be entertained. In the following article, "A Little Pregnant: The Impact of Rent Control in San Francisco," Edward Goetz uses the simple, interrupted time-series design to estimate whether the introduction of rent control had an impact on the construction of multifamily residences and the subsequent rents as advertised in local newspapers.

Supplementary Readings

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* Cambridge, England: Cambridge University Press, chap. 7.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 9.

READING

A Little Pregnant

The Impact of Rent Control in San Francisco

Edward G. Goetz
University of Minnestoa

The author uses an interrupted time series analysis to evaluate the housing market effects of moderate rent control in San Francisco. Using data on multifamily-housing construction and rent levels from 1960 to 1991, the analysis shows that rent control did not inhibit the development of new multifamily housing. In fact, the rate of multifamily-housing production has increased more rapidly after the program's initiation. Additionally, the analysis shows that rents for advertised units in San Francisco have significantly risen after the implementation of rent control.

AUTHOR'S NOTE: *I would like to thank Roger Herrera of the San Francisco City Planning Department and Mara Sidney of the University of Minnesota for their assistance in the collection of the data.*

RENT CONTROL IS an issue about which landlords, tenants, developers, and policy analysts are often passionate. It has been compared to a nuclear blast in slow motion by one economist (cited in Gilderbloom and Applebaum 1987). Another wrote, "next to bombing, rent control seems in many cases to be the most effective technique so far known for destroying cities" (Lindbeck 1972, 9). Yet, rent control remains popular with tenants' groups and low-income-housing advocates as a means of preserving affordable housing, averting rent gouging, and mediating runaway inflation in housing.

The claims made against rent control are many (see Gilderbloom 1981 for a list). Among the alleged negative effects are depression in the construction of multifamily-housing units (Phillips 1974; Sternlieb 1974; Lett 1976), decline in the maintenance of the housing stock (Kain 1975; Kiefer 1980; Lowry 1970; Rydenfelt 1972; Moorhouse 1972), greater levels of abandonment and demolition (Salins 1980; Lowry, DeSalvo, and Woodfill 1971; Phillips 1974; Sternlieb 1974), stagnation or decline in the local property tax base (Laverty 1976; Sternlieb 1974), and more recently (and more apocalyptically), homelessness (Tucker 1990). At the same time, a growing number of analyses show none of these adverse impacts (see, e.g., Gilderbloom 1981; Bartelt and Lawson 1982; Applebaum et al. 1991).

In this research note, I will use San Francisco as a case example to look at two issues: (1) whether rent control depressed the multifamily home-building market and (2) whether rent control had an impact on advertised rent levels. I use a time series statistical approach to analyze the empirical connection between rent control and these market effects.

Moderate Rent Control and Market Effects

San Francisco adopted rent control in 1979. An extremely heated housing market, low vacancy rates, highly publicized rent increases by large landlords, and a mounting campaign on the part of tenant groups to enact tough rent

control through the initiative process combined to persuade the board of supervisors to adopt their own, less strict form of rent control in June 1979 (see Hartman 1984, 236–45). The program has been in place continuously since then, with only minor modifications in the method of determining allowable rent increases. The program in San Francisco applies to all housing units, including single-family homes and condominiums, except those in buildings with four units or less in which the owner has lived for more than six months, units exempted through a special appeals process, and units constructed after June 1979, when the program was initiated.

The exemption for new construction, along with *vacancy decontrol,* which allows owners to raise rents to market level when apartments become vacant, makes the San Francisco program an example of *moderate rent control* (Gilderbloom 1981). Because of the special treatment of newly constructed units and vacated units, such programs can be expected to have little negative impact on the development of multifamily units and no dampening effect on rents for advertised units. On the other hand, Sternlieb (1981, 145) argued that "the term 'moderate rent control' is in the same dubious league as 'a little inflation' or 'a little pregnancy,'" suggesting that all rent-control programs will have uniformly negative effects on multifamily-housing construction. In the following analysis, I will use data from San Francisco to examine this issue.

Data

Data on the number of multifamily-housing units constructed in San Francisco were compiled from city records (Department of City Planning 1992).[1] Multifamily housing was defined as units in structures with two or more units. Although it is true that not all of these multifamily units are rental or that all single-family units are owner-occupied, construction figures broken down by tenure are not available. The correlation between housing structure and tenure is significant enough to justify the analysis. Figure 1 shows the pattern of multifamily home building in San Francisco from 1960 to 1992.

The figures for 1960 to 1966 are approximated. Figures from the city and county of San Francisco do not give the annual amount for these years, although totals broken down by building size are available for the entire seven-year period. The ratio of single- to multifamily units for the entire seven-year period was used to estimate multifamily-housing production on an annual basis. It was assumed that the ratio of single- to multifamily housing for the seven-year period was matched for each individual year.

The vertical line in the graph marks the adoption of rent control in San Francisco in 1979. My hypothesis is that rent control will have no negative effect on the rate of multifamily-housing construction. The data seem to indicate this is the case. There was a significant drop in production during the late 1960s, a slight recovery during the mid-1970s, another trough through the late 1970s and early 1980s, and a final rebound from 1984 to 1992. The decade of the 1970s saw a heavy migration into the city that put extreme pressure on the local housing market (Hartman 1984). The city also experienced huge levels of downtown commercial development between 1970 and 1985 that attracted new residents (DeLeon 1992). Housing production did not keep pace with population growth during the 1970s, and after the construction market bottomed out completely during the recession of 1981–1982, there was enormous built-up pressure for housing production.

Data on rents in San Francisco were collected for each year from 1960 to 1991. Listings of two-bedroom apartments in the *San Francisco Chronicle* were surveyed, using the

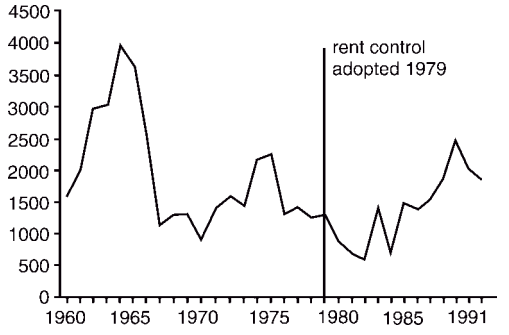


FIGURE 1
Multifamily Housing Units Constructed in San Francisco, 1960–1991

first Sunday in March for each year. Median rents for each year were computed and then adjusted for inflation. It is important to note that these rent figures are for apartments that are vacant. They do not reflect rents paid by tenants remaining in their units. In the analysis, I use the inflation-adjusted median-rent figures. Figure 2 shows the pattern of change for the median rent in San Francisco during this period.

Figure 2 shows a steady increase in rents, even when adjusting for the effects of inflation. The graph seems to indicate a higher upward slope in rents after adoption of rent control in 1979.

Analysis

Multifamily-Unit Construction

Time-series equations were estimated to determine the impact of rent control on multifamily-housing construction and rent levels in San Francisco. The initial equation estimated for each dependent variable was

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + b_3X_{3t} + e_t,$$

where Y_t = number of time-series observations of the dependent variable (either multifamily-housing unit completions [$N = 32$] or median rent [$N = 31$]); X_{1t} = a time counter from 1 to 31; X_{2t} = a dummy variable coded 0 for those years before adoption of rent control and 1 for years following adoption; X_{3t} = a dummy variable scored 0 for years before adoption of rent control and 1, 2, 3 … for years after adoption of the program; and e_t = the error term.

In this simple interrupted time series model, X_1 can be interpreted as the slope or trend in the dependent variable prior to the intervention, X_2 is a measure of the change in the intercept or level of the dependent variable attributable to the intervention, and X_3 is the change in the slope or trend of the dependent variable that was due to the intervention (Lewis-Beck 1986).

Equation 1 is

$$MFU_t = 2751.9 - 80.49X_{1t} - 670.3X_{2t} + 194.9X_{3t} + e_t$$
$$(8.7^*)\ (-2.9^*)\ (-1.4)\ (3.4^*)$$
$$R^2 = .42\ N = 32\ \text{D–W} = .86$$

where MFU_t indicates annual multifamily-housing-unit construction levels for the years 1960 to 1991; the figures in parentheses indicate t-ratios; * indicates t-ratio significant at the $p < .05$ level; D–W is the Durbin Watson statistic.

The data show a statistically significant downward trend in multifamily-housing construction of 80 units per year in San Francisco prior to the adoption of rent control ($-80.49X_{1t}$). The coefficient for the change in in-

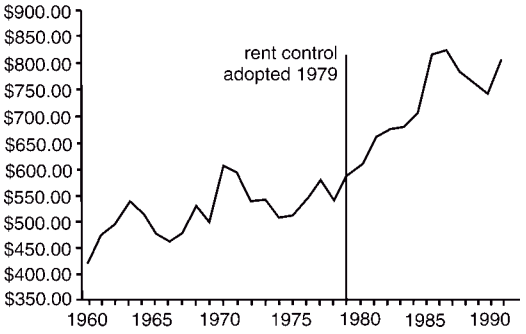


FIGURE 2
Median Rents, San Francisco, 1960–1990

tercept (670.3) is statistically insignificant. Finally, the coefficient for X_{3t} (194.9) indicates a statistically significant change in the trend of multifamily-housing completions in San Francisco. The post-rent-control period saw a greater upward trend in production than did the period before adoption of the program. This effect is clearly related to the relatively high levels of production in the city during the early 1960s. Yet, the data conclusively show that moderate rent control did not suppress multifamily-housing production.[2]

Equation 1 was rerun with two additional variables. The number of housing units completed in the entire western region of the United States was included as a measure of the strength of the construction industry. This measure was statistically insignificant in the equation and significantly reduced the efficiency of the overall equation, so it was dropped. The same results were produced when the average prime lending rate for each year was added to the equation to account for the state of the lending market. Given the lack of explanatory power of these variables, the simple interrupted time series was used.

Sternlieb (1981) suggested that the production effects of rent control are likely to be lagged a number of years, taking into account the long pipeline for most housing developments. Equation 1 was rerun with variables X_2 and X_3 lagged up to three years. No change in the results occurred.

The Durbin-Watson statistic indicates a problem with autocorrelation. Adjustments to the data were made to remove the autocorrelation of the error terms (Lewis-Beck 1986) and the equation was rerun.[3] The adjusted equation confirms the earlier results.

Equation 2 is

$$MFU_t = 1584.5 - 140.3X_{1t} - 208.7X_{2t} + 245.0X_{3t} + e_t$$
$$(5.2^*)\ (-2.7^*)\ (-.37)\ (2.5^*)$$
$$R^2 = .26\ N = 31$$

The downward trend in production prior to rent control remains significant (a decline of 140 units per year). The reduction in the mean number of units produced after the program was initiated (208.7) is statistically insignificant. Finally, the change in the trend in production is positive and statistically significant, the coefficient showing an increase in the rate of construction of 245 units after rent control was initiated. These findings demonstrate quite clearly that moderate rent control did not have a negative effect on the production of multifamily units in San Francisco.

Rent Levels

Equation 3 is

$$MDR_t = 470.7 + 4.76X_{1t} - 35.7X_{2t} + 13.94X_{3t} + e_t$$
$$(24.3^*)\ (2.80^*)\ (1.16)\ (3.7^*)$$
$$R^2 = .89\ N = 31\ \text{D–W} = 1.12$$

where MDR_t indicates the median advertised rent for two-bedroom apartments as listed on the first Sunday in March for the years 1960 to 1991; the figures in parentheses indicate t-ratios; * indicates t-ratio significant at the $p < .05$ level; and D–W is the Durbin Watson statistic.

This equation indicates a steady and significant upward trend in rents during the pre-rent-control period (4.76_{1t}), followed by a significantly accelerated rate of increase after adoption of rent control. The coefficient for X_{2t} is statistically insignificant, suggesting virtually no change in the intercept at the intervention point.

The Durbin-Watson statistic indicates trouble with autocorrelation, however. The model was rerun in the manner previously described. The resulting equation is similar to Equation 3 in all respects except that the coefficient for X_{1t} falls below the level for statistical significance. The data nevertheless suggest that there has been a significant change in the rate of rent increases ($15.82X_{3t}$, $p < .05$) since the adoption of moderate rent control in San Francisco.

Equation 4 is

$$MDR_t = 288.2 + 3.03X_{1t} + 40.6X_{2t} + 15.82X_{3t} + e_t$$
$$(14.3^*)\ (1.09)\ (1.11)\ (2.90^*)$$
$$R^2 = .77\ N = 30$$

A Little Pregnant in San Francisco

The findings suggest that it is, indeed, possible to be a little pregnant with rent control. That is, the particular form that rent-control programs take can have consequence for market outcomes. The evidence indicates that moderate rent control in San Francisco did not have an inhibiting effect on the production of multifamily units in the city as critics of rent control would predict. Vacancy decontrol has produced its own market effect as well. The data presented here indicate that rent levels for advertised units increased over the entire period from 1960 to 1991, even when adjusting for inflation. Furthermore, the increase in rent levels was significantly greater after adoption of the rent ordinance. The moderate form of rent control operated in San Francisco, therefore, does not protect a newcomer to the housing market or protect those who choose, or are forced by any circumstance, to move. For rent-control advocates, this clearly mitigates the positive impact the program presumably has on maintaining lower rents in occupied units. The vacancy-decontrol clause in the San Francisco program has clearly provided landlords and management companies the opportunity to inflate rents on unoccupied units. Rent-control advocates and opponents of vacancy decontrol suggest that these circumstances lead to an incentive for landlords to evict long-term residents to bring rents up to market levels. Though this analysis provides no data on this question, it does suggest the logic is sound.

Notes

1. This document contained figures dating to 1982. Previous figures were obtained from earlier such documents. Figures back to 1960 were obtained from Roger Herrera, planner, city of San Francisco.
2. The very high levels of multifamily-housing production in the early 1960s are important to the regression findings. These years are the estimated figures described earlier. Multifamily-housing production for each year between 1960 and 1966 was set at 84% of the total housing production, matching the percentage for the entire seven-year period. In most years, multifamily-housing production was an even greater proportion of total production. In fact, multifamily housing as a percentage of all housing built in San Francisco has exceeded 84% in 8 of the 12 post-rent-control years. It is highly unlikely, therefore, that the method used to estimate annual multifamily-housing production from 1960 to 1966 resulted in overinflated figures.
3. The autocorrelation coefficient, *p,* was estimated for the equation and then each variable in the equation was adjusted by the following equation, $X = X_t - pX_{t-1}$ (Lewis-Beck 1986).

References

Applebaum, R. P., M. Dolny, P. Dreier, and J. I. Gilderbloom. 1991. Scapegoating rent control: Masking the causes of homelessness. *Journal of the American Planning Association* 57 (2): 153–64.

Bartelt, D., and R. Lawson. 1982. Rent control and abandonment: A second look at the evidence. *Journal of Urban Affairs* 4 (4): 49–64.

DeLeon, R.E. 1992. *Left coast city: Progressive politics in San Francisco, 1975–1991*. Lawrence: University Press of Kansas.

Department of City Planning. 1992. *Housing information series: Changes in the housing inventory, 1991*. San Francisco, CA: Author.

Gilderbloom, J. I. 1981. Moderate rent control: Its impact on the quality and quantity of the housing stock. *Urban Affairs Quarterly* 17 (2): 123–42.

Gilderbloom, J. I., and R. P. Applebaum. 1987.

Rethinking rental housing. Philadelphia, PA: Temple Univ. Press.

Hartman, C. 1984. *The transformation of San Francisco.* Totowa, NJ: Rowman & Allanheld.

Kain, J. 1975. Testimony on rent control, Local Affairs Committee. (21 March). Boston, MA: State House.

Kiefer, D. 1980. Housing deterioration, housing codes and rent control. *Urban Studies* 17:53–62.

Laverty, C. 1976. *Effect on property valuation and assessed valuations for ad valorem taxation.* Cambridge, MA: Tax Assessor's Office.

Lett, M. 1976. *Rent control: Concepts, realities, and mechanisms.* New Brunswick, NJ: Center for Urban Policy Research.

Lewis-Beck, M. S. 1986. Interrupted time series. In *New tools for social scientists,* edited by W. D. Berry and M. S. Lewis-Beck, 209–40. Beverly Hills, CA: Sage.

Lindbeck, A. 1972. *The political economy of the New Left.* New York: Harper & Row.

Lowry, I., ed. 1970. *Rental housing in New York City: Confronting the crisis.* Vol. 1. New York: Rand Corporation.

Lowry, I., J. DeSalvo, and B. Woodfill: 1971. *Rental housing in New York City: The demand for shelter.* Vol. 2. New York: Rand Corporation.

Moorhouse, J. C. 1972. Optimal housing maintenance under rent control. *Southern Economic Journal* 39:93–106.

Phillips, D. 1974. *Analysis and impact of the rent control program in Lynn.* MA: Mayor's Office.

Rydenfelt, S. 1972. Rent control thirty years on. In *Verdict on rent control,* edited by A. Seldon. Worthing, Sussex, England: Cormorant.

Salins, P. 1980. *The ecology of housing destruction: Economic effects of public intervention in the housing market.* New York: New York Univ. Press.

Sternlieb, G. 1974. *The realities of rent control in the greater Boston area.* New Brunswick, NJ: Center for Urban Policy Research, Rutgers University.

____. 1981. Comment. *Urban Affairs Quarterly* 17 (2): 143–45.

Tucker, W. 1990. *The excluded Americans: Homelessness and housing policies.* Washington, DC: Regnery Gateway in cooperation with the Cato Institute.

Edward Goetz is an associate professor in the housing program at the University of Minnesota. He has written in the areas of local housing and economic development policy. He is the author of *Shelter Burden: Local Politics and Progressive Housing Policy* (Temple University Press, 1993) and coeditor of *The New Localism: Comparative Urban Politics in a Global Era* (with Susan Clarke; Sage, 1993). His work has appeared in the *Journal of Urban Affairs*, *Political Research Quarterly, Economic Development Quarterly*, and the *International Journal and Regional Research.*

Explanation and Critique

In this study Goetz uses the interrupted time-series design to estimate the impact of rent control in the case of San Francisco. Earlier studies had demonstrated mixed results, and Goetz was interested in determining whether the San Francisco model of rent control, "mild" (as in being a little pregnant) as opposed to "full blown" rent control, supported earlier findings or was a case unto itself. The "interruption" is the adoption of rent control as indicated in figures 1 and 2 of the article.

The first dependent variable examined is the number of multifamily units constructed each year for the 32 years between 1960 and 1991. The data for the number of multifamily units constructed between 1960 and 1966 are estimations that Goetz infers are conservative based on actual figures in the post–rent control period (see footnote 2 in article). The independent variables include two time counters, one that measures the passage of time throughout the 32 years from 0 to 31 and one that reflects the time subsequent to the adoption of rent control. The other independent variable, referred to as a "dummy" variable (chapter 3), takes on the value of "0" during the pre–rent control period and "1" for the post–rent control period.

On the basis of the time-series equations, Goetz concludes that the number of multi-

family units constructed went down even prior to the adoption of rent control (b = –0.49) and increased after rent control was in place. Although this finding is verified visually (see below), that is not the literal interpretation of the coefficient. As indicated in chapter 3, the coefficient is interpreted that as the value of X increases one unit, the value of Y changes in the direction of the sign by "b-units" of Y. Therefore, the number of units constructed goes down generally as time passes, even though it "appears" to increase after the adoption of rent control. Separate equations pooling the pre- and post-intervention years would demonstrate what we see visually. The postintervention effect is adequately demonstrated by the impact of the final count variable, which measures the postintervention impact. We do agree, though, that because the coefficient for the dummy variable indicating the change to rent control was insignificant, rent control itself had no impact. These findings remain consistent even when controlling for autocorrelation (equation 2 in article).

Let us assess these findings visually by looking at Goetz's figure 1. It is clear that there was a drop-off of constructed units after the high productivity of the 1990–1996 period. Coincidentally, this productive period was the one for which the number was estimated by Goetz. Although we do not have any problem with how he arrived at the estimates, we do have some concerns regarding how appropriate the application of a linear statistic is with data that do not exhibit a linear trend. Regression-based estimates are adversely affected by wide fluctuations as shown in the pre–rent control period. Oftentimes these fluctuations are "smoothed" out by using three-year moving averages, for example. The series for the post–rent control period does not exhibit such fluctuations and appears to be indicative of a trend.

The third and fourth equations estimate the impact of rent control on the median advertised rents for two-bedroom apartments. Recall that the effect of mild rent control is that newly constructed dwellings and new residents of heretofore rent-controlled units were not covered by the controls. Goetz finds that although rents were increasing before the institution of controls, the increases were significantly accelerated in the post–rent control period. This trend is clearly visible from figure 2 in the article. Moreover, since there are not extreme cases as in the first two equations, the interpretation of the coefficients is reasonable.

In the introduction to this chapter, we pointed out that the use of interrupted time-series designs is appropriate when no other rival explanations of the effect can be demonstrated. Here, Goetz acknowledges that other factors might be operating; however, it may be useful to point some of them out clearly. First, the decrease in the number of multifamily housing units in the pre–rent control era is dominated by the extreme cases of the period before 1966. If the time-series began in 1966, there actually appears to be an increase in the trend concomitant with the increased migration to San Francisco to which Goetz alludes. The decrease begins during a period of high inflation in the second half of the 1970s and bottoms out during the height of the recession of the early 1980s. These are rival hypotheses that seem to make sense.

In terms of the median rent increases during the post–rent control era, there are some items to be reconsidered. Recall that the rents that are not covered by the controls are those of new units and for new residents in previously rent-controlled dwellings. It makes sense that landlords would want to increase rents in tighter markets (which San Francisco was experiencing), and also to make up lost earnings due to rent control by raising the rents of new residents. What is not measured here is the overall impact on renters—both new and those under rent-control restrictions. Certainly, rent control had a

positive impact on those who continued to live in rent-controlled units. Therefore, there may be more of an impact than this evaluation suggests, albeit for only a subset of the renting population.

Once again, use of the single interrupted time-series design is justifiable in cases where no rival hypotheses could be entertained to produce the observed impact. The classic study in this regard was Robert Albritton's (1978) study of the impact of a permanent federal program to eradicate measles in the school-age population in 1966. Finding that the program had a substantial impact that continued over time, Albritton sought rival explanations for the impact. Because the number of school children had not declined and the impact remained even when statistically controlling for the private use of measles vaccine, he concluded that the impact of the program was real. Even though the federal program required improved measles reporting procedures, thus increasing reported incidence, the impact remained. Albritton pointed out that this attenuated his estimates of the impact. Therefore, in reality, the impact was even greater than that estimated. No rival hypothesis could possibly explain the impact. Therefore, when considering the results of this and any type of evaluation design, question whether any rival hypotheses could be introduced to falsify the findings.

References

Albritton, Robert B. 1978."Cost-Benefits of Measles Eradication: Effects of a Federal Intervention." *Policy Analysis* 4: 1–22.

Nachmias, David. 1979. *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's, 57.

PART V

Cost-Benefit and Cost-Effectiveness Analysis

The preceding parts of this book discussed the various methodologies for assessing the impact of a program or treatment. These experimental, quasi-experimental, and reflexive designs answer the fundamental question "Do these programs work as intended?" Cost-benefit and cost-effectiveness analyses move the process one step further by asking the questions "Is this program worth it?" and "How much do constituents receive as a result of the program?" Cost-benefit and cost-effectiveness analyses are thus not evaluation designs but are processes that use evaluation designs to evaluate outcomes.

Before examining cost-benefit and cost-effectiveness cases, it is first necessary to distinguish between the two. Each has the common purpose of examining the resources required (cost) in relation to the desired outcome (efficiency and/or benefits). The primary distinction between these two techniques is in how the benefits are valued. In cost-benefit analysis, all potential benefits are described in numerical terms (usually dollars). In cost-effectiveness analysis, however, benefits gained are usually measured in nonmonetary units (Pike and Piercy 1990, 375).

The designs described here and earlier can be (and are) used in tandem to evaluate the impact of programs. It makes little sense to estimate the financial benefits of a program that had no impact; the benefits did not exist. Likewise, it is important for decision makers to know which of competing programs work or have the greatest impact. If the costs to operate the best program are prohibitive, however, it would be more politically and fiscally responsible to offer a program with slightly less estimated impact that a political unit could afford than to offer no program at all.

Cost-benefit and cost-effectiveness measures are relative; it is sometimes difficult to interpret them outside of the evaluation site. One way to evaluate the measures, however, is to compare them with some established standard. Service-level standards, or goals, are often published by professional associations (e.g., the International City/County Management Association, the National Education Association, the American Library Association). For instance, the American Library Association publishes circulation and holdings goals by type of library and population characteristics of the catchment area. Peter Rossi and Howard Freeman call such goals "generic controls" (1985, 279–82). Care should be taken in determining the difference between a well-established norm and professional ag-

grandizement. In the absence of established norms, researchers may wish to compare their results with those from similar studies to know if they are "in the ballpark." Finally, they can use others' results to predict outcomes in their own evaluations.

The applications of cost-benefit and cost-effectiveness analysis are not confined to overtly political realms such as city councils and state legislatures. These tools are useful in such nonprofit realms as hospital and psychiatric settings. They can even be useful in the private sector where maximizing corporate profit while maintaining product and service integrity is important. The following two chapters provide examples of these two methods, which differ in assumption, problems, and utility. As with other designs, the appropriateness of either method depends on the questions the evaluation client needs answered and the constraints of the available data.

References

Pike, Connee L. , and Fred P. Piercy. 1990. "Cost Effectiveness Research in Family Therapy." *Journal of Marital and Family Therapy* 16 (October): 375–88.

Rossi, Peter H., and Howard E. Freeman. 1985. *Evaluation: A Systematic Approach.* 3d ed. Beverly Hills, Calif.: Sage.

Supplementary Reading

Rossi, Peter H., Howard E. Freeman, and Mark W. Lipsey. 1999. *Evaluation: A Systematic Approach.* 6th ed. Thousand Oaks, Calif.: Sage, chap. 11.

CHAPTER 13

Cost-Benefit Analysis

Cost-benefit analysis is an intuitively easy process to conceptualize and appreciate. One gathers all the costs of providing a good or service and weighs those costs against the dollar value of all the subsequent benefits provided by the good or service. If the benefits outweigh the costs, the good or service should be continued; if the costs of providing the service exceed the benefits obtained, the service should be terminated. All decision makers can relate to break-even points. Private-sector managers wish benefits (i.e., profit) to far exceed costs, whereas public-sector managers hope at least to break even.

This is not to say that public-sector managers never attempt to economize or provide efficient services. The nature of governmental involvement has been historically to provide "collective goods"—those for which exclusion is infeasible and joint consumption possible. The private market has no incentive to provide collective goods through normal market structures because they cannot make a profit by doing so. Therefore, government must get involved to provide these services (for a discussion, see Savas 2000).

However simplistic this analysis seems conceptually, the actual adding up of costs and benefits is anything *but* simple. First you must list all the costs and all the benefits associated with a program. Do you include only the costs to a *particular* agency when more than one agency supports the program? How do you apportion overhead costs? What about the costs of in-kind contributions? And as for benefits, do you consider short-term benefits or long-term benefits, or both? How do you divide programmatic benefits when a recipient participates in several related programs (e.g., general assistance, job training, child nutrition, and subsidized day care)? Once these costs and benefits are delineated, how do you attach a dollar amount to the numerator or denominator? What is the dollar value of one nonhungry child or the deterrent effect of keeping one child out of the juvenile justice system?

These are not easy questions. Yet these are the very questions confronted by any researcher involved in cost-benefit analysis. A fundamental assumption of this design is that the researcher is able to assign dollar values to both the numerator and denominator. Typically, costs are easier to assign than benefits. Some impacts of programs occur instantaneously, some over time, and some

extinguish quickly, whereas others deteriorate over time. When determining benefits, the researcher must be aware of the expected impact and judge benefits accordingly. This is not often an easy task.

When you can make such assignments, cost-benefit analysis becomes a very strong tool to assist decision making—especially when you can make comparisons among competing programs or the ways service is delivered. However, some major ethical considerations come into play when interpreting cost-benefit ratios. Should interpretation and utilization of the ratios be taken literally? Should you abolish a program merely on the basis of the cost-benefit result? If so, you can understand why researchers must be extremely careful in parceling out costs and benefits. Moreover, what role must government play in parceling basic social services? If there were positive benefit-to-cost ratios in all social programs, would not the marketplace step in to provide such services? Would strict attention to cost-benefit ratios nail the lid on the coffin of the delivery of social services?

"The Costs and Benefits of Title XX and Title XIX Family Planning Services in Texas," by David Malitz, is an excellent example of a careful, well-conceived cost-benefit analysis. Notice that Malitz is in a precarious position with regard to estimating benefits. The benefits are actually negative costs (expenditures) not incurred as a result of the estimated impact of the program. He cannot *directly* measure the benefits because the benefits are the consequences of averted pregnancies!

Malitz's article refers to the Title XIX and Title XX amendments to the Social Security Act. Title XIX refers primarily to Medicaid services (health care) for low-income people. Title XX provides health care and social service assistance (subsidized day care, maternal health and family planning, home health care) for low-income individuals—even those above "the poverty line."

READING

The Costs and Benefits of Title XX and Title XIX Family Planning Services in Texas

David Malitz
Consultant in Statistical Analysis and Evaluation

The Texas Department of Human Resources provided family planning services to more than a quarter of a million women in the fiscal year 1981 through its Title XX and XIX programs. In order to evaluate the impact of these services, estimates were made of the number of pregnancies, births, abortions, and miscarriages averted by the program. These estimates were then used to calculate the costs and benefits attributable to the program's services.

AUTHOR'S NOTE: *This research was funded under contract with the Texas Department of Human Resources, although the opinions and conclusions expressed are those of the author. The results of this study were reported at the 1982 meeting of the American Public Health Association, Montreal, Canada. The author wishes to acknowledge the invaluable assistance of Beth Weber, Director of Family Self Support Services, Texas Department of Human Resources; Dick Casper, Vice President, James Bowman Associates; and Peggy Romberg, Executive Director, Texas Family Planning Asso-*

ciation. A special debt for assistance in the design of this study is due Joy Dryfoos.

PUBLICLY FUNDED FAMILY PLANNING services in Texas are administered by two state agencies: the Texas Department of Human Resources (TDHR) and the Texas Department of Health. In fiscal year (FY) 1981, TDHR spent about $22 million in Title XX and XIX funds to provide services to more than a quarter of a million patients, representing about 79% of all women receiving family planning services from organized providers in Texas (Alan Guttmacher Institute, 1982). In 1982, TDHR commissioned a study to evaluate the impact of this large program (Malitz et al., 1981). One component of this evaluation was a cost-benefit analysis, the results of which will be reported in this article.

The costs of the programs were relatively easy to ascertain and consisted of expenditures for the programs during FY 1981. The benefits, however, were considerably more difficult to measure. For this study, consideration was limited to direct, first-year cost savings to TDHR which arose through the prevention of unwanted births. These cost savings consisted of expenditures which would have been incurred within the first year following birth in TDHR's Aid to Families of Dependent Children (AFDC), Food Stamp, and Medicaid programs.

It is important to note some of the possible benefits that were not considered in the analysis. These include prevention of long-term welfare dependency and adverse social and health consequences due to births by adolescent mothers, as well as the health benefits which may result from health screening by family planning providers. In addition, although estimates were made of the number of abortions and miscarriages averted, no attempt was made to quantify these benefits in financial terms.

The methodology is patterned after Chamie and Henshaw's (1982) analysis of governmental expenditures for the national family planning program. It requires estimates of the number of births averted by the program, the proportion of such births which lead to the receipt of public assistance, and the cost of this assistance.

The number of births averted was estimated by examining patterns of contraceptive use among family planning patients (Forrest et al., 1981). Data were collected regarding the use of various contraceptive methods (including no method) by family planning patients before they entered the program (the premethod) and after their last visit to the program (the postmethod). Use-effectiveness rates for each of these contraceptive methods were applied to the preprogram distribution to estimate the number of women who would become pregnant within a year if they continued using the contraceptive methods observed just prior to their first program visit. Similarly, estimates were made of the number of women who would become pregnant within a year using the postmethods of contraception. The difference between these two estimates represents the number of pregnancies averted by the program.

Using available data regarding the outcomes of unwanted pregnancies in the United States (Dryfoos, 1982), calculations were made of the number of births, abortions, and miscarriages which would have resulted from the pregnancies averted. Finally, public assistance costs were estimated for the births averted and compared with the costs of the program to derive the cost-benefit ratio.

Methodology

As noted above, the number of averted pregnancies was estimated from data on pre and postprogram use of contraceptive methods by family planning patients. To gather these data, a survey was conducted of a randomly

selected sample of case records maintained by the 78 providers of Title XX family planning services in Texas. Technical problems made it impossible to draw a separate sample of patients from the Title XIX program in the time available. Although estimates of the costs and benefits of the Title XIX program will be made below, it should be understood that these are based upon contraceptive use data gathered from the Title XX program.

The sampling frame for the Title XX survey consisted of billing records for all 227,253 Title XX patients served in FY 1981 (September, 1980 through August, 1981). Patients were stratified by provider and by age: adolescents (19 years and younger) and adults (20 years and over). Patient records within each age group were sampled from the 78 providers in proportion to the number of patients served by each provider.

Based upon this sampling plan and upon an expected rate of return, the sample size was determined and a sample drawn from the billing records. The final sample consisted of 1606 adolescents (about 2.5% of the Title XX adolescent population) and 1605 adults (about 1.0% of the adult population). Given an expected response rate of about 70%, a 95% confidence interval of about ±18 pregnancies per 1,000 women would be achieved.

Survey forms were printed which identified each patient sampled, and which solicited information from the providers to determine the date when the patient first visited the agency, the date of the last visit in FY 1981, and the method(s) of contraception used by the patient before her first visit and after her last visit in FY 1981. Agency staff were assured that patient confidentiality would be maintained, and the survey form was designed in such a manner that patient names could be detached so they could not be identified with patient information.

Survey forms were returned by 65 of the 78 providers. Among those returning forms, the completion rate was quite high. Overall, 1252 complete and usable survey forms for adolescents (78.0%) were returned and 1283 (79.9%) complete forms for adults were obtained.

Results

Pregnancies, Births, Abortions and Miscarriages Averted

Table 1 presents the survey results for adolescents and adults. On the left-hand side of Table 1, the various contraceptive methods are listed along with the number of pregnancies expected within a year's time among 1,000 sexually active women using each method (Forrest et al., 1981). It can be seen that the use-effectiveness rates range from a low of zero for sterilization to a high of from 490 to 640 pregnancies per 1,000 women using no method of contraception.

Following the use-effectiveness rates are data from the sample of case records for adolescent Title XX patients. Shown first is the percent distribution of preprogram contraceptive use, followed by the postprogram distribution. Similar pre and postdistributions are shown for adult patients. When multiple methods, either at the pre or postlevel were indicated by agency staff, the patient's most effective method was chosen as the primary method for purposes of data analysis.

It can be seen from Table 1 that, among adolescents, the most common premethod was "none," indicated for 68.7% of the 1252 patients from whom data were available. This percentage is considerably higher than the 50.4% reported by Forrest et al. (1981) for a 1975 national sample of adolescents visiting family planning clinics. However, data collected by the Alan Guttmacher Institute (AGI) (AGI, 1981, AGI, 1982) show a fairly steep rise between 1972 and 1980 in the proportion of family planning patients using no method of birth control before entering clinics. When patients of all ages are combined, the percentage rose from 31% in 1976 to 55%

in 1979. Among adolescents, data for 1980 show 70% using no premethod in a national sample. Thus, the percentage for Texas adolescents is slightly lower than the national average for 1980, the latest year for which data were available.

Examination of the distribution of postprogram usage among adolescents in Table 1 reveals a dramatic shift from preprogram usage. Whereas 68.7% used no preprogram method, only 14.4% used no method after the program, a drop of more than 50 percentage points. Thus, by the time of the last agency visit in FY 1981, the majority of the adolescent patients (85.6%) were using one of the more effective contraceptive methods. The percentage using no method after the program (14.4%) is slightly higher than the 12% figure reported among adolescents in 1980 (Alan Guttmacher Institute, 1982). It is also higher than the 8.1% reported in 1975 (Forrest et al., 1981).

Among adolescents using an effective method after the program, the vast majority left the agency using the pill. This method was used by 74.7% of the sample, slightly more than the 69% reported by AGI for 1979. The condom was the next most common method, and was used by 6.5% of the Texas sample after leaving the program.

It should be noted that among the 14.4% of the adolescent patients who left the agency using no method, most (88.3%) also entered the agency using no method on their first

TABLE 1

Contraceptive Method Use Patterns and Expected Number of Pregnancies Averted Among Title XX Patients Served

Contraceptive Method	*Expected Number of Annual Pregnancies per 1,000 Women*	***Percent Using Each Method Before First Visit and After Last Visit to Program***			
		Adolescents		*Adults*	
		Pre	*Post*	*Pre*	*Post*
Pill	25	21.4	74.7	42.9	62.9
IUD	71	0.8	1.8	3.7	6.7
Diaphragm	172	0.3	1.1	0.6	1.9
Foams, creams, jellies	184	1.9	0.8	2.3	2.4
Rhythm	250	0.2	0.0	0.2	0.2
Sterilization	0	0.0	0.1	1.7	6.9
Condom	123	5.9	6.5	4.8	7.9
Other	189	0.8	0.6	1.1	1.2
None	490 to 640	68.7	14.4	42.6	9.9
Total	—	100.0	100.0	100.0	100.0

Expected Number of Annual Pregnancies per 1,000 Patients	***Adolescents***		***Adults***	
	Pre	*Post*	*Pre*	*Post*
Low estimate	357	103	237	90
High estimate	460	124	301	104
Midpoint estimate	408.5	113.5	269.0	97.0
Pregnancies averted	295		172	

Source: Forrest, Hermalin, and Henshaw (1981) for expected number of annual pregnancies per 1,000 women.

Note: Based upon surveys of contraceptive failures among married women using the methods listed. The estimates of pregnancy rates among users of no method are based upon surveys of unmarried, sexually active adolescents.

visit (not shown in Table 1). To a somewhat lesser degree, this was true for adults, for whom 66.4% of those leaving with no method also entered with no method.

Patients who entered and exited with no method would appear to represent program failures because ideally all patients should leave the agency using an effective contraceptive method. However, inquiries made to clinic personnel indicated that many of these patients came to the agencies for pregnancy tests. Some of these patients were found to be pregnant and therefore could not be given contraceptives. Unfortunately, it was not possible, with the data at hand, to determine how many of those exiting with no method were pregnant and how many refused or were not given contraceptives for other reasons. However, AGI data (1982) indicate that among patients who had no contraceptive method at their last clinic visit, about half were pregnant, while the remainder were seeking pregnancy, came to the clinic for infertility services, did not receive a method for medical reasons, or, for nearly one-third, failed to receive a method for some other reason.

At the bottom of Table 1 is the expected number of pregnancies per 1,000 patients based upon the pre and postmethod distributions. These estimates were derived by multiplying the number of pregnancies among 1,000 users of each method by the percent using each method, and summing the resulting products across methods. Given that a range is indicated for the pregnancy rate among users of no contraceptive method, low and high estimates were calculated in each column.

It can be seen that among adolescent users of the preprogram methods, between 357 and 460 pregnancies per 1,000 patients would be expected (the midpoint of this range is 408.5). Nearly all of these pregnancies would be attributable to patients using no method, who accounted for 337 to 440 of these pregnancies.

Among adolescent users of the postprogram methods, fewer pregnancies (103 to 124 per 1,000 patients with a midpoint of 113.5) would be expected. This is due primarily to the shift from no method to some effective method.

To simplify future calculations, the midpoints of each range were used as the best estimate of the expected number of pregnancies. The difference between the pre and postmidpoints represents the number of pregnancies averted. Thus, approximately 295 pregnancies per 1,000 adolescent patients were averted by the Title XX program. This is higher than the 240 pregnancies averted among U.S. adolescents in 1975 reported by Forrest et al. (1981). It also compares favorably with the 1979 figure of 282 pregnancies averted per 1,000 adolescents aged 15–19 reported by AGI (1981).

Turning to the adult data presented in the second part of Table 1, it can be seen that the percentage of patients using no preprogram method (42.6%) is lower than was found among adolescents, but is still substantial. Only 9.9% of the adults used no postprogram method, the remainder using the pill most frequently, followed by condoms, sterilizations, and the IUD. Forrest et al. do not present data for adults, but AGI (1982) data from 1980 show that among women aged 20 and over, 39% used no preprogram method and 12% used no postprogram method. Thus, the Texas figures show slightly more adult women using no preprogram method and slightly fewer using no postprogram method than the national figures for this age group.

In terms of pregnancies averted, the estimate in Table 1 of 172 pregnancies per 1,000 patients is higher than the estimates reported by AGI for 1970–1972, the latest years for which adult estimates were made. These national estimates are 141 pregnancies averted per 1,000 women aged 20–29 and 52 pregnancies averted among women 30 and over.

Table 2 uses the Title XX estimates of

pregnancies averted to make additional estimates of the number of births, abortions, and miscarriages averted by the Title XX and XIX programs. At the top of the table is shown the number of pregnancies averted, computed by multiplying the program counts by the midpoint of the pregnancies-averted rates from Table 1. It can be seen that almost 19,000 adolescent pregnancies and more than 28,000 adult pregnancies were averted, for a total of nearly 47,000 pregnancies averted by the Title XX program in FY 1981.

At the bottom of Table 1 are listed the percent distribution of births, abortions, and miscarriages for unintended pregnancies to women of low and marginal income (Dryfoos, 1982). This distribution is based on the assumption that pregnancies that occur to women who are visiting family planning clinics would be unintended. The figures are also appropriate because they were computed for women of low and marginal income in 1979, income levels similar to those among Title XX women in 1981.

These percentages were multiplied by the number of pregnancies averted to estimate the number of births, abortions, and miscarriages averted by the program. At the top of Table 1 it can be seen that nearly 7,000 adolescent births and more than 15,000 adult births were averted by the program, for a total of about 22,000 births averted. About 18,000 abortions were averted, approximately half by adolescents and half by adults. About 6,000 miscarriages, were also averted by the program.

The final column at the top of Table 2 expresses the births averted as a rate per 1,000 patients served. It can be seen that the program averted about one birth for every ten patients served.

The second section of Table 2 repeats the calculations for the Title XIX program based upon the Title XX pregnancies-averted rates. It can be seen that about 9,000 pregnancies

TABLE 2

Estimated Number of Pregnancies, Births, Abortions, and Miscarriages Averted by the Title XX and Title XIX Family Planning Programs

Age Group	*Number in Program*	*Pregs. Averted*	***Total Births, Abortions, and Miscarriages Averted***			*Births Averted per 1,000 Patients*
			Births	*Aborts.*	*Miscar.*	
Title XX Program						
Adolescents	63,176	18,637	6,784	9,542	2,311	107
Adults	164,077	28,221	15,614	8,592	4,015	95
Total	227,253	46,858	22,398	18,134	6,326	99
Title XIX Program						
Adolescents	11,653	3,438	1,251	1,760	426	107
Adults	32,823	5,646	3,122	1,716	802	95
Total	44,476	9,084	4,342	3,516	1,226	99

U.S. Pregnancy Outcomes Age Group	***Percent Distribution of Unintended Pregnancies by Outcome***		
	Births	*Aborts.*	*Miscar.*
Adolescents	36.4	51.2	12.4
Adults	55.3	30.4	14.2
Total	47.8	38.7	13.5

Source: Dryfoos (1982) for distribution of unintended pregnancies by outcome.
Note: Pregnancies averted are calculated by multiplying program counts by Title XX rates in Table 1. Rows may not sum to totals due to rounding.

were averted by the Title XIX program, and that these pregnancies would have resulted in about 4,300 births, 3,500 abortions, and 1,200 miscarriages.

Cost-Benefit Analysis of the Title XX Program

Direct first-year costs to TDHR are incurred by births to Title XX women in two ways. First, a certain percentage of Title XX women will become Title XIX eligible due to the birth of a child. For these women, TDHR will pay, through Medicaid, certain delivery and birth-related expenses. In addition, the department will pay AFDC and Medicaid benefits for each woman and her child for one year. Finally, because of the birth of a child, some women may be entitled to additional food stamp allotments. Births in this category will be termed "AFDC births."

TDHR can incur direct first-year expenses for non-AFDC births as well. Although at least some of these women will depend upon city and county funds to pay delivery and birth-related expenses, these costs will not be paid by TDHR and will therefore not be included in this analysis. However, some of the Title XX family planning patients will be food stamp recipients, and may therefore be entitled to an increased allotment due to an increase in family size.

Let *a* equal the costs associated with an AFDC birth, *b* equal the cost associated with a non-AFDC birth, *p* equal the proportion of Title XX women who would go onto AFDC due to a birth, and *1-p* equal the proportion who would not go onto AFDC.

Given these parameters, the total average cost per Title XX birth, *t*, will be:

$$t = (p \times a) + [(1-p) \times b].$$

In other words, the two costs are weighted by the proportion of women incurring each cost, and summed to produce a total average cost per Title XX birth. Looked at another way, this figure represents the average cost savings per Title XX birth averted. This figure can then be multiplied by the probability of averting a birth for each Title XX patient. The resulting figure will be the average cost savings per patient, which can be compared with the average cost of serving a patient to derive the cost-benefit ratio.

The following discussion will first outline the costs that can be incurred by TDHR due to births to Title XX women who become AFDC recipients. Following this, there will be a discussion of the increased food stamp costs to the remaining Title XX recipients. Estimates derived for these two groups will then be combined in the manner described above and will be used to estimate the overall cost savings brought about by averting births to Title XX women.

The first step involves computing the annual Title XIX cost in FY 1981 that would be expected for each case that went onto AFDC due to the birth of a child. This includes $1304, the average Title XIX payment per delivery; and $807, the average cost for inpatient hospital care for infants aged zero to one year of age (due primarily to premature births and birth defects); for a total Title XIX cost of $2,111 associated with the birth of a child.

The next figure needed is the cost of welfare maintenance for a mother and one child during FY 1981. Although some of the cases which go onto AFDC due to the birth of a child will have family sizes larger than this (and will therefore incur additional expenses), a conservative approach was taken by assuming all cases would consist of only two recipients.

Based upon a grant amount of $86 per month for twelve months, TDHR will pay $1032 to maintain a mother and one child on AFDC. In addition to this, it was calculated that the department will pay an estimated $1,006 in the first year for non-birth-related Medicaid benefits for the family. Thus, the cost for AFDC and Medicaid is $2,038 per

year. Adding this to the birth-related expenses yields the total first-year Title XIX cost for an AFDC birth: $4,149.

The department will also incur added food stamp costs for the new AFDC cases that receive food stamps. The maximum allotment for increasing the family size from one to two recipients is $50 per month ($600 per year). About 87% of all AFDC cases receive food stamps, and the average allotment for these cases is about 91% of the maximum. Therefore, the average per-case yearly increase in food stamp allotments for a mother and one child who go on onto AFDC is about $475 ($600 X .87 X .91).

To summarize, the costs to TDHR for cases that would go onto AFDC include $4,149 for birth-related costs and welfare maintenance, plus $475 in increased food stamp allotments. This totals to $4,624 per AFDC birth.

The next parameter needed is the percentage of Title XX women who would go onto AFDC due to the birth of a child. A direct estimate was not available from TDHR. Therefore, an indirect method was used to derive this parameter.

National and local data were used to ascertain the number of family planning patients (both Title XIX and XX) who were mothers. Also estimated was the proportion of these family planning mothers who were already Title XIX recipients (i.e., the proportion who were on AFDC). It was reasoned that this proportion, the proportion of family planning mothers already on AFDC, represented the probability that a Title XX nonmother would go onto AFDC due to the birth of a child. This reasoning was based on the assumption that Title XX women will go onto AFDC in the same proportions as current family planning mothers who are enrolled in AFDC.

In this way it was estimated that 30.4% of all family planning mothers are on AFDC in Texas. Among adolescents, the proportion was 35.8%, whereas among adults, the proportion was 28.3%. These figures are somewhat lower than national estimates made by Chamie and Henshaw (1982) who estimated that among women of low and marginal income, 47.8% of adolescents and 31.0% of adults would go onto AFDC if they had a child. However, because Texas income requirements for AFDC eligibility are among the lowest in the country (McManus and Davidson, 1982), it makes sense that the Texas proportions would be lower than the national estimates.

These estimates of the percent of Title XX women who would go onto AFDC were multiplied by the total cost per AFDC birth, resulting in the first component of the average cost per birth to a Title XX family planning patient: $1,309 for adults and $1,655 for adolescents, with an average of $1,406 for all Title XX births (this average is weighted in accordance with the size of the adolescent and adult Title XX populations).

The second manner in which the department can incur costs is due to increased food stamp allotments for those Title XX women who give birth but do not go onto AFDC. It will be recalled that the maximum increase in the food stamp allotment for increasing the family size from one to two would be $600 per year. However, only about 27.4% of Title XX family planning patients receive food stamps, and the average allotment for a non-AFDC case is only about 50% of the maximum. Thus, the average yearly increase in the food stamp allotment for non-AFDC births is about $82 ($600 X .274 X .50).

The proportion of women not going onto AFDC would simply be 100% minus the proportion that would go onto AFDC. Thus, about 64.2% of adolescents, 71.7% of adults, and 69.6% of all Title XX women would not go onto AFDC due to the birth of a child. Applying these proportions to the $82 figure yields the second component of the cost of a Title XX birth: $43 for adolescents,

$59 for adults, and about $57 for all patients combined.

The total cost per Title XX birth equals the sum of the two components (for AFDC and non-AFDC births): $1,708 for adolescents, $1,368 for adults, and $1,463 for births to all Title XX women combined. Table 3 uses these estimates to compute the overall costs and benefits attributable to the Title XX program. The above figures represent the average cost savings per birth averted. Multiplying these figures by the births-averted rates (in Table 2) yields the average cost savings per patient. Table 3 shows these costs savings to be $183 per adolescent, $130 per adult, and about $145 per patient of any age.

In FY 1981, 227,253 contraceptive and sterilization patients were served at a cost of $17,187,782; an average cost per patient of about $75. Dividing the average cost savings per patient by the average cost of delivering services, cost-benefit ratios were computed to be 1:2.44 for adolescents, 1:1.73 for adults, and 1:1.93 for all Title XX patients. Thus, for every dollar spent on the Title XX family planning program, about $1.93 is saved. For every dollar spent on services to adolescents, about $2.44 is saved, and about $1.73 is saved from services to adults.

Total first-year savings to TDHR attributable to births averted by the Title XX program were computed by multiplying the total program cost by the ratio of benefits to costs. These total savings amounted to nearly $33 million, with $11.7 million attributable to adolescent services and about $21.5 million attributable to adult services. Subtracting the cost of delivering services from total savings yields a net savings of nearly $16 million to the Department. Roughly $6.9 million of these savings can be attributed to adolescent services, and approximately $9 million can be attributed to adult services.

The cost-benefit ratio of 1:2.44 for adolescents is somewhat lower than the national estimate of 1:2.92 computed by Chamie and Henshaw (1982). Although the methodologies were similar, there are two reasons why the estimates reported here are not directly comparable to Chamie and Henshaw's. First, the cost

TABLE 3

Costs and Estimates Savings Associated with the Title XX Family Program for Adolescents and Adults

	Adolescents Only (19 and Under)	*Adults Only (20 and Over)*	*All Patients*
Savings			
Total estimated savings per birth averted	$1,708	$1,368	$1,463
Births averted per family planning patient	.107	.095	.099
Savings per family planning patient	$183	$130	$145
Costs			
Average Title XX expenditure per family planning patient (including sterilizations)	$75	$75	$75
Cost-Benefit Ratio	*1:2.44*	*1:1.73*	*1:1.93*
Total Cost and Savings in FY 81			
Number of patients	63,176	164,077	227,253
Total cost	$4,778,000	$12,410,000	$17,188,000
Total estimated savings	$11,658,000	$21,469,000	$33,173,000
Net estimated savings	$6,880,000	$9,059,000	$15,985,000

Note: Summing figures for adolescents and adults may not equal totals due to rounding.

per patient served was higher in the Title XX program than the national average. The Texas Title XX program served patients at an average cost of about $75, compared to the 1979 national average of about $63. This will, of course, lower the ratio of benefits to costs.

The second reason the Title XX ratio is lower is that Chamie and Henshaw's cost-savings calculations include estimates of birth-related costs to all public sector entities. In contrast, only a portion of these birth-related costs, those billed directly to TDHR, were included here. Thus, costs incurred by city and county hospitals, for example, to help women of low and marginal income pay for deliveries and first-year care were included in Chamie and Henshaw's estimates, but not in those used for this study.

Cost-Benefit Analysis of the Title XIX Program

Having estimated the costs and cost savings associated with the Title XX program, similar estimates can be made for the Title XIX program. For Title XIX women receiving family planning services, TDHR incurs direct costs due to the birth of a child through birth-related expenses and welfare maintenance and increased food stamp allotments for the child (because the mother is already a Title XIX recipient, her costs represent no increased expense). The costs associated with birth-related expenses (delivery and first-year care) were outlined above and were seen to be approximately $2,111 per birth. AFDC grants for an additional child are $30 per month, or $360 per year, and Medicaid expenses for the new child recipient are about $503 per year. Added to this is $475 in increased food stamp allotments, for a total of $3,449 per birth to a Title XIX recipient.

In Table 4, this cost per birth is multiplied by the births-averted rates to yield the cost savings per Title XIX family planning patient: approximately $369 per adolescent patient, $328 per adult, and $341 per patient of any age. In FY 1981, $4,959,614 Title XIX dollars were spent to deliver services to 44,476 patients through organized family planning providers

TABLE 4

Costs and Estimated Savings Associated with the Title XIX Family Program for Adolescents and Adults

	Adolescents only (19 and Under)	*Adults Only (20 and Over)*	*All Patients*
Savings			
Total estimated savings per birth averted	$3,449	$3,449	$3,449
Births averted per family planning patient	.107	.095	.099
Savings per family planning patient	$369	$328	$341
Costs			
Average Title XIX expenditure per family planning patient (including sterilizations)	$112	$112	$112
Cost-Benefit Ratio	*1:3.29*	*1:2.93*	*1:3.04*
Total Cost and Savings in FY 81			
Number of patients	11,653	32,823	44,476
Total cost	$1,300,000	$3,660,000	$4,960,000
Total estimated savings	$4,277,000	$10,724,000	$15,078,000
Net estimated savings	$2,977,000	$7,064,000	$10,118,000

Note: Summing figures for adolescents and adults may not equal totals due to rounding.

(this includes patients who received sterilizations, other contraceptive methods, or drug refills, but does not include services delivered by private physicians). This averages to about $112 per patient.

When this average cost is compared with the average cost savings, a cost-benefit ratio of 1:3.04 results for all Title XIX patients, 1:3.29 for adolescents, and 1:2.93 for adults. Thus, for every Title XIX dollar spent on family planning, the department saves about $3. For adolescents, the savings are even greater. Total first-year savings in FY 1981 are calculated to be about $15.1 million. Net savings, after subtracting the cost of service, are about $10.1 million: $3.0 million attributable to adolescent services and $7.1 million to adult services.

It should be noted that the per-patient cost of delivering Title XIX services ($112) is substantially higher than either the average Title XX cost ($75) or the 1979 national average ($63). Nevertheless, the cost-benefit ratio was higher than that calculated for Title XX or for the nation as a whole. This is because TDHR pays delivery and birth-related expenses for all Title XIX recipients who give birth, but for only a portion of Title XX patients who give birth (those who go onto AFDC). Even in Chamie and Henshaw's national study (1982), it was estimated that, on the average, only 65% of the birth-related expenses for women of low and moderate income would be assumed by the public sector. Thus, it is the high expense to TDHR for each Title XIX birth that makes the Title XIX family planning program so cost-effective.

Discussion

Many potential benefits of the family planning program's services were not included in this analysis. These include certain human and long-range benefits that can be attributed to the prevention of about 56,000 unwanted pregnancies and 27,000 births. Also not included are the benefits, financial and otherwise, of preventing nearly 22,000 abortions and over 7,000 miscarriages.

Even certain financial cost savings were not included in this analysis. For example, many Title XX women who give birth but who are not eligible for AFDC and Medicaid depend upon the financial resources of city and county governments to pay for their births. By preventing births in this group of women, cost savings are realized by city and county governments that are not included in this analysis. Nevertheless, even when consideration is limited only to direct, first-year, financial cost savings experienced by a state welfare department, the demonstrable benefits of the program are substantial.

It is important to note that, as in many cost-benefit analyses, many estimates and assumptions were required to calculate the cost savings attributable to the program. Although every attempt was made to do this as accurately and, when necessary, as conservatively as possible, the conclusions of the study are only as valid as these estimates and assumptions. Nevertheless, the fact that the final cost-benefit ratios and the estimates of pregnancies and births averted are generally consistent with previous estimates lends credence to the results.

One key parameter was the estimate of the number of pregnancies averted based upon the survey of contraceptive use among Title XX patients. Leridon (1977) states that this method of calculating pregnancies and births averted "often leads to an over-optimistic estimate, since one does not know whether the methods prescribed are effectively and correctly used" (p. 130). To some extent, this problem was overcome by employing use-effectiveness rates for the various contraceptive methods, for these rates reflect contraceptive effectiveness by taking into account factors that lead to less than maximal

performance. Perhaps more problematic is the issue of discontinuation or switching of methods because the calculation of pregnancies averted assumes that women have used the premethods for one year in the absence of the program, and that they will also continue to use the postmethod for a period of one year following their last clinic visit.

Data from the National Survey of Family Growth (Vaughan et al., 1980) indicate that most women seeking to prevent an unwanted pregnancy continue using contraceptives over a period of a year, and that only between 3% and 9% will abandon use of a contraceptive method altogether. However, this abandonment could affect the estimate of the number of pregnancies averted, probably leading to an overestimate. Switching among methods could also affect the estimates, although the effect should be relatively minor as long as the switching occurs primarily among the more effective methods. The reason for this is that most of the pregnancies averted, according to the estimation procedure used in this study, are due to the switch from no method before coming to a clinic to some effective method at the last clinic visit. Any of the effective methods are so much more effective than no method, that switching among them will not have a great effect on the estimate of the number of pregnancies averted.

Despite these apparent problems with the method of estimation used in this study, Forrest et al. (1981) found that in one instance, at least, this method yielded an estimate of pregnancies and births averted which was quite comparable to one obtained using areal multivariate analysis. Although the two methods used completely different analytical methods and data bases, both yielded fairly similar estimates of the number of births averted in the United States in 1975 (75 to 98 births per 1,000 patients based upon the contraceptive-use methodology compared with approximately 101 births from the areal multivariate analysis methodology).

Nevertheless, further research is needed comparing the contraceptive-use methodology with other methods that estimate averted pregnancies. The contraceptive-use methodology is a simple and economical method to employ, and could be implemented in a variety of settings if found to be valid. Although not routinely reported in data systems that currently exist, it was found in the course of conducting this study that the pre and postmethod data were readily accessible in clinic records. Using these data, it would be possible to estimate the number of pregnancies, births, abortions, and miscarriages averted on any level of analysis from the level of the individual clinic or provider all the way up to the state or national level. It is therefore hoped that additional research will be undertaken to validate this simple method of assessing the impact of family planning services.

References

Alan Guttmacher Institute (1982) Current Functioning and Future Priorities in Family Planning Services Delivery. New York: The Alan Guttmacher Institute.

Alan Guttmacher Institute (1981) Data and Analyses for 1980 Revisions of DHHS Five-Year Plan for Family Planning Services. New York: The Alan Guttmacher Institute.

Chamie, M. and S. K. Henshaw (1982) "The costs and benefits of government expenditures for family planning programs." *Family Planning Perspectives* 13: 117–126.

Dryfoos, J. G. (1982) "Contraceptive use, pregnancy intentions and pregnancy outcomes among U.S. women." *Family Planning Perspectives* 14: 81–94.

Forrest, J. D., A. I. Hermanlin, and S. K. Henshaw (1981) "The impact of family planning clinic programs on adolescent pregnancy." *Family Planning Perspectives* 13: 109–116.

Leridon, H. (1977) *Human Fertility: The Basic Components*. Chicago: Univ. of Chicago Press.

Malitz, D., R. L. Casper, and P. Romberg (1981) *Impact Evaluation of the Texas Department of Human Resources Family Planning Program*. Austin, TX: Texas Department of Human Resources.

McManus, M. A. and S. M. Davidson (1982) *Medicaid and Children: A Policy Analysis*. Evanston, IL: The American Academy of Pediatrics.

Vaughan, B., J. Trussel, J. Menken, and E. F. Jones (1980) *Contraceptive Efficacy Among Married Women Aged 15–44 Years*. Hyattsville, MD: National Center for Health Statistics.

David Malitz received his Ph.D. in quantitative psychology from the University of Texas at Austin in 1980. He is currently an independent consultant in statistical analysis and program evaluation specializing in public welfare, health, and safety research as well as statistically oriented software development.

Explanation and Critique

David Malitz's study attempts to estimate the benefits of a $22 million family planning program funded by the Texas Department of Human Resources (TDHR). In the beginning paragraphs he sets up the parameters of the study, outlining which aspects were and were not to be considered. He acknowledges that the cost estimates were easily obtained—the TDHR commissioned the study to see how *their* funds were spent. He limits the benefits to direct, first-year cost savings due to the prevention of unwanted pregnancies among Title XIX and XX family planning patients. He immediately makes his benefits estimate more conservative by eliminating long-term dependency costs and costs due to predicted abortions and miscarriages. In other words, he resisted the urge to "pad" the benefits, making the favorable relationship more difficult to obtain.

The benefits are the lack of expenditures that would have been incurred had the program not been in place, which complicates the matter. Family planning agencies cannot withhold services from eligible recipients, assigning them randomly into treated and untreated groups. Therefore, Malitz could not estimate averted pregnancies by counting the number of pregnancies in an untreated group; nor could he pull the figures out of thin air. Instead, he used national standards regarding contraceptive use among family planning patients and data regarding the outcomes of unwanted pregnancies in the United States as the basis from which he estimated averted pregnancies and their associated costs. Using these measures as standards has face validity because the Texas women *were* family planning patients and the consequences of nontreatment (unwanted pregnancies) had been tabulated in another setting. In a sense, Malitz could use these standards as comparison group in determining whether his findings were consistent with other research, in addition to using these prior results to project expenditures. In other words, he used generic controls.

Malitz reasonably assumed that in the absence of the family planning services, women would use the contraceptive methods they had been using when they entered the program. He used this preprogram information on Title XX recipients to estimate the number of averted pregnancies for both sets of recipients as a result of FY 1981 services. One could argue that in the absence of the program, women could have learned about or received other contraceptive advice result-

ing in averted pregnancies; however, the short time frame reduces the impact of this random effect. From the survey results and the baseline study that estimates the expected number of pregnancies per 1,000 women per method, Malitz estimated the number of adolescent and adult pregnancies averted (table 1 of the article). "These estimates were derived by multiplying the number of pregnancies among 1,000 users of each method by the percent using each method, and summing the resulting products across methods." Malitz used the midpoint estimates (again a relatively conservative measure) as the basis for extrapolating to the population served (table 2 of the article). Again, notice that he compared his findings with the Forrest and colleagues (1981) and AGI (1981) studies to determine whether the Texas experience was at all similar to the other studies. To the extent they are similar, he is confident in his estimation techniques. Again, this illustrates a use of generic controls. Because the costs to TDHR differ by whether one is a Title XX or Title XIX recipient, the estimates were calculated separately for each population. But what does the use of generic controls mean in terms of the evaluation design selected? Is it really a pretest-posttest control group design or something else?

All of this was accomplished before the actual cost-benefit analysis. The cost-benefit analysis takes into account the TDHR's responsibility for AFDC births as opposed to the reduced expenditure for non-AFDC births—those Title XX women who would not become AFDC eligible due to a birth. The only TDHR increment for those remaining Title XX–eligible women would be an increase in eligibility for food stamps. Therefore, three calculations were necessary to estimate benefits (nonadditional costs to TDHR)—one for current AFDC recipients, one for Title XX eligibles who would become AFDC eligible with a birth, and one for those who would remain only Title XX eligible. Notice how carefully (and often conservatively) Malitz uses existing information to calculate portions of the benefits. For example, consider the added food stamp costs for new AFDC cases that receive food stamps:

> The maximum allotment for increasing family size from one to two recipients is $50 per month ($600 per year). About 87% of all AFDC cases receive food stamps, and the average allotment for these cases is about 91% of the maximum. Therefore, the average per-case yearly increase in food stamp allotments for another child who go on AFDC is about $475 ($600 x .87 x .91).

In this and other examples Malitz tried to incorporate existing knowledge from the Texas program and the standards to arrive at realistic benefits. One seeming anomaly is that the current cost of supporting the averted-birth AFDC family was not subtracted from the AFDC birth calculation (the portion TDHR would pay anyway for maintenance without a birth). Although Malitz did not say so directly, it seems that this amount was taken care of by estimating maintenance at a two-person family amount (by Malitz's terms a conservative approach). This is because AFDC assumes at least a three-person family. Moreover, Title XX eligibles who became AFDC eligible are a 100 percent liability to TDHR—those expenditures are all new expenditures.

Malitz then took the average cost savings per Title XX birth averted, multiplied that by the birth averted rates (his table 2) to show the average cost savings per patient (his table 3). From there a traditional cost-benefit analysis was straightforward. For example, TDHR cost to the adolescent family planning patient was $75 divided by the cost savings for that group (benefits) of $183, which yielded a cost-benefit ratio of 1 to 2.44. That ratio means that for every $1 (cost) provided by TDHR for family planning services to adolescents, the department saved $2.44

(benefits) in costs it would have had to provide for a Title XX adolescent birth. This procedure was duplicated for adults, all patients and groups of Title XIX eligibles (Malitz's table 4).

At the bottom of his tables 3 and 4, Malitz calculated the total cost savings (the "bottom line") of providing family planning services to the population of recipients. The net savings estimate was calculated by taking the total cost from the total estimated savings to arrive at the amount TDHR was ahead given the current level of services.

The only question Malitz could not adequately address was the "What if ?" question. What would Title XX–eligible (or Title XIX–eligible, for that matter) women do in the absence of the family planning services if they were truly concerned about averting pregnancies? How many would pay a private doctor for contraceptive services? How many would require partners to supply condoms? Because of the sliding fee scale used to provide many Title XX services, one would expect that those women at the top of the scale (those with higher incomes) would be able to provide a few extra dollars for private assistance. In the absence of information on this, Malitz could not make an informed guess. He tried to overcome this weakness by consistently using conservative benefit estimates.

All in all, Malitz presents an analysis suggesting that the benefits of providing family planning services in Texas were greater than the costs incurred. Does this automatically mean that Texas will continue funding those services? That is a political question. Social conservatives may argue that government has no business being in the family planning business—or supporting indigent welfare mothers for that matter. Cost-benefit analysis is a tool to assist decision makers to make informed decisions based on sound empirical evidence regardless of whether one believes the ultimate decision to be sound.

Malitz's research is not the first article in the book that might benefit from cost-benefit analysis. Given the data presented by Ann Solberg in chapter 7, it would have been relatively simple to calculate the cost-benefit of posthospital follow-up of psychiatric patients.

Reference

Savas, E.S. 2000. *Privatization and Public-Private Partnerships*. New York: Chatham House.

CHAPTER 14

Cost-Effectiveness Analysis

In the previous chapter, cost-benefit analysis was used to determine if program objectives were economically beneficial or justifiable. In contrast, in cost-effectiveness analysis, the results of a project are presumed to be worthwhile per se. Cost-effectiveness analysis "explores how such results might be efficiently achieved and which costs are attached to them for reaching different levels of the desired outcomes" (Peterson 1986, 31). Like cost-benefit analysis, cost-effectiveness analysis assumes that all of the costs (both direct and indirect) of a program can be calculated. The difference arises in the determination of the denominator. Rather than attaching a dollar amount, a meaningful unit of effectiveness measure (or unit of service) is substituted. For example, one would speak in terms of the number of dollars spent per ton of garbage picked up per day or dollars per life saved.

Cost-effectiveness designs are reasonably employed when it is difficult to attach a dollar amount to the denominator (e.g., the value of a life saved). Their utility is traditionally associated with the exploration of a number of different alternatives to a specified, desired outcome. Like cost-benefit analysis, then, cost-effectiveness results are valuable tools for decision makers, particularly during budgetary deliberations. In addition, cost effectiveness can act as a very good measure of internal evaluation. It allows one to determine how well the organization or program is performing as compared to its performance before a change in the program was initiated. An example would be comparing the cost per ton of garbage collected in 1996 to that collected in 1999, controlling for inflation. Generic controls are often used to compare a program's cost effectiveness with national or professional standards (e.g., library book circulation standards, parkland acquisition standards). Finally, elected officials in one jurisdiction may wish to gauge their performance relative to that of other jurisdictions that are similar to them in order to see how well they are performing.

Whether it is a before-after, cross-community, or across-alternative process, cost effectiveness is based on comparisons. It is seldom useful if no comparisons are made. R.D. Peterson outlines three approaches to cost-effectiveness comparisons: constant-cost analysis, least-cost analysis, and objective-level analysis (1986, 32). In a constant-cost analysis, the job of the researcher is to determine how much of an objective can be obtained within

given cost constraints. Peterson's example is a road accident reduction program. He writes:

> a legislature may vote to allocate only $2.5 million to accomplish the objective. A constant-cost analysis would focus on the total number of deaths that could be prevented by spending $2.5 million for each alternative that could achieve that objective.

The least-cost analysis goal is to reach a prespecified goal for the least cost. Peterson goes on to say:

> For example, suppose that the stated objective was "to reduce deaths due to automobile accidents by 10 percent each year." Based on this criterion, the cost effectiveness question involves finding the alternative that will achieve that end in the least expensive way. Each alternative way of achieving the stated objective would be analyzed in terms of the dollar amount of expenditures required to reach the specified goal.

The objective-level analysis compares varying performances levels under a single alternative. As Peterson says:

> In preventing highway deaths, the analyst might seek to determine costs according to 10%, 20%, 40% and 80% reductions in deaths for each specific program alternative. . . . The costs per unit of objective reached are then computed for each successive level.

A researcher involved in cost-effectiveness analysis usually identifies alternatives for reaching objectives and calculates cost-comparison ratios. In the following article, Elaine Graveley and John Littlefield (1992) perform a version of the least-cost analysis to compare the costs of three staffing models for the delivery of low-risk prenatal care.

READING

A Cost-Effectiveness Analysis of Three Staffing Models for the Delivery of Low-Risk Prenatal Care

Elaine A. Graveley, *DBA, RN*
John H. Littlefield, *PhD*

Abstract

Background. Health care costs are increasing at more than twice the rate of inflation, thus, public officials are seeking safe and economic methods to deliver quality prenatal care to poor pregnant women. This study was undertaken to determine the relationship between the cost and effectiveness of three prenatal clinic staffing models: physician based, mixed staffing, and clinical nurse specialist with physicians available for consultation.

Methods. Maternal and neonatal physiological outcome data were obtained from the hospital clinical records of 156 women attending these clinics. The women were then interviewed concerning their satisfaction with their prenatal care clinic. The financial officer from each clinic provided data on the clinic staffing costs and hours of service.

Results. There were no differences in outcomes for the maternal-neonatal physiological variables, although newborn admission to the Neonatal Intensive Care Unit (NICU) approached significance among the clinics. The clinic staffed by clinical nurse specialists had the greatest client satisfaction and the lowest cost per visit.

Conclusions. The use of clinical nurse specialists might substantially reduce the cost of providing prenatal care while maintaining quality, and might thereby save valuable resources. (*Am J Public Health* 1992;82:180–184)

The authors are with the University of Texas Health Science Center, San Antonio.

Requests for reprints should be sent to John H. Littlefield, PhD, Director, Instructional Development, University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78284-7948.

This paper was submitted to the journal November 2, 1990, and accepted with revisions July 17, 1991.

Introduction

WITH HEALTH CARE COSTS growing at more than twice the rate of inflation, public officials are seeking safe and economic methods to deliver quality health care to indigent populations. The percentage of indigent individuals who receive quality health care is decreasing. Lewin[1] reported that, from 1976 to 1983, Medicaid covered a decreasing number of the poor (of the population with incomes below the poverty line, 65% were covered by Medicaid in 1976 and 53% in 1983) and that poor pregnant women have been increasingly underserved as a result of budget cuts and policy changes.

The relationship between low-birthweight (LBW) babies and the absence of prenatal care is well documented.[2–4] The Office of Technology Assessment (OTA) estimated that for every LBW birth averted, between $14,000 and $30,525 in newborn and long-term health care costs could be saved.[5] OTA also analyzed the cost-effectiveness of expanding Medicaid coverage to all poverty-level pregnant women in an effort to provide them with early prenatal care. Their conclusion was that such expansion would be cost effective and would be a good investment for the nation.

The National Commission to Prevent Infant Mortality (NCPIM) also studied the problem of LBW and proposed universal access to early maternity care for all expectant mothers.[6] However, neither NCPIM nor OTA recommended the type of health care provider for these additional 194,000 women.[5] With our increasingly litigious society, the number of obstetricians willing to serve Medicaid patients is declining. There are, however, other models of prenatal care that need to be studied for their cost-effectiveness: individual or group physician-based clinics and, depending on state laws, nurse-based clinics that tie medical practice with a nurse midwife, nurse practitioner, or clinical nurse specialist. The focus of this study was a comparative cost-effectiveness analysis of three low-risk prenatal clinic staffing models: physician based, mixed staffing (physician, nurse practitioner, registered nurses, nurse aides), and nurse based. Low-risk prenatal care was defined in this study as care given to a pregnant woman "whose complaint or problem has been resolved, or does not pose a threat to life."[7] The results of this study should be helpful to public policymakers who are implementing OTA's and NCPIM's proposal to increase the availability of prenatal care for indigent patients.

Methods

Clinics

The three clinics served predominantly Hispanic clients of low socioeconomic status and referred all women with complications to the Complicated Obstetrical Clinic at the community health center. The women primarily delivered at the hospital operated by the county hospital district.

Clinic 1. This private-not-for-profit entity was funded under Titles 329 and 330 of the Public Service Health Act and served as the physician-based-clinic (MDC) in this study. Women saw the same physician for all their prenatal care and counseling and received a standard clinic appointment. Fees were charged on a sliding scale, according to the woman's ability to pay.

Clinic 2. This clinic, one of 19 operated by the city health department, served as the mixed-staffing clinic (MSC). The clinic was located within the same catchment area as the MDC. Each patient saw a nurse aide, a registered nurse for prenatal teaching, and the nurse practitioner and/or contract physician at each visit. The clinic operated twice a week from 8 am to 12 noon. All women received an 8 am appointment, and prenatal care and counseling were provided for $50 unless the woman was unable to pay.

Clinic 3. This clinic was jointly operated by the county hospital district and the local University Health Science Center's School of Nursing and Department of Obstetrics and Gynecology and served as the nurse-based clinic (RNC). The clinic was located within the county community health center. Three clinical nurse specialists delivered total prenatal care, including referrals to community agencies. Fees were charged on a sliding scale, according to the woman's ability to pay, and the women received a standard appointment time.

Subjects

The sample size, 156 subjects, was determined by a power analysis of 3 months of birth-weight data from women who attended the clinics and delivered at the county hospital ($\alpha = .05$, effect size = .25, power = .8; 3 groups). The study sample was drawn over a 3-month period in 1989 from women who had delivered at the county hospital. To be included, subjects had to: (1) be 18 years of age or older, or emancipated minors; (2) have obtained their prenatal care at one of the three clinics with a minimum of three prenatal visits (three visits was considered the minimum for the woman to have sufficient experience

to validly report her satisfaction with her prenatal care clinic); and (3) have delivered within 48 hours of the interview.

Potential subjects were identified from their hospital record and asked to participate in a study of prenatal care clinics. No potential subject was excluded because of language, a Spanish interpreter was available to all women, and none of the women declined to participate.

Determining Pregnancy Outcomes and Satisfaction

Four broad categories of patient variables were analyzed: demographic, physiological, satisfaction with care, and cost (see Table 1). The physiological variables related to low-risk prenatal care were gathered from previous prenatal research studies.[5,6,8] The Kessner Index[9] was used to assess the adequacy of prenatal visits. This index categorizes prenatal care as adequate, intermediate, or inadequate depending on the weeks of gestation and the number of prenatal visits. Differences in the number of laboratory tests among the clinic could not be determined, as one clinic denied access to its patient records. However, both the MSC and RNC nurses used similar screening protocols and had similar guidelines for physician consultation.

Maternal satisfaction with access to care was assessed via a patient satisfaction tool[10] (PST) consisting of 27 statements that the subject rated on a 6-point Likert scale. The statements addressed five categories of satisfaction: accessibility, affordability, availability, acceptability, and accommodation. Five scores, one for each category, were generated by the PST. The readability index of the PST was at the fifth-grade level, and the internal consistency reliability coefficient was reported as .83. Both an English and a Spanish version of the tool were available to the subject. The interview took approximately 20 minutes to complete.

Determining Costs

The financial officer of each clinic provided data on five variables for study: number of staff, hourly wages (see Table 2), number of prenatal appointments made and kept, and number of hours spent delivering prenatal care. The RNC nurses kept a flowchart of the number of times they conferred with a physician and the time involved in order to determine the cost of consultation. The cost of the consultation time was computed on the basis of the hourly salary for the MSC contract physician. No physical facility costs were included in the analysis, as they were not con-

TABLE 1
Variables Studied

Demographic	*Physiological*	*Satisfaction*	*Cost*
Ethnicity	Maternal	Patient satisfaction tool	Personnel
Age	Weight gain		Hours delivering prenatal care
Education	Hemoglobin		Consultation (nurse-based clinic)
Medicaid	Complications at the time of delivery		No. of appointments made/kept
Marital status	Neonatal		
Gravida	Weeks of gestation		
No. of prenatal visits	Birth weight		
Adequacy of prenatal visits	Apgar score		
Prenatal classes	Admission to Neonatal Intensive Care Unit		

sidered germane to determining the clinics' staffing model costs.

Cost per clinic visit was determined by dividing the personnel costs by the number of kept appointments. Clinic productivity was determined by dividing the number of patients seen for prenatal care by the number of hours spent delivering prenatal care. In a cost-effectiveness analysis, the differences among the clinic costs and appointment outcomes are calculated using percentage differences. The staffing model with the lowest outcome value or cost item is identified, and then the percentage differences between this model and the other two models are calculated.[11]

Results

Demographic Variables

Subjects at the three clinics did not differ significantly on any of the demographic variables except ethnicity, years of education, and Medicaid coverage (see Table 3). Compared with the MDC and MSC, subjects at the RNC were less likely to be Hispanic, more likely to have graduated from high school, and more likely to be on Medicaid.These significant variables were further evaluated using x^2 tests were significant.

Analysis of variance (ANOVAs) were computed to assess the relationship between patient ethnicity, educational level, and Medicaid coverage with satisfaction as measured by the PST. Ethnicity and Medicaid coverage were not significantly related to any of the PST statements.

Educational level was statistically significant for two PST statements ("I saw the same physician/registered nurse for all my prenatal care" and "The fees at the clinic were more than I wanted to pay"). These two statements were further analyzed using a Student Newman-Keuls post hoc *t* test to investigate pairwise differences. For the first statement ("saw the same physician/registered nurse"), those subjects with 12 years of education differed significantly from the group with 8 to 11 years of education. Cross tabulation, controlling for the clinic, was done for each group; the 19 subjects who responded "never" to the statement were from the MSC. A Student Newman-Keuls test for the statement referring to the fees being too high showed that all educational levels differed significantly from the group with the 0 to 7 years of education. The lower education group was the only one in which all members (n = 4) indicated that the fees charged were "never" fair. Each clinic was represented in this group. Ninety-four percent (n = 147) of the subjects responded that the fees were "always" fair.

Analysis of the subject's reported cost for her prenatal care was done using an ANOVA. The ANOVA was significant, $F(2, 156) = 28$, $P = .000$. The grand mean for the three clinics was $29.40; the means for the individual clinics were $59.54 for MDC, $8.71 for MSC, and $19.16 for RNC. Fifty-six percent of the subjects (n = 88) reported that they paid nothing. A Pearson's *r* showed no significant relationship between subject-reported costs and the number of prenatal visits. Also, there was no significant relationship between the clinic's charge to each subject for her prenatal care, the maternal/neonatal physiological variables, and NICU admission. These differences in subject demographic data, therefore, did not appear to be related to the categorical dependent variables in the study.

Physiological Outcomes

There were no significant differences among the clinics for any of the maternal physiological variables. Twenty-eight percent (n = 20) of the subjects had an inadequate number of prenatal visits; the mean hemoglobin level at the time of hospital admission for delivery was 12 g, the mean weight gain was 30 lb, and there were no documented maternal complications at the time of delivery.

TABLE 2
Clinic Staffing and Hourly Salaries

	Clinic					
	MDC		MSC		RNC	
	No. of Personnel	Hourly Salary, $	No. of Personnel	Hourly Salary, $	No. of Personnel	Hourly Salary, $
Physician						
Family practice	3	$41.96	—		—	
Contract	1	$65.00	1	35.00	—	
Obstetrical resident for consultation	—		—		1	35.00
Nurse Practitioner	—		1	15.48	—	
Clinical nurse specialist	—		—		3	20.04
Registered nurse	—		5	13.53	—	
Licensed vocational nurse	3	$8.45	—		—	
Nurse Aide	—		3	7.06	—	
Social worker	—		1	11.24	—	
Caseworker	1[a]	$5.19	—		—	
Clerk	1[a]	$5.35	1[a]	7.11	1[a]	6.47

Note: MDC = physician-based clinic; MSC = mixed-staffing clinic; RNC = nurse-based clinic
a. Part-time.

Babies of the subjects of the three clinics did not differ on any of the five neonatal variables: 85% (n = 133) were between 38 and 42 weeks of gestation, 86% (n = 134) weighed between 2500 and 4000 g, 28% (n = 43) were admitted to the NICU, 85% (n = 132) were discharged with their mother, and there were no fetal deaths or baby deaths within the first 48 hours of life.

Maternal Satisfaction

The PST internal reliability coefficient in this study was .76. ANOVA tests revealed no significant differences among the clinics for the satisfaction categories of accessibility and affordability. A Student Newman-Keuls post hoc *t* test showed the significant differences among the mean scores for the other groupings (see Table 4).

Cost Data

Table 5 shows the cost and outcome data per clinic and Table 6 shows the percentage differences among the clinics for the supplied data. Although the MSC had the lowest personnel costs, the RNC had the lowest cost per clinic visit and the highest productivity.

Discussion

The absence of significant differences among the three clinics for the maternal-neonatal physiological variables was expected and supports other reports that nurses prepared for a specific area of health care can provide quality care.[7,12–15] The number of NICU admissions approached significant differences, with the MSC having 35% (n = 20) of its subjects' babies admitted compared with 21% (n = 11) for the MDC and 23% (n = 12) for the RNC. The number of complicated pregnancies referred to the health center by each clinic was not analyzed. It is possible that the MSC referred fewer complicated pregnancies and therefore increased its number of NICU admissions. As 35% is higher than the 28% average number of newborn NICU admissions

(level 1, 2, or 3) for this hospital, this may deserve further research.[16] However, it did not affect the conclusions of this study.

The RNC subjects indicated significantly higher satisfaction than the MDC subjects in the access-to-care dimension of accommodation and overall satisfaction. This finding supports the studies, cited by OTA,[5] that reported patients often found that nurse practitioners exhibited more interest, reduced the professional mystique of health care delivery, conveyed more information, and were more skillful than physicians when providing services that depended on communication with patients.

However, the subjects at the MDC, which had no nurse practitioners, indicated significantly greater satisfaction with availability and accommodation and had a higher total satisfaction score than the subjects at the MSC, which had one nurse practitioner. One explanation for this might be that at both the MDC and RNC, subjects had only one caregiver throughout their pregnancy. All MDC and RNC subjects were able to name their care provider, while not one MSC subject could name a single individual who either had given her care or had provided health education during her prenatal visits. Subjects at the MSC saw a physician and/or nurse practitioner at each visit; however, it was not necessarily the same individual. In addition, the MSC policy of not making specific patient appointments, with the subjects sometimes waiting several hours for an appointment, was cited by many subjects as being very difficult. Lowered subject satisfaction with the MSC was probably due to its policy regarding block appointments and limited time spent with any care provider.

Prenatal care costs were based on data provided by the clinic financial officers. Costs were based on 15 minutes per visit at the MDC and 30 minutes per visit at the RNC; for the MSC costs were based on preset times for each personnel category during each 4-hour period of clinic operation. For example, during this

TABLE 3
Comparison of Clinics for Significant Demographic Data

	Clinic							
	MDC		***MSC***		***RNC***			
	n	%	*n*	%	*n*	%	x^2	*P value*
Ethnicity							26.8	.0002
Anglo		—	1	02	9	17		
Black		—	4	08	5	10		
Hispanic	52	100	47	90	38	73		
Education, y							16.5	.04
0–7	7	13	7	13	3	06		
8–11	26	50	28	54	14	27		
12	13	25	10	19	23	44		
13+	3	06	5	10	9	17		
General equivalency diploma	3	06	2	04	3	06		
Medicaid							11.4	.003
Yes	14	27	24	46	31	60		
No	38	73	28	54	21	40		

Note: MDC = physician-based clinic; MSC = mixed-staffing clinic; RNC = nurse-based clinic.

TABLE 4
Student Newman-Keuls Significant PST Category Means by Clinic

	Clinic Mean Score		
Significant Category	*MDC*	*MSC*	*RNC*
Availability	22.3	18.5[a]	23.5
Acceptability	52.3	51.6[b]	63.9
Accommodation	36.2[b]	32.3[a]	37.9
Total PST score	124.5[b]	116.6[a]	129.1

Note: PST = patient satisfaction tool; MDC = physician-based clinic; MSC = mixed-staffing clinic; RNC = nurse-based clinic.
a. Significantly lower score compared with MDC and RNC.
b. Significantly lower score compared with RNC.

study the clinic clerk at the MSC was assigned only 26 hours of the scheduled 104 hours of clinic operation. In addition, this clinic was frequently open for more than 4 hours, which was not accounted for in the supplied cost data. Therefore, the MSC personnel costs and cost per visit might actually be higher and their productivity lower than reported in Table 5. The lower cost per RNC appointment was primarily due to limited use of physician time, which is the most expensive resource in delivering low-risk prenatal care.

Discussions with the RNC staff concerning the time spent waiting for physician consultation and the physician's changes to a proposed plan of care suggest the potential value of a yearly review of clinic patient care protocols by the nurse-physician team. This review would help to reduce the nonproductive waiting time for the clinical nurse specialists.

Both the RNC and the MSC had 27% missed appointments, compared with 18% at the MDC. A high incidence of missed appointments ultimately increases clinic costs because fixed personnel expenses are expended without delivering care. The lower missed-appointment incidence at the MDC may be attributed to the effectiveness of its caseworker, who called patients with missed appointments, discussed with them the importance of their clinic visits, and then tried to reschedule them.

Qualitative data can be very useful in studies of health care delivery.[17,18] This study of staffing models found significant differences in patient satisfaction and cost-effectiveness. Qualitative data from patient interviews helped ascertain that lower satisfaction with the MSC was due in part to clinic scheduling policies. In a similar vein, differences in the number of missed appointments were due in part to an effective caseworker in the MDC. These qualitative data are critically important to field studies of health care delivery that, by their nature, lack tight experimental controls.

Conclusion

Escalating health care costs, coupled with decreasing governmental economic resources, are reducing the availability of health care for underserved populations such as pregnant women.

TABLE 5
Cost and Outcome for Three Staffing Models

	Alternatives		
Measure	*MDC*	*MSC*	*RNC*
Cost			
Personnel costs	$18,965	$12,381	$13,006[a]
Outcome			
Appointments made	1695	1341	1665
Appointments kept	1397	984	1216
Hours	804	962	590
Productivity	1.7	1.02	2.06
Cost-effectiveness			
Cost per appointment	$11.19	$9.23	$7.81
Cost per kept appointment	$13.58	$12.58	$10.70

Note: MDC = physician-based clinic; MSC = mixed-staffing clinic; RDC = nurse-based clinic
a. Includes 6 hours of consultation time

TABLE 6

Percentage Differences in Cost and Outcome Data for Clinic Staffing Models for 3 Study Months

	Clinic, %		
Measure	*MDC*	*MSC*	*RNC*
Personnel costs	53	0	5
Kept appointments	42	0	24
Hours delivering Prenatal Care	35	63	0
Productivity	67	0	101
Cost per kept appointment	27	18	0

Note: MDC = physician-based clinic; MSC = mixed-staffing clinic; RNC = nurse-based clinic

This study found that increasing the availability of low-risk prenatal care professionals through the use of nonphysician maternal health providers, with physicians available for consultation, might substantially reduce the cost of providing this care while maintaining quality. Therefore, such a system might save valuable resources. The results have direct public policy implications for the staffing of low-risk prenatal care clinics. The findings regarding the mixed-staffing model underscore the importance of analyzing environmental issues such as clinic policies regarding individual patient appointments and lack of continuity of care.

Acknowledgment

This paper was presented at the Southern Nursing Research Conference, Atlanta, Georgia, 1990.

References

1. Lewin ME. Financing care for the poor and underinsured: an overview. In Lewin ME, ed. *The Health Policy Agenda.* Washington, DC. American Enterprise Institute for Public Policy Research; 1985:26–43.
2. Moore TR, Origel W, Key TC, Resnick R. The perinatal and economic impact of prenatal care in a low-socioeconomic population. *Am J Obstet Gynecol.* 1986;154:29–33.
3. Scholl TO, Miller LK, Salmon RW, Cofsky MC, Shearer J. Prenatal care adequacy and the outcome of adolescent pregnancy: effects on weight gain, preterm delivery, and birth weight. *Obstet Gynecol.* 1987;69:312–316.
4. Showstack JA, Budetti PP, Minkler D. Factors associated with birthweight: an exploration of the roles of prenatal care and length of gestation. *Am J Public Health.* 1984;74:1003–1008.
5. Office of Technology Assessment. *Healthy Children Investing in the Future.* Washington, DC: Office of Technology Assessment; 1988.
6. National Commission to Prevent Infant Mortality. *1985 Indirect Costs of Infant Mortality and Low Birthweight.* Washington, DC; National Commission to Prevent Infant Mortality; 1988.
7. University of Texas Health Science Center, San Antonio School of Nursing. *Developments in Prenatal Care.* San Antonio, Tex: University of Texas Health Science Center, San Antonio School of Nursing.
8. Williams RL, Binkin NJ, Clingman EJ. Pregnancy outcomes among Spanish-surname women in California. *Am J Public Health.* 1986;76:387–391.
9. Kessner DM, Kalk CE. Infant death: An analysis of maternal risk and health care. In: *Contrast in Health Status.* Washington, DC: National Academy of Sciences; Institute of Medicine; 1973;1.
10. Reed SB, Draper SJ. *The Mothers of the Nurses Prenatal Clinic (1978–1982).* Unpublished manuscript, University of Texas Health Science Center San Antonio School of Nursing.
11. Reynolds J, Gaspari KC. *Operations Research Methods: Cost-effectiveness Analysis.* Chevy Chase, Md: Primary-Health Care Operations Research; 1985. Monograph Series: Methods Paper 2.
12. Watkins LO, Wagner EH. Nurse practitioner and physician adherence to standing orders criteria for consultation or referral. *Am J Public Health.* 1982;72:22–29.
13. Hastings GE, Vick L, Lee G, Sasmor L, Natiello TA, Sanders JH. Nurse practitioners in a jailhouse clinic. *Med Care.* 1980; 18:731–744.
14. Bibb BN. Comparing nurse practitioners and physicians: a simulation study on process of care. *Eval Health Professions.* 1982;5:29–42.

15. Denton FT, Spencer BG, Gafni A, Stoddart GL. Potential savings from the adoption of nurse practitioner technology in the Canadian health care system. *Socio-Economic Plan Sci.* 1983;17:199–209.
16. Smith C. Saving newborn lives—otherwise lost. *San Antonio Express-News Sunday Magazine.* May 1, 1988;18–20, 23, 26–29, 33.
17. Lincoln Y, Guba E. *Naturalistic Inquiry.* Beverly Hills, Calif: Sage; 1985.
18. Jacob E. Clarifying qualitative research: a focus on traditions. *Educ Researcher.* 1988;17(1):16–24.

Explanation and Critique

"A Cost-Effectiveness Analysis of Three Staffing Models for the Delivery of Low-Risk Prenatal Care" by Elaine Gravely and John Littlefield is an example of a straightforward cost-effectiveness study. Graveley and Littlefield wanted to determine the comparative cost-effectiveness of three different models of staffing low-risk (healthy mother and fetus) prenatal clinics: (1) physician staffed (MDC), (2) nurse staffed (RNC), and (3) mixed staffing (MSC). At the MDC, women saw the same physician for all their prenatal care and counseling, and they received standard clinic appointments. At the RNC, patients saw the same clinical nurse specialists, who delivered total prenatal care. Patients also received standard clinic appointments. Procedures were a little different at the MSC. Each patient saw a nurse aid, a registered nurse for prenatal teaching, and the nurse practitioner and/or physician at each visit. In addition, patients were all expected to report at 8:00 A.M. for appointments and waited their turn to see clinic personnel.

A total of 156 subjects were selected shortly after giving birth at the county hospital (52 from MDC, 52 from RNC, and 52 from MSC). All 156 agreed to participate in the study. We are not told how the subjects came to attend the particular clinic they attended, but we assume it was not randomly.

Four broad categories of variables were compared—demographic, physiological, patient satisfaction, and cost. There were few differences between the groups with regard to the demographic variables and none for the physiological variables.

Patient satisfaction was a different story, however. MDC and RNC patients were satisfied with their services. Lower satisfaction with MSC was apparently due to its policy regarding block appointments (8:00 A.M. for everyone) and limited time spent with any caregiver. (All MDC and RNC subjects were able to name their care providers, while none of the MSC subjects could name a single individual who had either provided instruction or given care.)

So both the MDC and RNC models were found to be approximately equally effective in terms of service delivery. The question then is this: Which method of delivery is most cost effective?

The financial officer of each clinic provided data on the following—number of staff, hourly wages, number of appointments made, number of appointments kept, and number of hours spent delivering prenatal care. Obviously doctors are more expensive than nurses, but other factors come into play. Both the RNC and MSC had 27 percent missed appointments compared with only 18 percent at the MDC.

The financial data are shown in table 5 of the article. The productivity of the various options were compared by dividing the num-

ber of clients seen (appointments kept) by the number of hours delivering the services:

	MDC	RNC	MSC
Appointments kept	1,397	1,216	984
Hours	804	590	962
PRODUCTIVITY (HOURS PER KEPT APPOINTMENT)	1.70	2.06	1.02

The higher the score, the more productive the facility. The RNC model was the most productive, followed by the MDC.

Cost-effectiveness was determined by dividing personnel costs by appointments kept:

	MDC	RNC	MSC
Personnel costs	$18,965	$13,006	$12,381
Appointments kept	1,397	1,216	984
COST PER KEPT APPOINTMENT	$13.58	$10.70	$12.58

Thus, as with productivity, the RNC model was the most cost-effective. The authors conclude that

> increasing the availability of low-risk prenatal care professionals through the use of nonphysician maternal health providers, with physicians available for consultation, might substantially reduce the cost of providing this care while maintaining quality.

This is a nice little cost-effectiveness study; it is straightforward, and there is nothing fancy about it. It deals with a small problem. But it is a good example of a thoughtful, thorough, and effective cost-effectiveness study. What is the evaluation design the authors use?

References

Graveley, Elaine A., and John H. Littlefield. 1992. "A Cost-Effectiveness Analysis of Three Staffing Models for the Delivery of Low-Risk Prenatal Care." *American Journal of Public Health* 82 (February):180–84.

Peterson, R.D. 1986. "The Anatomy of Cost Effectiveness Analysis." *Evaluation Review* 10 (February): 291–44.

PART VI

Other Designs

Part VI deals with designs that do not easily fit into the categories of the preceding four parts of the book; however, they build from the material presented there. We call the first one a "patched" design, since it patches together elements from the previous chapters. A good evaluator will be able to assess the existing or available data relating to a program evaluation and attempt to maximize the strengths of various designs to construct the best design given the data. Of course, ideally, you would want to have the ability to randomly assign observations to experimental and control groups and then gather pretest and posttest time-series data for both groups. But evaluators seldom have that luxury. Now that you know the strengths and weaknesses of a variety of evaluation designs, however, you are in a position to construct and assess patched designs as you will see in chapter 15.

As mentioned in chapter 1, meta-evaluations are systematic attempts to make sense out of a number of outcome evaluations of similar programs. Chapter 16 provides an example of this type of evaluation. It examines a program that has been widely implemented in local governments, Project DARE.

Still other types of "Other Designs" exist—most notably, those based on case studies and other qualitative methods. Since this book is designed to examine quantitative methods, review of qualitative methods is outside its scope. If you plan to perform case studies, however, we suggest you consult with the leading scholar in the field, Robert Yin (1993; 1994). Yin suggests that the main difference between case studies and other forms of quasi-experimentation is that case studies consider "context" to be an essential part of the phenomenon being evaluated (1994, 64). The types of data collected in case studies may be either qualitative, quantitative, or both. Therefore, many of the techniques covered in this text are equally applicable to case study evaluations.

References

Yin, Robert K. 1993. *Application of Case Study Research.* Thousand Oaks, Calif.: Sage.

Yin, Robert K. 1994. *Case Study Research: Design and Methods.* Thousand Oaks, Calif.: Sage.

Supplementary Readings

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* Cambridge, England: Cambridge University Press, chaps. 8–9.

Mohr, Lawrence B. 1988. *Impact Analysis for Program Evaluation.* Chicago: Dorsey, chap. 8.

CHAPTER 15

Patched Designs

In each of the preceding chapters, one specific evaluation design was examined discretely—that is, by itself. This was done to make learning about evaluations easier. It is obviously easier to learn about one technique or problem than it is to try to sort out several designs at one time.

Sometimes the evaluator will have the opportunity (although some would say the bad luck) to draw from more than one pure design—to "patch together" a design that maximizes the validity of the design under the constraints of the current study. For instance, say there are data already collected at one point during the pretest stage; however, the evaluator has access to quarterly data collected during the first three years of the program. One could then patch together a pretest-posttest, time-series design. It would have a one-point (non–time-series) pretest and a time-series of twelve points for the posttest. In drawing from the available data, the evaluator is only constrained by imagination and ingenuity in patching together what works best in a given situation.

The following article, "Attitude Change and Mental Hospital Experience" by Jack Hicks and Fred Spaner, is an example of a patched design. The article features not one, but two designs. The authors identify the first design as a "recurrent institutional cycles design" developed by Donald Campbell and Julian Stanley (1962, 171–246)—a design never discussed in this book. Furthermore, the second design does not perfectly fit the designs presented in Parts II, III, IV, or V. But this should pose no problem. Each real-world evaluation problem is unique, and you should not expect textbook designs to be completely appropriate for all problems. For each design described and illustrated, there are clearly numerous modifications. The "recurrent institutional cycles design" is a modification of a design already presented, as is the design adopted for the second experiment.

While reading the article, it might help to view the two experiments as two cases (they are). The authors were dissatisfied with the results of the first experiment and adopted a new design to overcome the weaknesses of the first evaluation.

Try to identify what the recurrent institutional cycles design actually is, based on your reading of the earlier chapters. Then focus on the threats to validity: What are these threats? Which threats could not be overcome, thus forcing the authors to adopt a new design and conduct a second experiment?

Then, what is the second design? Does it overcome the problems of the initial experiment? At the end are you struck with serious reservations about the results? If so, there must be certain sources of invalidity that were not overcome successfully. Are there any "neat" aspects to the experiment? Have the authors been able to measure any external influences that are not attributable to the program?

We believe that if a reader can satisfactorily critique, not criticize, this case, then he or she understands how patched evaluations may be constructed, and our objectives in writing and editing this book have been met.

READING

Attitude Change and Mental Hospital Experience

Jack M. Hicks
Northwestern University

Fred E. Spaner*
Veteran's Administration Hospital, Downey, Illinois

IT IS THE THESIS of this investigation that favorable attitudes toward the mentally ill may develop as a function of intimate exposure to the mental hospital environment. Specifically, attitudes of the hospital employees will shift over time in a direction more favorable to the patient. Empirical support of this notion is suggested by Bovard (1951), who found that as frequency of interaction between persons increases, the degree of liking for one another increases. His study, however, involved college classes and may not be generalizable to the mental hospital.

Impressionistic evidence of the positive effects of mental hospital experience is provided by extensive work with student nurses (Perlman & Barrell, 1958). These observers report that at the beginning of psychiatric training student nurses see patients as "maniacal" and tend to emphasize the differences between patients and themselves. But, as their training progresses, "they begin to maximize the similarities between themselves and patients and see the differences as predominantly differences in degree and not kind." Some experimental corroboration of these observations is offered by Altrocchi and Eisdorfer (1960).

Dissenting evidence of the positive effects of the hospital environment upon attitudes is suggested by Middleton (1953) who found greater prejudice toward mental patients among older, more experienced mental hospital employees than among younger, less experienced employees. However, this finding is difficult to interpret as evidence of attitude change due to the absence of pretest information.

In view of the paucity of systemic research and ambiguity of finding in this area, the present investigation was undertaken.

Experiment I

This study was an initial step toward further exploration of the effects of mental hospital

*The authors would like to express their indebtedness to Donald T. Campbell who critically read this paper and made many valuable suggestions.

experiences upon attitudes toward the mentally ill. Attitude modification was measured among student nurses exposed to 12 weeks of psychiatric nursing training at the Veterans Administration Hospital, Downey, Illinois. Attitudes elicited immediately prior to training were compared to attitudes immediately following training.

The following hypothesis was tested:

There will be a favorable shift in attitude toward mental patients over a 12-week period of psychiatric nursing training. Favorable attitude shift was defined as a change in response to attitude statements in the direction of judges' opinions of the most desirable attitudes for student nurses to have.

Psychiatric nursing training consisted of two main subdivisions: classroom instruction, and ward experience. Classroom instruction involved 10 hours per week devoted to exploration of nursing care problems. Ward experience entailed 6 weeks on each of two clinical areas—the male section and the female section—of the Acute and Intensive Treatment Service.

Method

Subjects. Seventy-eight student nurses, sent to Downey from several schools of nursing in the Chicago area as part of their regular nursing training schedule, were used. They were there in groups of 48 and 30, 3 months apart. All had received an average of about 2 years' previous "Nonpsychiatric" nursing training.

Materials. A five-point Likert questionnaire designed to elicit attitudes associated with mental patients was used. The scale contained 66 statements some of which were borrowed verbatim or modified from Gilbert and Levinson's (1956) CMI scale (Items 5, 7, 11, 15, 18, 19, 25, 31, and 38) and Middleton's (1953) "Prejudice Test" (Items 1, 2, 5, 10, 14, 17, 22, 23, 26, 34, 36, and 37). The remaining items were original with this study.

Design. The design for this study, presented in Table 1, was a "recurrent institutional cycles design" developed by Campbell and Stanley (1962). It is appropriate to those situations in which an institutional process (in this case, psychiatric nursing training) is presented continually to new groups of respondents. Group A was a posttest only aggregate whose only exposure to the questionnaire was on the final day of their 12-week psychiatric training period. Group B was a pretest-posttest aggregate. These respondents received the questionnaire on the first morning of their arrival into the psychiatric setting, prior to any formal orientation or exposure to patients (Group B_1). They also received an identical form of the questionnaire on the final day of their training period (Group B_2).

Procedure. First exposure: For the pretest and posttest only conditions, the experimenter was introduced to the group by a nursing education instructor. The experimenter then asked their cooperation in participating in a survey of attitudes and opinions about mental illness. Confidentiality of results was assured. Group B_1 were asked to sign their names on the questionnaire. Group A were asked not to record names.

Second exposure: Group B_2 respondents, having already been exposed to the test situation, were not given the introduction and full instructions. The group was simply asked to participate in filling out another questionnaire concerning attitudes and opinions about mental illness. They were reminded to put their names on the questionnaires. The same test room and same experimenter were used for all conditions.

Composite scoring key: A total attitude score was derived for each respondent by the method of summated ratings. The directionality of each item was determined by asking five nursing education instructors to respond to each statement in accordance with their professional opinions as to the most desirable

attitude for a student nurse to have. These ratings produced at least 80% interjudge agreement on 38 of the total 66 items. Item response categories were scored from zero through four, depending on the directionality of the statement.

Results and Discussion

The major statistical tests of the hypotheses shown in Table 1 were *t* tests of difference in mean composite scores between (a) Group B_1 (pretest) and Group B_2 (posttest), and (b) Group B_1 and Group A (posttest only).

The mean difference between B_1 and B_2 was 15.25. A direct-difference *t* test between these two means was highly significant and in the predicted direction ($p < .0005$). This finding strongly supported the hypothesis. The mean difference between B_1 and A was 4.5, reaching significance in the same direction as the B_1–B_2 difference ($p < .05$). Hence, support of the hypothesis was provided by both comparisons.

However, these results became less clear-cut in consideration of an unexpected divergence between the two posttest means. A *t* test between A and B_2 produced a convincingly significant difference ($p < .001$), with B2 higher than A. This finding seemed noteworthy despite the fact that a *t* test did not show the unpredicted A–B_2 difference to be significantly greater than the predicted A–B_1 difference ($p > .05$). There appeared to be several plausible explanations. One possibility stemmed directly from the fact that B_2 received a pretest, whereas A did not. That is, the pretest may have influenced performance on the subsequent testing. According to Campbell (1957), a pretest may increase or decrease sensitivity to the experimental variable. Thus, there may have been an interaction between the attitude pretest and psychiatric experience.

Another possible cause of the A-B_2 discrepancy was a respondent selection bias. Group B may have had a higher mean attitude than Group A at the outset of psychiatric training. This consideration became particularly important since the nature of the field situation prevented random preselection by the experimenter. Although psychiatric nursing trainee assignments were made unsystematically, and were not expected to vary from group to group, this assumption could only be made on a priori grounds at this time.

A third source of confoundment pertained to names being asked of Group B_2 but not Group A. This fact may have "freed" Group A respondents to more readily express less favorable attitudes actually felt. Group B_2, on the other hand, because of the presence of their names, may have been more inclined to present themselves in a favorable light, giving answers which they knew the staff thought desirable.

A fourth possibility concerned the sheer fact of taking a pretest sufficing to produce systematic changes on the posttest. This phenomenon differs from test-treatment interaction in that it refers to the main effects of testing per se.

Finally, it is possible that "history" effects of a powerful extraneous influence unknown to the experimenter occurred prior to B_2, but not prior to A. An example would be a mental health publicity campaign occurring contemporaneously with the Group A training pe-

TABLE 1
Comparisons of Composite Score Means

Comparison	*N*	*M*	*t*
B_1 (Pretest)	48	103.1	11.55***
B_2 (Posttest)	48	118.3	
A (Posttest only)	30	107.6	1.96*
B_1	48	103.1	
A	30	107.6	4.67**
B_2	48	118.3	

* Significant at .05 level, one-tailed.
** Significant at .001 level.
*** Significant at .0005 level, one-tailed.

riod. A related possibility is that psychiatric training itself differed in some important way for Groups A and B. It is certainly possible to imagine that some change in hospital personnel, goals, or other conditions, may have affected the nursing education program.

Experiment II

This investigation was an attempt to (a) corroborate and refine the basic findings of Experiment I, and (b) clear up ambiguities in the results of Experiment I. It will be recalled that a small but significant difference in the predicted direction was found between a posttest only and an independent pretest group, but that a highly significant difference between the two posttests created problems of interpretation. Basically, the same quasi-experimental design was used (institutional cycles design), but with additional controls and a modified measuring instrument. This design also resembles Campbell and Stanley's (1962) "nonequivalent control groups" design in that random preselection was not possible.

The primary advantage of Experiment II was in the introduction of control groups. This addition made two important contributions: First, it provided for the testing of a hypothesis more precisely related to the effects of psychiatric training. The hypothesis for Experiment I was limited to predicting "absolute" attitude shift over 12 weeks of psychiatric training. The study was not designed to differentiate between change as a function of psychiatric training, and change produced by other variables. As much change may conceivably have occurred with no psychiatric training at all. Equipped with control groups, Experiment II may test the more meaningful hypothesis of whether or not attitude change will be relatively greater as a function of psychiatric training than some other treatment. Hypothesis I for this investigation is:

There will be greater favorable shift in attitudes toward mental patients over a 12-week period of psychiatric nursing training than over an equivalent period of nonpsychiatric nursing training. (Attitude change was defined as in Experiment I.)

A second advantage of control groups permitted light to be cast on factors related to the unpredicted mean difference between the posttests of Experiment I. In particular, it was possible to test for pretest "sensitization" to psychiatric experience. For Experiment I, there were pretested and unpretested groups for the experimental treatment only. For Experiment II, there were pretested and unpretested groups for both experimental and control conditions. Due to the suggestive finding of Experiment I, the following hypothesis may be stated:

There will be a significant interaction between pretest attitude and psychiatric experience. This interaction should have a favorable rather than a dampening effect upon posttest attitudes.

An alternative possibility which can be tested as part of the analysis of the above hypothesis concerns the main effects of testing. In this consideration, pretested respondents would have significantly more favorable attitudes than unpretested respondents to a comparable degree on both experimental and control conditions.

Experiment II was also designed to test for recruitment differences among groups of respondents, another of the suggested confounders of Experiment I. This analysis was made by testing for nonlinearity among mean attitudes of all pretest conditions, both Experimental and Control.

It was also conjectured that the unpredicted difference between the posttests of Experiment I was due to instructions to record names on the questionnaires for one group but not the other. In order to check out this possibility, an additional experimental group was tested with "no name" instructions.

One further test designed to clarify Experiment I was considered. The reader will

recall the suggestion that the two posttest means differed due to "history" effects of extraneous influences occurring concomitantly with psychiatric training, or subtle changes in psychiatric training itself. Such confounders were controlled for in Experiment II by the inclusion of two pretest-posttest psychiatric groups differing in terms of a 3-month separation in experimental treatment. If either or both of these effects are manifest, differential attitude change between the two groups would be expected.

Method

Subjects. A total of 354 student nurses were used. Two hundred twenty-four were members of the experimental groups receiving 3 months of intensive psychiatric nursing training. The remaining 130 were control respondents who, not yet having reached the psychiatric phase of nursing training, were encountered somewhat earlier in their overall programs. There was considerable overlap in this respect, however, since entrance of students into psychiatric training was distributed fairly evenly over the second half of the 3-year registered nurses training program. The only prerequisite for psychiatric training was at least 18 months of previous general hospital training.

All Experimental subjects took their psychiatric training at the Veterans Administration Hospital, Downey, Illinois, as part of their normal nursing training schedule. They came in aggregates of approximately 50 at 3-month intervals from seven schools of nursing in the Chicago area. Control respondents were tested while undergoing pediatric, surgical, obstetrical, or other phases of nursing training at two of the seven home hospitals of the nursing schools mentioned above. Respondents were not randomly assigned by the experimenter.

Materials. The Experiment I questionnaire was used except for slight modifications in format and content. In terms of format, the six-point forced-choice Likert adaptation, employed by Cohen and Struening (1959) was adopted. This represented an elimination of the ? category used too liberally by Experiment I respondents, being replaced by two categories: not sure but probably agree, and not sure but probably disagree.

The content of the questionnaire was modified to the extent of eliminating 5 items and adding 26 new ones, making a total of 87 for the Experiment II questionnaire. The 5 eliminated items were unsatisfactory because of highly skewed response distributions. One of these items was Number 22 from Middleton's (1953) Prejudice Test. The addition of 26 items was an attempt for better representation of the attitude universe of interest. Of these items, 4 were borrowed from Middleton's Prejudice Test (Items 4, 18, 29, and 30), and 8 were taken from Cohen and Struening's (1959) Opinions about Mental Illness questionnaire (Items 8, 14, 34, 35, 39, 41, 42, and 48). The remaining 14 items were original with this study. A special feature of the revised questionnaire was the retention of 21 non-shifting items from Experiment I. These items were included as an indicator of the degree to which Experimental respondents discriminate among items in the process of demonstrating attitude change. Fourteen items which showed significant change in the same direction on both $B_1 - A$ and $B_1 - B_2$ comparisons in Experiment I were also retained.

Design. The 354 respondents made up eight conditions, five treatment conditions and three control conditions. Experimental groups are referred to in Table 2 as Groups A, B, F, G, and H. Groups C, D, and E are Control groups. The lettering, A through H, is related to the chronological order in which the different groups were encountered over a 9-month period. The Experimental groups were of two types: pretest-posttest and posttest only.

Groups A and H (comparable to Group A of Experiment I) were posttest only (or unpretested) whose only exposure to the questionnaire was on the final day of their respective 12-week psychiatric training periods. Groups A and H differed in that A was tested 9 months prior to H; and Group A respondents were asked to record their names on the questionnaires, while H respondents remained anonymous. Groups B and G were Experimental pretest-posttest aggregates, comparable to Group B of Experiment I. These subjects were pretested immediately upon arrival at Downey and prior to any of the scheduled activities (Conditions B_1 and G_1). They were posttested on an identical form of the questionnaire after completion of all scheduled psychiatric activities 12 weeks later (Conditions B_2 and G_2). A 3-month separation in the span of time (12 weeks) during which training was received denotes the principal difference between these two groups.

In order to determine the longer range effects of psychiatric experience upon attitude changes, an additional Experimental group (Group F) was tested. Of the 24 individuals in this group, 14 had participated previously in Experiment I and 10 in Group B of Experiment II, discussed above. These subjects ranged from 1 to 10 months of "post-psychiatric" general hospital training, with a mean of 5.25 months.

Control groups were also of two types: pretest-posttest and pretest only. There were two pretest only groups (C and E), so-called as a convenient label for "one-shot" testing administered at times prior to psychiatric training. Minor differences between Groups C and E were in terms of time and hospital at which testing took place. The final condition to be discussed pertains to Group D, a nonpsychiatric pretest-posttest aggregate. This group, representing the basic control group for testing of Hypothesis I, received 12 weeks of general hospital training between pretest and posttest.

Procedure. The identical procedure was used as described in Experiment I. The same experimenter presided over all sessions, with one exception. A substitute examiner was used for Condition G_2. Instructions were all the same. All groups were requested (on both pre-and posttest) to record their names in the upper right-hand corner of the questionaire except Group H. Group H was asked not to record names.

As in Experiment I, the method of summated ratings was used. Directionality of the new attitude statements was established by

TABLE 2
Experimental Design

Groups[a]							
A	B	C	D	E	F	G	H
Treatment Posttest	↓ Pretest Treatment Posttest	↓ Pretest	↓ Pretest ↓ Posttest	↓ Pretest	(Treatment) (Posttest) ↓ Post posttest	↓ Pretest Treatment Posttest	↓ Treatment Posttest[b]

a. In chronological order.
b. Respondents asked not to record names.

the same five nursing education instructors used in Experiment I. Items retained from Experiment I were not "rekeyed." Eighty percent interjudge agreement was achieved on 59 of the 87 items. Items were scored from one through six for analysis purposes.

Results

A post hoc measure of test-retest stability of the 59 keyed items was obtained by matching pretest-posttest scores for Control Group D. The Pearson coefficient was .82. Similar Pearson *r*'s were obtained between pretests and posttests of Experimental Groups B and G of .55 and .57, respectively.

Internal consistency of the keyed items was estimated as a by-product of average item-total correlations (Guilford, 1950, p. 494). Two independent matrices of item-total correlations were computed. Groups A, B, and G were combined , as well as Groups C, D, and E, yielding average item-total correlations of .34 and .30, respectively. Resulting internal consistency reliabilities were .88 and .86. The average item-total *r*'s were also squared to provide estimates of .11 and .09 mean item intercorrelations.

TABLE 3
Multiple Comparisons Test of Differential Attitude Shift

Comparison	*Difference*	$\overline{WSD}$
$(\bar{X}_{B_1} - \bar{X}_{B_2}) - (\bar{X}_{D_1} - \bar{X}_{D_2})$	16.71	9.37*
$(\bar{X}_{G_1} - \bar{X}_{G_2}) - (\bar{X}_{D_1} - \bar{X}_{D_2})$	16.57	8.81*
$(\bar{X}_{B_1} - \bar{X}_{B_2}) - (\bar{X}_{G_1} - \bar{X}_{G_2})$	.14	8.81

* Significant at .01 level.

	Date Tested 1960	*Mean*	*Test Condition*	*Treatment*
A	2–20	263.65	Post-only	Psych.
B_1	2–28	249.58	Pre	"
B_2	4–21	271.15	Post	"
C	3–11	247.36	Pre-only	General
D_1	3–11	252.83	Pre	"
D_2	6–10	257.69	Post	"
E	6–8	251.26	Pre	"
F	6–8	271.12	Post-Post	Psych.
G_1	4–30	247.84	Pre	"
G_2	8–20	269.27	Post	"
H	11–19	265.02	Post-only (no–name)	"

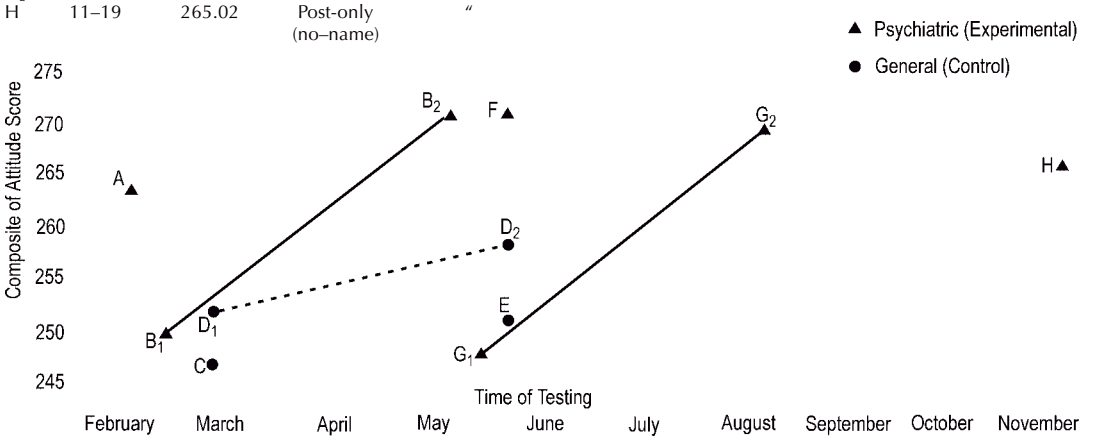


FIGURE 1
Mean attitudes and attitude shifts for all groups

An overall picture of mean attitudes and attitude shifts is presented in Figure 1. A cursory inspection reveals both markedly higher attitude scores for psychiatric groups than Control groups overall, and sharp increments in attitude favorability over the 12-week psychiatric training period. Mean attitude increments for Experimental Groups B and G were 21.57 and 21.43, respectively, both of which are highly significant ($p < .0005$, one-tailed), thus, strongly supporting the hypothesis of Experiment I.

The Tukey test for multiple comparisons, reviewed by Ryan (1959), was employed for the assessment of Hypothesis I. A three-way comparison was made between Experimental Groups B and G, and Control Group D in terms of degree of favorable attitude shift.

Groups B and G exceeded D in magnitude of favorable attitude shift by mean differences of 16.71 and 16.57, respectively. Both of these increments exceeded their respective "wholly significant difference" ($\overline{WSD}$) set at the .01 level of confidence indicated in Table 3. The third comparison between Groups B and G produced no significant difference, as expected. These comparisons unequivocally supported Hypothesis I as shown graphically in Figure 2.

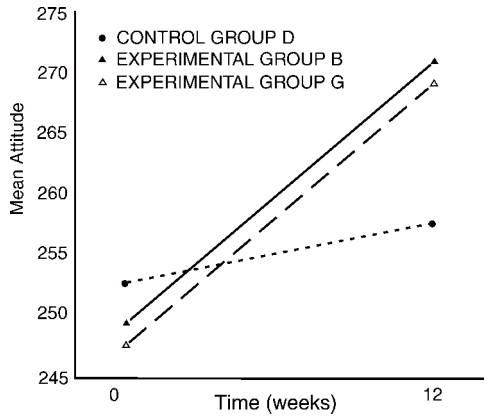


FIGURE 2
Mean attitude shift over 12-week period

TABLE 4
Analysis of Variance of Mean Attitude Differences as a Function of Testing and Treatment

Source	*df*	*MS*	*F*
Pretest	1	3932.5	11.00*
Treatment	1	14,543.3	40.68*
Interaction	1	140.2	.39
Error (w)	326	357.5	

* Significant at .001 level.

Pretest-treatment interaction was tested by a fourfold analysis of variance for unequal and disproportionate cell frequencies. Experimental versus Control groups made up the row main effects, whereas pretested versus unpretested groups were represented in the columns. As noted in Table 4, and also graphically in Figure 3, there was no significant interaction between pretesting and psychiatric nursing experience ($F < 1$). Therefore, Hypothesis II was rejected in that there was no evidence of pretest sensitization to the experimental variable.

Table 4 does present, however, strong evidence of the main effects of testing. Mean posttest attitudes of pretested respondents were found to be significantly higher than corresponding mean attitudes of posttest only subjects ($p < .001$).

The significant main effects of the experimental variable are also shown in Table 4. Overall mean posttest attitudes of psychiatric nurses are shown to be significantly more favorable than posttest attitude scores of the nonpsychiatric groups. This analysis does not reflect attitude *shift*, however, and does not bear directly on Hypothesis I.

TABLE 5
Analysis of Variance of Recruitment Differences

Source	*df*	*MS*	*F*
Between groups	4	252.5	<1
Error	234	338.9	

Results of a one-way analysis of variance are presented in Table 5, testing for nonlinearity among group mean pretest attitudes. The between-groups variance was less than the within-groups variance, hence rejecting the nonlinearity hypothesis ($F < 1$) regarding recruitment nonequivalence.

The hypothesized difference between "names" and "no names" was tested by comparing the attitude means of posttest only Experimental Groups A and H. The means for the two groups were 263.65 and 265, respectively. An uncorrelated t was < 1, failing to support the hypothesis.

The effects of "history" and/or undetected changes in the experimental variable were tested by comparing the mean shifts of two psychiatric groups, B and G. A test for this hypothesis was conveniently provided by Table 3. As previously mentioned, the mean difference (.14) in shifts between Groups B and G is far below the $\overline{\text{WSD}}$ of 8.81. Thus, no history effects or changes in psychiatric training appear evident.

The testing of 24 "post-posttest" respondents (Group F) made it possible to estimate the stability of attitude change. Since most of these respondents were previous members of psychiatric pretest-posttest conditions, they were compared accordingly with pretest-posttest Groups B and G. The post-posttest mean attitude of Group H was 271.12, a level remarkably comparable to a posttest mean of 271.15 for Group B and 269.27 for Group G. The durability of attitude change, at least for a period of several months, seems clearly demonstrated.

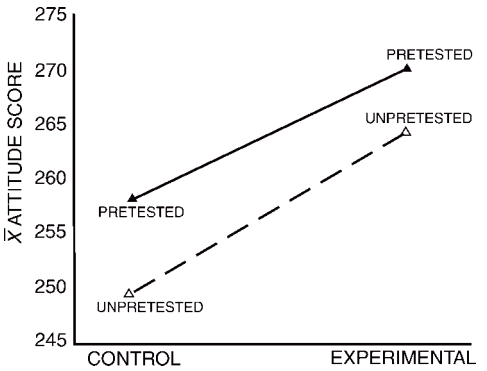


FIGURE 3
Conjunctive effects of testing and treatment on posttest scores

A comparison of degree of shifts was also made between shifting and nonshifting items of Experiment I. The criterion of item shift on Experiment II was a significant mean difference on the B_1–B_2 comparison. A Yates-corrected x^2 was a nonsignificant 1.73 ($p > .05$). Hence, items which shifted on Experiment I did not shift to a significantly greater extent on Experiment II than items which did not shift on Experiment I.

Discussion

Of the several additional subsidiary hypotheses concerning the unpredicted difference between the posttests of Experiment I, only that which concerned pretest main effects was supported. It appears, therefore, that the significantly higher mean attitude of Condition B as compared to Condition A of Experiment I may be traced to transfer effects from Pretest Condition B with Group A not being so benefited by a pretest "booster." This interpretation does not follow from two investigations of attitude change by Lana (1959a, 1959b), who found neither pretest sensitization nor testing effects. A third study by Lana and King (1960), however, did show testing main effects without test-treatment interaction when a learning task was used. These results are directly in line with those of this investigation, suggesting the presence of a learning factor in the attitude questionnaire used.

In conclusion, then, it may be stated with confidence that 12 weeks of psychiatric training was demonstrated to be unequivocally effective in producing attitude change toward

mental illness over and above the noise level of pretest transfer effects. Furthermore, the absence of troublesome interactions strengthens the generalizability of these findings beyond this particular experimental arrangement.

Summary

Two quasi-experiments were performed as part of an on-going project to investigate the effects of overall mental hospital experience upon modification of attitudes toward the mentally ill. Four hundred thirty-two student nurses served as respondents over all. Attitudes were measured by the Likert technique before and after 12 weeks of psychiatric nursing training. In Experiment I, attitudes on both posttests were observed to be significantly higher in a favorable direction than a pretest. A complication arose, however, with respect to an unpredicted significant difference between the posttests. Experiment II was basically a refinement of Experiment I. Control groups were added to the design, the questionnaire was revised, and a larger *N* used. Control respondents were student nurses in general hospitals not as yet having reached the psychiatric phase of their training. The basic hypothesis that psychiatric subjects would make a significantly greater favorable shift in attitudes over a 12-week period than nonpsychiatric subjects was strongly supported. The evidence suggests that the posttest discrepancy in Experiment I was probably due to pretest main effects, with the "pretested" posttest condition manifesting a higher mean attitude than the unpretested group.

References

Altrocchi, J., & Eisdorfer, C. Changes in attitudes toward mental illness. Unpublished manuscript, 1960.

Bovard, E. W. Group structure and perception. *J. Abnorm. Soc. Psychol.*, 1951, 46, 398–405.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, 1957, 54, 297–312.

Campbell, D. T., & Stanley, J. C. Experimental designs for research in teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand-McNally, 1962, in press.

Cohen, J., & Struening, E. L. Factors underlying opinions about mental illness in the personnel of a large mental hospital. Paper read at American Psychological Association, Cincinnati, 1959.

Gilbert, Doris, & Levinson, D. J. Ideology, personality, and institutional policy in the mental hospital. *J. Abnorm. Soc. Psychol.*, 1956, 53, 263–271.

Guilford, J. P. *Fundamental statistics in psychology and education.* (2nd ed.) New York: McGraw-Hill, 1950.

Lana, R. E. A further investigation of the pretest-treatment interaction effect. *J. Appl. Psychol.*, 1959, 43, 421–422. (a)

Lana, R. E. Pretest-treatment interaction effects in attitudinal studies. *Psychol. Bull.*, 1959, 56, 293–300. (b)

Lana R. E., & King, D. J. Learning factors and determiners of pretest sensitization. *J. Appl. Psychol.*, 1960, 44, 189–191.

Middleton, J. Prejudices and opinions of mental hospital employees regarding mental illness. *Amer. J. Psychiat.*, 1953, 10, 133–138.

Perlman, M., & Barrell, L. M. Teaching and developing nurse-patient relationships in a psychiatric setting. *Psychiat. Quart. Suppl.*, 1958, 31, 1–13.

Ryan, T. A. Multiple comparisons in psychological research. *Psychol. Bull.*, 1959, 56, 26–47.

(Received June 3, 1961)

Explanation and Critique

In the face of conflicting research results concerning the effect of student nurses' hospital experience on attitudes toward the psychiatric patients, Jack Hicks and Fred Spaner sought to systematically assess the changes in attitudes of nurses that occurred between their entering the psychiatric training program and their leaving it. Hicks and Spaner used the *institutional cycles design,* which is appropriate when an institutional process (in this instance, psychiatric nursing training) is presented continuously to new groups. This would be appropriate in a number of instances. For instance, any training process—student teaching, geriatric nursing training, apprenticeship programs—would be a candidate.

But what actually is this design? It is merely a label attached to a design that patches together aspects of the Solomon Four-Group Design and the pretest-posttest comparison group designs. The label is instructive to point out instances when the use of this patchwork makes sense. In such a design, a number of comparisons can be made—as Hicks and Spaner do. Gathering data in this manner allows a researcher to make comparisons between pretest and posttest scores ($B_1 - B_2$), between pretest and posttest-only scores ($B_1 - A$), and between posttested scores ($B_2 - A$).

Hicks and Spaner use sixty-six Likert items scored ordinally (e.g., a five-point scale in which 1 was unfavorable and 5 was favorable), which were drawn from existing indexes or were modifications of them. (This is a common practice. There is a whole body of testing literature that deals specifically with measuring the validity and reliability of measures and scales. You should use a tried-and-true measure rather than reinventing the wheel with each new research endeavor. In addition, it adds to the cumulative nature of this research.) "Favorable" attitudes were determined by a panel of judges who examined possible attitude items and ranked answers from "favorable" to "not favorable" toward patients. Use of expert opinions was reasonable in this instance or in any instance in which there are no existing measures which perfectly match the study's needs. The attitude index was a summation of the responses to these items after all items were recoded to reflect the same scale.

The subjects were nurses on two psychiatric training rotations. They were judged to be similar on a number of characteristics; for example, all had two years' nursing training. Random assignment was impossible because of the nature of the timing of the rotation. Nurses were not systematically or randomly assigned to class cohorts. They entered the nursing school when they wished. Thus, the investigators had no reason to believe there were systematic differences between the cohorts.

Using a difference of means test (*t*-test), Hicks and Spaner compare the pre- and posttest results. If training were to have a positive effect on favorable attitudes, the B_2 and A mean scores would have to be significantly higher than the B_1 mean score. That, in fact, is what they found. To rule out any effects of the pretest on the posttest results, the A and B_2 mean scores would have to be the same (not significantly different from each other). The latter result did not occur. Not only are the A and B_2 scores different, but the B_2 score is higher. The substantive meaning of this is that the pretested nurses had favorable scores that were even higher than those without the pretest but with the same treatment (training).

These results troubled Hicks and Spaner. Rather than being satisfied that the training appeared to have a positive impact, they sought to explain the differences between the A and B_2 scores. The plausible explanations revolved around threats to internal validity, mostly the effect of testing. The various threats resulting from testing and what they

suggest are interesting. First, what were the effects of the pretest on the posttest response? This is what one typically asks when doubtful about testing effects. Second, did the pretest experience in any way interact with the psychiatric training experience to, in this case, enhance the favorable result? This interaction effect is above and beyond the main effects postulated in the first example. So, was it the pretest that affected the result, or did the pretest do something to enhance the training effects? Did the pretest sensitize the Group B nurses during the training to make them react more favorably to the patients? This question was compounded by the fact that the responses by the Group B nurses were not confidential. This situation cannot be helped when one needs to link the responses at B_1 to those at B_2 without using a numbered coding scheme (which might have helped). Did this lack of confidentiality result in nurses' answering questions in line with what they believed their evaluators would regard as favorable? Even though Hicks and Spaner postulated no systematic differences between the two groups, could any selection bias be explored? Finally, did anything (i.e., history) occur prior to B_2 that did not occur before A?

Typically there is no easy way to evaluate these other effects after the fact; they are often estimated if addressed at all. The beauty of an institutional cycles design, however, is that there will be new entrants into the cycle and the effects can be measured. The nursing rotation allowed Hicks and Spaner to redesign the initial experiment with the result being Experiment II.

Hicks and Spaner physically controlled for other explanatory variables by adding nonequivalent control groups (see Part III). The members of the basic control groups were nurses who were in a nonpsychiatric rotation and who had not completed a psychiatric rotation. These nurses had slightly less hospital experience than those in the treatment groups. Within the treatment groups, there were also control groups to control for confidentiality and time. The attitude instrument was also modified slightly on the basis of the results in Experiment I. By forcing respondents to make a favorable versus unfavorable decision (i.e., by removing the "don't know/no opinion" category), Hicks and Spaner increased the variation in the attitudinal score. In other words, they reduced the propensity to give a neutral response, which really is not unique. Apparently Hicks and Spanner believed that nurses would have some feeling rather than no feeling regarding these attitudes. This is very similar to the idea among some political scientists that there are fewer pure independent voters than Republican and Democratic leaners. The leaners are not allowed to hide their leanings under the cloak of fashionable independence.

Systematically, Hicks and Spaner constructed control groups to test each of the possible threats suggested in Experiment I. Their table 2 displays the groups in terms of their experimental status and the timing of the test. Although the experiment is designed to verify the findings in Experiment I, Hicks and Spaner also did a good job of showing how to test the validity threats.

The findings are clearly presented and very straightforward. (Unfortunately, most evaluation results are not so clear cut!) First, did the treatment favorably affect attitudes? The results in both table 3 and figure 2 indicate that attitudes improved with the training program. The pretest scores are similar, and the change in the scores between the pre- and posttest indicates that the treatment versus control change is significant. In addition, the effects across the two treated groups appear to be similar, and their respective differences from the control group are similar. The pattern of the attitude change does not change over time (the three months between groups B and G).

The fourfold analysis of variance in table 4 of the article was used to determine whether

and which testing effects were operative. Hicks and Spaner find no significant interaction between the pretest and the treatment, but there were consistent and significant effects of the pretest results as compared with the posttest results. There is an effect of taking the pretest and more favorable posttest scores for both the experimental and control groups.

Was there a recruitment effect? If there were a difference, it would show up on the pretest scores. Hicks and Spaner performed an analysis of variance to see if the scores on the pretests were different among the groups. The idea was that if the difference between groups was greater than the difference within the groups, then there were systematic differences between the groups (i.e., the groups were not equivalent or similar initially). Hicks and Spaner found more variation within the groups than between them. Thus, they conclude that there is no systematic variation—that is, no selection bias. This does not mean that there were no differences between the groups in Experiment I. Because of the results in Experiment II, Hicks and Spaner assumed there were no selection effects in Experiment I.

Did lack of confidentiality enhance posttest scores? To determine the answer, Hicks and Spaner compared groups that received the treatment and posttest-only attitude test. The groups differed only regarding whether they were asked to identify themselves on the instrument. Thus, the testing effect was controlled for; no pretest was given. Again, they found no difference in the posttest scores, which indicated that there was no enhancement based on pleasing the examiner. The inference is that it did not make a difference in Experiment I either. Rather, the difference that emerged in Experiment I was due to the main testing effects.

One of the reasons Hicks and Spaner had two pretest-posttest groups was to determine whether history played a part in the differences between A and B_2. Apparently all the posttested As were in one class cohort and the Bs were in another. In Experiment II the two pretest-posttest groups (B and G) were from different cohorts that began their training three months apart. By looking at the mean shifts in both groups, Hicks and Spaner concluded that history was not a factor because there were no significant differences between the shifts. History was not a contributing factor across that three-month period.

This finding is bolstered by the fact that the post-posttest results from group F were similar to those of B and G. Not only did there appear to be no history effect across those three periods, but the results indicated that the effect was fairly durable. The authors did not indicate how they chose the nurses from group F. Perhaps they tried to follow up on all, used participants from one participating school or hospital, or chose convenient subjects. The threat to validity they did not address in this aspect of the study was experimental mortality.

One of the neat and orderly aspects of this design was that with each new nursing cohort, the researchers were able to have a clean slate with which to work (i.e., untreated nurses with similar pretraining), so the experiments could be done over and over again. This luxury is not normally afforded evaluators. When the opportunity is presented, however, it allows one to take a second shot at the research and improve on the initial design.

Reference

Campbell, Donald T., and Julian C. Stanley. 1962. "Experimental Design for Research in Teaching." In *Handbook of Research on Teaching,* edited by N.L. Gage. Chicago: Rand McNally, 171–246.

CHAPTER 16

Meta-Evaluation Designs

Meta-evaluation studies use statistical procedures to combine two or more empirical studies that report findings on the same or very similar programs (Hunter and Schmidt 1990). Conceptually, meta-evaluations are similar to literature reviews in that they attempt to make sense out of a body of literature developed in different contexts. Unlike literature reviews, however, meta-evaluations compare and combine empirical results using statistical techniques.

The key to meta-evaluations is the calculation of the *effect size*. The effect size computation is a "standardization process through which the strength of the X-Y [independent variable, dependent variable] relationship is expressed as standard deviation units" (Bowen and Bowen 1998, 74) so that the results can be integrated across studies. In program or policy evaluations, the effect size is the difference between experimental and control group means divided by their pooled standard deviations. Sample size variation across studies can be controlled by weighting the individual means by multiplying them by the inverse of the variance. Effect size can be calculated for individual dependent variable scores (when there is more than one measure of the variable), which also can be pooled to report composite scores. Effect scores are typically reported showing the weighted mean of each variable of interest and its 95 percent confidence interval. Effect size can be either negative or positive; however, if a confidence interval overlaps with 0, it has no statistically significant effect. Variables for which the confidence intervals do not contain 0 are considered statistically significant effects, although their magnitude should be scrutinized as well.[1]

The first step in performing a meta-evaluation is to clearly define the type of program and effects you are interested in comparing and then do an exhaustive search for both published and unpublished empirical evaluations. Meta-evaluations report the bibliographic information for all studies discovered that were ultimately used. This is important for anyone who wishes to determine the appropriateness of your sample. The specific criteria used to include or exclude studies must also be reported. A number of criteria are used to include studies or to exclude them. For instance, you want to include studies that implement the same program or same kind of program. You want to include studies that measure the same variable of interest in the same way. You might exclude studies that have a small sample, that are not

methodologically sophisticated, or in which the program is implemented in a setting that is out of the ordinary (e.g., implementation of a drug rehabilitation program inside prison as opposed to one for the general public).

In the following article, "How Effective Is Drug Abuse Resistance Education?" Susan Ennett and her colleagues perform a meta-evaluation on results from evaluation studies of the widely implemented DARE program.

READING

How Effective Is Drug Abuse Resistance Education?

A Meta-Analysis of Project DARE Outcome Evaluations

Susan T. Ennett, *PhD* Nancy S. Tobler, *MS, PhD*
Christopher L. Ringwalt, *DrPH* Robert L. Flewelling, *PhD*

Abstract

Objectives. Project DARE (Drug Abuse Resistance Education) is the most widely used school-based drug use prevention program in the United States, but the findings of rigorous evaluations of its effectiveness have not been considered collectively.

Methods. We used meta-analytic techniques to review eight methodologically rigorous DARE evaluations. Weighted effect size means for several short-term outcomes also were compared with means reported for other drug use prevention programs.

Results. The DARE effect size for drug use behavior ranged from .00 to .11 across the eight studies; the weighted mean for drug use across studies was .06. For all outcomes considered, the DARE effect size means were substantially smaller than those of programs emphasizing social and general competencies and using interactive teaching strategies.

Conclusions. DARE's short-term effectiveness for reducing or preventing drug use behavior is small and is less than for interactive prevention programs. (*Am J Public Health.* 1994; 84:1394–1401)

Susan T. Ennett, Christopher L. Ringwalt, and Robert L. Flewelling are with the Research Triangle Institute, Research Triangle Park, NC. Nancy S. Tobler is with the State University of New York, at Albany, NY.

Requests for reprints should be sent to Susan T. Ennett, PhD, Center for Social Research and Policy Analysis, Research Triangle Institute, PO Box 12194, Research Triangle Park, NC 27709-2194.

This paper was accepted February 15, 1994.

Note. The views expressed here are the authors' and do not necessarily represent the official position of the US Department of Justice or the US Department of Health and Human Services.

Introduction

SCHOOL-BASED DRUG USE prevention programs have been an integral part of the US antidrug campaign for the past two decades.[1,2] Although programs have proliferated, none is more prevalent than Project DARE (Drug Abuse Resistance Education).[3] Created in 1983 by the Los Angeles Police Department and the Los Angeles Unified School District, DARE uses specially trained law enforcement officers to teach a drug use prevention curriculum in elementary schools[4] and, more recently, in junior and senior high schools. Since its inception, DARE has been adopted by approximately 50% of local school districts nationwide, and it continues to spread rapidly.[3] DARE is the only drug use prevention program specifically named in the 1986 Drug-Free Schools and Communities Act. Some 10% of the Drug-Free Schools and Communities Act governor's funds, which are 30% of the funds available each fiscal year for state and local programs, are set aside for programs "such as Project Drug Abuse Resistance Education,"[5] amounting to much of the program's public funding.

Given its widespread use and the considerable investment of government dollars, school time, and law enforcement effort, it is important to know whether DARE is an effective drug

use prevention program. That is, to what extent does DARE meet its curriculum objectives, most prominently "to keep kids off drugs"?

DARE's core curriculum, offered to pupils in the last grades of elementary school, is the heart of DARE's program and the focus of this study. We evaluate here the core curriculum's short-term effectiveness by using meta-analytic techniques to integrate the evaluation findings of several studies.[6,7] We searched for all DARE evaluations, both published and unpublished, conducted over the past 10 years and selected for further review those studies that met specified methodological criteria. We calculated effect sizes as a method for establishing a comparable effectiveness measure across studies.[7–9] In addition, to put DARE in the context of other school-based drug use prevention programs, we compared the average magnitude of the DARE effect sizes with those of other programs that target young people of a similar age.

DARE's *Core Curriculum*

The DARE core curriculum's 17 lessons, usually offered once a week for 45 to 60 minutes, focus on teaching pupils the skills needed to recognize and resist social pressures to use drugs.[4] In addition, lessons focus on providing information about drugs, teaching decision-making skills, building self-esteem, and choosing healthy alternatives to drug use.[4] DARE officers use teaching strategies, such as lectures, group discussions, question-and-answer sessions, audiovisual material, workbook exercises, and role-playing.[4]

The training that DARE officers receive is substantial. They are required to undergo 80 training hours in classroom management, teaching strategies, communication skills, adolescent development, drug information, and curriculum instruction.[4] In addition, DARE officers with classroom experience can undergo further training to qualify as instructors/mentors.[4] These officers monitor the program delivery's integrity and consistency through periodic classroom visits.

Methods

Identification of Evaluations

We attempted to locate all quantitative evaluations of DARE's core curriculum through a survey of DARE's five Regional Training Centers, computerized searches of the published and unpublished literature, and telephone interviews with individuals known to be involved with DARE. Eighteen evaluations in 12 states and one province in Canada were identified. Several evaluations were reported in multiple reports or papers. (See Appendix A for a bibliography of the studies considered.)

Evaluation Selection Criteria

To be selected for this meta-analysis, an evaluation must have met the following criteria: (1) use of a control or comparison group; (2) pretest-posttest design or posttest only with random assignment; and (3) use of reliably operationalized quantitative outcome measures. Quasi-experimental studies were excluded if they did not control for preexisting differences on measured outcomes with either change scores or covariance-adjusted means.[10] In addition, to ensure comparability, we focused on results based only on immediate posttest. Because only four evaluation studies were long term (two of which were compromised by severe control group attrition or contamination), we were unable to adequately assess longer-term DARE effects.

We examined several other methodological features, such as the correspondence between the unit of assignment and analysis, the use of a panel design, matching of schools in the intervention and control conditions, and attrition rates. Although these factors were considered in assessing the studies' overall methodological rigor, we did not eliminate evaluations on the basis of these criteria.

Data Analysis

For each study, we calculated an effect size to quantify the magnitude of DARE's effectiveness with respect to each of six outcomes that reflect the DARE curriculum's aims. An effect size is defined as the difference between the intervention and the control group means for each outcome measure, standardized by dividing by the pooled standard deviation [effect size = $\text{mean}_I - \text{mean}_C/\text{SD}$].[7–9] If means and standard deviations were not available, we calculated effect sizes using formulas developed to convert other test statistics and percentages to effect sizes.[9] In all cases, we used statistics reflecting covariance-adjusted means, with pretest values as covariates rather than unadjusted means so that any differences between the comparison groups before the intervention would not be reflected in the effect sizes.[10]

The six outcome measure classes include knowledge about drugs, attitudes about drug use, social skills, self-esteem, attitude toward police, and drug use. Some studies did not include all six, and some outcomes were measured by more than one indicator. When multiple indicators were used (e.g., two measures of social skills), we calculated separate effect sizes and then averaged them.[6,10] This procedure yielded one effect size per study for each measured outcome type. In the one study that reported only that a measured outcome was not statistically significant (and did not provide any further statistics), we assigned a zero value to that effect size.[10] To calculate effect sizes for drug use, we considered only alcohol, tobacco, and marijuana use; we averaged effect sizes across these substances. In a supplementary analysis, we considered use of these substances separately. The prevalence of other drugs, such as cocaine, was too small to produce meaningful effects.

In addition to calculating one effect size per outcome per study, we calculated the weighted mean effect size and 95% confidence interval (CI) for each outcome type across programs. The weighted mean is computed by weighting each effect size by the inverse of its variance, which is a reflection of the sample size.[8,9] The effect size estimates from larger studies are generally more precise than those from smaller studies.[8] Hence, the weighted mean provides a less biased estimate than the simple, unweighted mean because estimates from larger samples are given more weight. The 95% CI indicates the estimated effect size's accuracy or reliability and is calculated by adding to or subtracting from the mean 1.96 multiplied by the square root of 1 divided by the sum of the study weights.[8]

Comparison of DARE with Other Drug Use Prevention Programs

For comparison with DARE, we used the effect sizes reported in Tobler's meta-analysis of school-based drug use prevention programs.[10] To allow the most appropriate comparisons with DARE effect sizes, we obtained Tobler's results for only those programs (excluding DARE) aimed at upper elementary school pupils. These programs are a subset of 25 from the 114 programs in Tobler's meta-analysis, whose studies are referenced in Appendix B.

We selected this meta-analysis for comparison because of its greater similarity to ours than other meta-analyses of drug use prevention programs.[11–14] Tobler's studies

TABLE 1
DARE Evaluation Studies Selected for Review

Location	*References* [a]
British Columbia (BC)	Walker 1990
Hawaii (HI)	Manos, Kameoka and Tanji 1986
Illinois (IL)	Ennett et al. 1994 (in press)
Kentucky-A (KY-A)	Clayton et al. 1991a, 1991b
Kentucky-B (KY-B)	Faine and Bohlander 1988, 1989
Minnesota (MN)	McCormick and McCormick 1992
North Carolina (NC)	Ringwalt, Ennett, and Holt 1991
South Carolina (SC)	Harmon 1993

a. See Appendix A for full references.

met the same methodological standards that we used for the DARE studies. The only differences were that Tobler excluded studies that did not measure drug use and considered results from later posttests, whereas we considered only immediate posttest results. Neither of these differences, however, should seriously compromise the comparison.

The evaluation studies included in Tobler's meta-analysis are classified into two broad categories based on the programs' content and process. Process describes the teaching approach (how the content is delivered). Programs classified by Tobler as "noninteractive" emphasize intrapersonal factors, such as knowledge gain and affective growth, and are primarily delivered by an expert. "Interactive" programs emphasize interpersonal factors by focusing on social skills and general social competencies and by using interactive teaching strategies, particularly peer to peer. Consistent with other meta-analyses showing that programs emphasizing social skills tend to be the most successful,[11–13,15] interactive programs produced larger effect sizes than noninteractive programs. We compared DARE with both categories of programs.

Results

Characteristics of Evaluations

Of the original 18 studies, 8 met the criteria for inclusion. One additional study met the methodological criteria but did not administer the first posttest until 1 year after DARE implementation; therefore, it could not be included in our analysis of immediate effects.[16,17] The location and primary reference for each evaluation are shown in Table 1, and study characteristics are summarized in Table 2.

Each evaluation represents a state or local effort. The number of student subjects in all studies was large, each study comprising at least 10 schools with approximately 500 to 2000 students. Although demographic information was not given for three studies, the

TABLE 2
Sample and Methodological Characteristics of the DARE Evaluations (n = 8)

Study	*Schools n*	*Subjects n*	*Research Design*	*Matching*	*Unit of Analysis*	*Pretest Equivalency*[a]	*Scale Reliabilities*	*Attrition*
BC	11	D = 287 C = 175	Quasi, cross-sectional	Yes	Individual	Yes	No	Not applicable
HI	26	D = 1574 C = 435	Quasi, panel	No	Individual	No	No	No
IL	36	D = 715 C = 608	Experimental/quasi, panel	Yes	School based	Yes	Yes	Yes[b]
KY-A	31	D = 1438 C = 487	Experimental, panel	No	Individual	Yes	Yes	Yes[b]
KY-B	16	D = 451 C = 332	Quasi, panel	Yes	Individual	Yes	Yes	No
MN	63	D = 453 C = 490	Quasi, panel	No	Individual	Yes	Yes	Yes[c]
NC	20	D = 685 C = 585	Experimental, panel	No	School based	Yes	Yes	Yes[b]
SC	11	D = 295 C = 307	Quasi, panel	Yes	Individual	Yes	Yes	Yes[c]

Note: See table 1 for information on study locations and references. D = DARE; C = comparison.
a. Pretest equivalency on demographic variables assessed and controlled if necessary.
b. Attrition rates reported and differential attrition across experimental conditions analyzed.
c. Attrition rates reported only.

TABLE 3
Unweighted Effect Sizes Associated with Eight DARE Evaluations

Study	Knowledge	Attitudes about Drugs	Social Skills	Self-Esteem	Attitude toward Police	Drug Use*
BC	.68	.00	—	—	—	.02
HI	—	.07	.34	—	—	—
IL	—	.03	.15	.15	.12	.05
KY-A	—	.11	.10	.07	—	.00
KY-B	.58	.19	.30	.14	.27	—
MN	.19	.06	.08	-.03	.05	—
NC	—	.19	.17	.00	—	.11
SC	—	.32	.19	.06	.08	.10

Note: See Table 1 for information on study locations and references.
*Limited to alcohol, tobacco, and marijuana.

remaining five studies in the sample primarily consisted of White subjects.

Assignment of DARE to intervention and control groups was by school for all eight studies. In one study, DARE also was assigned by classroom in certain schools.[18] Because of potential contamination in this study of the control group classrooms by their close proximity to DARE classes, we eliminated these control classrooms; only control schools with no DARE classes were included. Two studies used a true experimental design in which schools were randomly assigned to DARE and control conditions; a third study used random assignment for two thirds of the schools. The remaining five evaluations used a nonequivalent control group quasi-experimental design.

Because there were relatively few sampling units across studies—ranging from 11 to 63 schools, with all except one study involving fewer than 40 schools—it is unlikely that equivalence between groups was obtained without prior matching or blocking of schools, even with randomization. Only half the studies matched comparison schools on selected demographic characteristics. Most studies (75%), however, assessed the equivalency of the comparison groups at pretest and made adjustments for pretest differences on demographic characteristics. All studies adjusted for pretest differences on outcome measures.

All but one study used a panel design that matched subjects from pretest to posttest with a unique identification code.

Outcome measures used in the DARE evaluations were based on responses to self-administered questionnaires. Seven studies used standardized scales or revised existing measures; six studies reported generally high scale reliabilities (usually Cronbach's alpha). Validity information, however, was rarely reported, and no study used either a biochemical indicator or "bogus pipeline" technique to validate drug use self-reports.[19]

Most studies (75%) did not use a data analysis strategy appropriate to the unit of assignment. Because schools, not students, were assigned to DARE and control conditions, it would have been appropriate to analyze the data by schools with subjects' data aggregated within each school or to use a hierarchical analysis strategy in which subjects are nested within schools.[20,21] Six studies ignored schools altogether and analyzed individual subjects' data, thereby violating the statistical assumption of independence of observations. Ignoring schools as a unit of analysis results in a positive bias toward finding statistically significant program effects.[21] This bias may

be reflected in CIs reported for each outcome's weighted mean effect size.

Five studies reported generally small attrition rates. None of the three studies that analyzed attrition found that rates differed significantly across experimental and control conditions. In addition, subjects absent from the posttest were not more likely to be drug users or at risk for drug use. Although attrition usually is greater among drug users,[22] given the sample's young age (when school dropout is unlikely and drug use prevalence is low), these results are not surprising.

DARE Effect Sizes

Study effect sizes are shown in Table 3. In general, the largest effect sizes are for knowledge and social skills; the smallest are for drug use.

Figure 1 shows the mean weighted effect size and 95% CI for each outcome based on the eight studies combined. The largest mean effect size is for knowledge (.42), followed by social skills (.19), attitude toward the police (.13), attitudes about drug use (.11), self-esteem (.06), and drug behavior (.06). The effect sizes for knowledge, social skills, attitude toward the police, attitudes about drug use, and self-esteem are statistically significant. The CI for the mean drug use effect size overlaps with zero (i.e., it is not significantly different from zero).

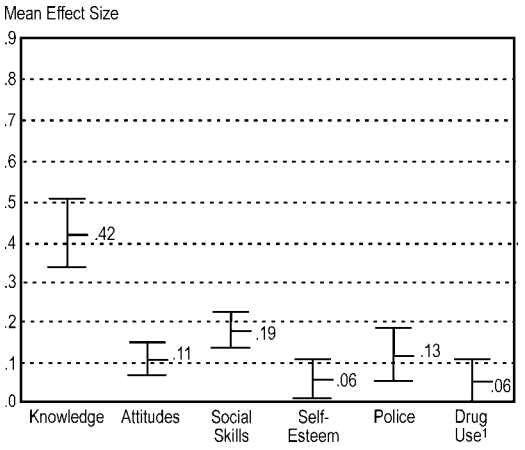


FIGURE 1

Magnitude of DARE's weighted mean effect size (and 95% CI), by outcome measure

1. Drug use includes alcohol, tobacco, and marijuana.

Because averaging alcohol, tobacco, and marijuana use for the drug use effect size could obscure substantial differences among the substances, we calculated DARE's mean weighted effect sizes separately for these substances. The weighted mean effect size for alcohol use is .06 (95% CI = .00, .12); for tobacco use, .08 (95% CI = .02, .14); and for marijuana use, -.01 (95% CI = -.09, .07). Only the mean for tobacco use is statistically significant.

Mean Effect Sizes for DARE vs Other Drug Use Prevention Programs

We compared by type of outcome the mean weighted DARE effect size with the mean weighted effect size for noninteractive (n = 9) and interactive (n = 16) programs; effect sizes for the comparison programs are derived from Tobler.[10] The comparison programs target youth of the same grade range targeted by DARE. The outcomes assessed by both DARE and the comparison programs are knowledge, attitudes, social skills, and drug use behavior.

Across the four outcome domains, DARE's effect sizes are smaller than those for interactive programs (Figure 2). Most notable are DARE's effect sizes for drug use and social skills; neither effect size (.06 and .19, respectively) is more than a third of the comparable effect sizes for interactive programs (.18 and .75, respectively). DARE's effect size for drug use is only slightly smaller than the noninteractive programs' effect size. DARE's effect size for knowledge, attitudes, and social skill, however, are larger than those for noninteractive programs.

Comparison of effect sizes separately for alcohol, tobacco, and marijuana use shows that DARE's effect sizes are smaller than

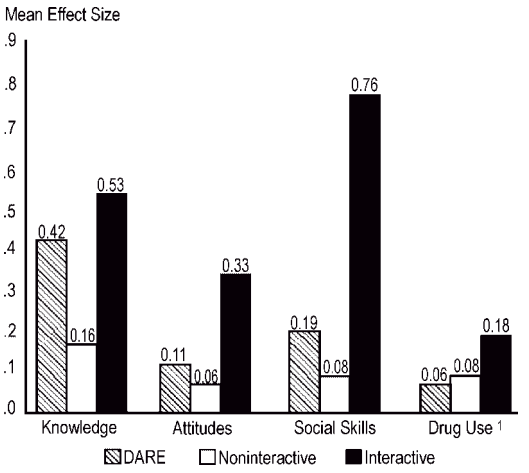


FIGURE 2

Weighted mean effect size, by outcome, for DARE and other drug use prevention programs

Note: Comparison programs selected from Tobler.[10]
1. Drug use includes alcohol, tobacco, and marijuana.

those for interactive programs (Figure 3). Except for tobacco use, they also are smaller than those for noninteractive programs.

Discussion

The results of this meta-analysis suggest that DARE's core curriculum effect on drug use relative to whatever drug education (if any) was offered in the control schools is slight and, except for tobacco use, is not statistically significant. Across the studies, none of the average drug use effect sizes exceeded .11. Review of several meta-analysis of adolescent drug use prevention programs suggests that effect sizes of this magnitude are small.[10-14]

The small magnitude of DARE's effectiveness on drug use behavior may partially reflect the relatively low frequency of drug use by the elementary school pupils targeted by DARE's core curriculum. However, comparison of the DARE effect sizes with those of other school-based drug use prevention programs for same-age adolescents suggests that greater effectiveness is possible with early adolescents. Compared with the programs classified by Tobler as interactive, DARE's effect sizes for alcohol, tobacco, and marijuana use, both collectively and individually, are substantially less.[10] Except for tobacco use, they also are less than the drug use effect sizes for more traditional, noninteractive programs.

It has been suggested that DARE may have delayed effects on drug use behavior once pupils reach higher grades.[23,24] Longer-term follow-up studies are needed to test this possibility. Only four reviewed studies administered multiple posttests, and for two of these the results from some later posttests are uninterpretable. However, based on two experimental studies for which reliable information 1 and 2 years after implementation is available, there is no evidence that DARE's effects are activated when subjects are older.[25,26] Most long-term evaluations of drug use prevention programs have shown that curriculum effects decay rather than appear or increase with time.[27,28]

DARE's immediate effects on outcomes other than drug use were somewhat larger (especially for knowledge) and were statistically significant. These effect sizes, however, also were less than the comparable effect sizes for same-age interactive programs. That DARE's effect sizes for knowledge, attitudes, and skills were greater in magnitude than those of noninteractive programs may not be particularly meaningful because many of these types of programs, such as programs using "scare tactics" or emphasizing factual knowledge about drug use, have been discredited as unsuccessful.[29,30]

Comparison of DARE's core curriculum content with the interactive and noninteractive programs' curricula may partially explain the relative differences in effect sizes among these programs. Interactive programs tend to emphasize developing drug-specific social skills and more general social competencies,

whereas non-interactive programs focus largely on intrapersonal factors. Because DARE has features of both interactive and noninteractive programs, it is perhaps not surprising that the effect sizes we reported should fall somewhere in between. Perhaps greater emphasis in the DARE core curriculum on social competencies and less emphasis on affective factors might result in effect sizes nearer to those reported for interactive programs. However, it is difficult to speculate on the effect of adding or subtracting particular lessons to or from DARE's curriculum. Most school-based prevention program evaluations have assessed the effectiveness of an overall program rather than various program components or combinations of components.

Who teaches DARE and how it is taught may provide other possible explanations for DARE's limited effectiveness. Despite the extensive DARE training received by law enforcement officers, they may not be as well equipped to lead the curriculum as teachers. No studies have been reported in which the DARE curriculum was offered by anyone other than a police officer; results from such a study might suggest whether teachers produce better (or worse) outcomes among pupils.

Regardless of curriculum leader, however, the generally more traditional teaching style used by DARE has not been shown to be as effective as an interactive teaching mode.[10,14] Although some activities encourage pupil interaction, the curriculum relies heavily on the officer as expert and makes frequent use of lectures and question-and-answer sessions between the officer and pupils. In fact, it is in teaching style, not curriculum content, that DARE most differs from the interactive programs examined by Tobler. The DARE core curriculum recently was modified to introduce more participatory activities, which may lead to greater program effectiveness.

Several limitations should be considered in evaluating our findings. The number of evaluations reviewed (eight) is not large when compared with the vast number of sites where DARE has been implemented. The consistency of results across studies, however, suggests that the results are likely to be representative of DARE's core curriculum. Even so, we would have preferred a full set of eight effect sizes for each outcome.

It is possible that the effect sizes for the DARE studies may have been attenuated compared with the drug use prevention programs reviewed by Tobler because the control groups were not pure "no treatment" groups. As documented by Tobler, effect sizes are lower when the control group receives some sort of drug education.[10,14] The DARE evaluations generally lacked information on alternative treatments received by the control groups, but it is likely that most control groups received some drug education because the studies occurred after the 1986 Drug-Free Schools and Communities Act. However, approximately half (54%) of the programs reviewed by Tobler also were conducted between 1986 and 1990, suggesting that they may suffer from the same effect.[10]

Most of the drug use prevention programs evaluated by Tobler were university research-based evaluation studies, whereas DARE is a commercially available curriculum. Although the magnitude of the resources invested in DARE is considerable, the intensity of effort devoted to smaller-scale programs may be greater. Some diminished effectiveness is perhaps inevitable once programs are widely marketed.

Although we found limited immediate core curriculum effects, some features of DARE may be more effective, such as the middle school curriculum. In addition, DARE's cumulative effects may be greater in school districts where all DARE curricula for younger and older students are in place. Other DARE outcomes, such as its impact on community law enforcement relations, also may yield important benefits. However, due to the absence of evaluation studies, consideration of these features is beyond this study's scope.

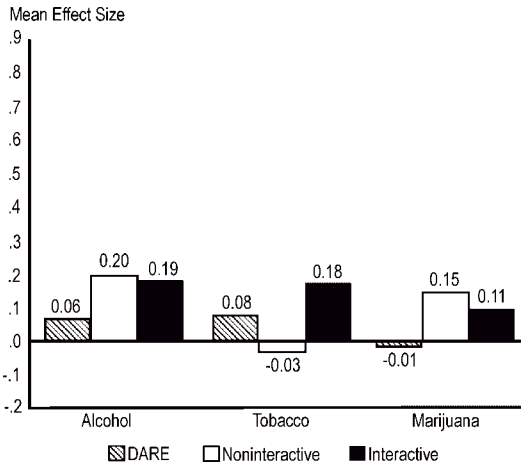


FIGURE 3

Weighted mean effect size, by drug, for DARE and other drug use prevention programs

Note: Comparison programs selected from Tobler.[10]

DARE's limited influence on adolescent drug use behavior contrasts with the program's popularity and prevalence. An important implication is that DARE could be taking the place of other, more beneficial drug use curricula that adolescents could be receiving. At the same time, expectations concerning the effectiveness of any school-based curriculum, including DARE, in changing adolescent drug use behavior should not be overstated.[31]

Acknowledgments

This research was supported by awards from the National Institute of Justice, Office of Justice Programs, US Department of Justice (91-DD-CX-K053) and the National Institute on Drug Abuse, US Department of Health and Human Services (5 R01 DA07037).

We thank Susan L. Bailey, Karl E. Bauman, George H. Dunteman, Robert L. Hubbard, Ronald W. Johnson, and J. Valley Rachal for their thoughtful comments on an earlier draft of this paper. In addition, we acknowledge the support and assistance of Jody M. Greene on this project and the editorial assistance of Beryl C. Pittman, Linda B. Barker, and Richard S. Straw.

Appendix A—Bibliography of DARE Evaluations

Aniskiewicz RE, Wysong EE. *Project DARE Evaluation Report: Kokomo Schools Spring, 1987.* Kokomo, Ind: Indiana University-Kokomo; 1987. Unpublished report.

Aniskiewicz R, Wysong E. Evaluating DARE: drug education and the multiple meanings of success. *Policy Stud Rev.* 1990;9:727–747.

Anonymous. *Project DARE Evaluation, Spring, 1987.* Pittsburgh, Pa; 1987. Unpublished paper.

Becker HR, Agopian MW, Yeh S. Impact evaluation of Drug Abuse Resistance Education (DARE). *J Drug Educ.* 1990;22:283–291.

Clayton RR, Cattarello A, Day LE, Walden KP. Persuasive communication and drug abuse prevention: an evaluation of the DARE program. In: Donohew L, Sypher H, Bukoski W, eds. *Persuasive Communication and Drug Abuse Prevention.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1991a:295–313.

Clayton RR, Cattarello A, Walden KP. Sensation seeking as a potential mediating variable for school-based prevention intervention: a two-year follow-up of DARE. *Health Commun.* 1991b;3:229–239.

DeJong W. A short-term evaluation of Project DARE (Drug Abuse Resistance Education): preliminary indicators of effectiveness. *J Drug Educ.* 1987;17:279–294.

Dukes R, Matthews S. *Evaluation of the Colorado Springs Citywide DARE Program.* Colorado Springs, Colo: University of Colorado, Center for Social Science Research; 1991. Unpublished report.

Earle RB, Garner J, Phillips N. *Evaluation of the Illinois State Police Pilot DARE Program.* Springfield, Ill: A.H. Training and Development Systems, Inc; 1987. Unpublished report.

Ennett ST, Rosenbaum DP, Flewelling RL, Bieler GS, Ringwalt CR, Bailey SL. Long-term evaluation of Drug Abuse Resistance Education. *Addict Behav.* 1994;19:113–125.

Evaluation and Training Institute. *DARE Evaluation Report for 1985–1989.* Los Angeles, Calif: Evaluation and Training Institute; 1990. Unpublished report.

Faine JR, Bohlander E. *Drug Abuse Resistance Education: An Assessment of the 1987–88 Kentucky State Police DARE Program.* Bowling Green, Ky: Western Kentucky University, Social Research Laboratory; 1988. Unpublished report.

Faine JR, Bohlander E. *DARE in Kentucky Schools 1988–89.* Bowling Green, Ky: Western Kentucky University, Social Research Laboratory; 1989. Unpublished report.

Harmon MA. Reducing the risk of drug involvement among early adolescents: an evaluation of Drug Abuse Resistance Education (DARE). *Eval Rev.* 1993;17:221–239.

Kethineni S, Leamy DA, Guyon L. *Evaluation of the Drug Abuse Resistance Education Program in Illinois: Preliminary Report.* Normal, Ill: Illinois State University, Department of Criminal Justice Sciences; 1991. Unpublished report.

McCormick, FC, McCormick ER. *An Evaluation of the Third Year Drug Abuse Resistance Education (DARE) Program in Saint Paul.* Saint Paul, Minn: Educational Operations Concepts, Inc; 1992. Unpublished report.

McDonald RM, Towberman DB, Hague JL. *Volume II: 1989 Impact Assessment of Drug Abuse Resistance Education in the Commonwealth of Virginia.* Richmond, Va: Virginia Commonwealth University, Institute for Research in Justice and Risk Administration; 1990. Unpublished report.

Manos MJ, Kameoka KY, Tanji JH. *Evaluation of Honolulu Police Department's Drug Abuse Resistance Education Program.* Honolulu, Hawaii: University of Hawaii-Manoa, School of Social Work, Youth Development and Research Center; 1986. Unpublished report.

Nyre GF, Rose C. *Drug Abuse Resistance Education (DARE) Longitudinal Evaluation Annual Report, January 1987.* Los Angeles, Calif: Evaluation and Training Institute; 1987. Unpublished report.

Nyre GF, Rose C, Bolus RE. *Drug Abuse Resistance Education (DARE) Longitudinal Evaluation Annual Report, August 1987.* Los Angeles, Calif: Evaluation and Training Institute; 1987. Unpublished report.

Ringwalt C, Ennett ST, Holt KD. An outcome evaluation of Project DARE (Drug Abuse Resistance Education). *Health Educ Res.* 1991;6:327–337.

Walker S. *The Victoria Police Department Drug Abuse Resistance Education Programme (D.A.R.E.) Programme Evaluation Report #2 (1990).* Victoria, British Columbia: The Ministry of the Solicitor-General, Federal Government of Canada; 1990. Unpublished report.

Appendix B—Bibliography of Comparison Program Evaluations

Allison K, Silver G, Dignam C. Effects on students of teacher training in use of a drug education curriculum. *J Drug Educ.* 1990;20:31–46.

Dielman T, Shope J, Butchart A, Campanelli P. Prevention of adolescent alcohol misuse: an elementary school program. *J Pediatr Psychol.* 1986;11:259–281.

Dielman T, Shope J, Campanelli P, Butchart A. Elementary school-based prevention of adolescent alcohol misuse. Pediatrician: *Int J Child Adolesc Health.* 1987;14:70–76.

Dielman T, Shope J, Leech S, Butchart A. Differential effectiveness of an elementary school-based alcohol misuse prevention program. *J Sch Health.* 1989;59:255–262.

Dubois R, Hostelter M, Tosti-Vasey J, Swisher J. *Program Report and Evaluation: 1988–1989 School Year.* Philadelphia, Pa: Corporate Alliance for Drug Education; 1989. Unpublished report.

Flay B, Koepke D, Thomson S, Santi S, Best J, Brown S. Six-year follow-up of the first Waterloo school smoking prevention trial. *Am J Public Health.* 1989;79:1371–1376.

Flay B, Ryan K, Best J, et al. Cigarette smoking: why young people do it and ways of preventing it. In: McGrath P, Firestone P, eds. *Pediatric and Adolescent Behavioral Medicine.* New York, NY: Springer-Verlag, 1983:132–183.

Flay B, Ryan K, Best J, et al. Are social psychological smoking prevention programs effective? The Waterloo study. *J Behav Med.* 1985;8:37–59.

Gersick K, Grady K, Snow D. Social-cognitive development with sixth graders and its initial impact on substance use. *J Drug Educ.* 1988;18:55–70.

Gilchrist L, Schinke S, Trimble J, Cvetkovich G. Skills enhancement to prevent substance abuse among American Indian adolescents. *Int J Addict.* 1987;22:869–879.

Johnson C, Graham J, Hansen W, Flay B, McGuigan K, Gee M. *Project Smart After Three Years: An Assessment of Sixth-Grade and Multiple Grade Implementations.* Pasadena, Calif: University of Southern California, Institute for Prevention Research; 1987. Unpublished report.

McAlister A. Approaches to primary prevention. Presented at the National Academy of Science National Research Council Conference on Alcohol and Public Policy; May 1983; Washington, DC.

Moskowitz J, Malvin J, Schaeffer G, Schaps E. Evaluation of an affective development teacher

training program. *J Primary Prev.* 1984;4: 150–162.

Sarvela P. Early adolescent substance use in rural northern Michigan and northeastern Wisconsin. In: *Dissertation Abstracts International.* 1984:46 01, 2000A. University Microfilms No. 84–22, 327.

Sarvela P, McClendon E. An impact evaluation of a rural youth drug education program. *J Drug Educ.* 1987; 17:213–231.

Schaeffer G, Moskowitz J, Malvin J, Schaps E. *The Effects of a Classroom Management Teacher Training Program on First-Grade Students: One Year Follow-Up.* Napa, Calif: The Napa Project, Pacific Institute for Research; 1981. Unpublished report.

Schaps E, Moskowitz J, Condon J, Malvin J. A process and outcome evaluation of an affective teacher training primary prevention program. *J Alcohol Drug Educ.* 1984;29:35–64.

Schinke S, Bebel M, Orlandi M, Botvin G. Prevention strategies for vulnerable pupils: school social work practices to prevent substance abuse. *Urban Educ.* 1988;22:510–519.

Schinke SP, Blythe BJ. Cognitive-behavioral prevention of children's smoking. *Child Behav Ther.* 1981;3:25–42.

Schinke S, Gilchrist L. Primary prevention of tobacco smoking. *J Sch Health.* 1983;53:416–419.

Schinke S, Gilchrist L, Schilling R, Snow W, Bobo J. Skills methods to prevent smoking. *Health Educ Q.* 1986;13:23–27.

Schinke S, Gilchrist L, Snow W. Skills interventions to prevent cigarette smoking among adolescents. *Am J Public Health.* 1985;75:665–667.

Schinke S, Gilchrist L, Snow W, Schilling R. Skills-building methods to prevent smoking by adolescents. *J. Adolesc Health Care.* 1985;6: 439–444.

Shope J, Dielman T, Leech S. An elementary school-based alcohol misuse prevention program: two year follow-up evaluation. Presented at the American Public Health Association Annual Meeting; November 1988; Boston, Mass.

References

1. Glynn TJ. Essential elements of school-based smoking prevention program. *J Sch Health.* 1989;59:181–188.
2. *Healthy People 2000.* Washington, DC: US Dept of Health and Human Services; 1988. DHHS publication no. PHS91-50212.
3. Ringwalt CL, Greene JM. Results of school districts' drug prevention coordinator's survey. Presented at the Alcohol, Tobacco, and Other Drugs Conference on Evaluating School-Linked Prevention Strategies; March 1993; San Diego, Calif.
4. *Implementing Project DARE: Drug Abuse Resistance Education.* Washington, DC: Bureau of Justice Assistance; 1988.
5. Drug-Free Schools and Community Act. Pub L No. 101-647, §5122(e)(1).
6. Bangert-Drowns RL. Review of developments in meta-analytic method. *Psychol Bull.* 1986;99:388–399.
7. Rosenthal R. *Meta-Analytic Procedures for Social Research.* Applied Social Research Methods Series. Vol 6. Newbury Park, Calif: Sage Publications; 1991.
8. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis.* New York, NY: Academic Press; 1985.
9. Perry PD, Tobler NS. Meta-analysis of adolescent drug prevention programs: final report. In: *Manual for Effect Size Calculation: Formula Used in the Meta-Analysis of Adolescent Drug Prevention Programs.* Rockville, Md: National Institute on Drug Abuse; 1992: appendix 3.
10. Tobler NS. *Meta-Analysis of Adolescent Drug Prevention Programs: Final Report.* Rockville, Md: National Institute on Drug Abuse; 1992.
11. Bangert-Drowns RL. The effects of school-based substance abuse education: a meta-analysis. *J Drug Educ.* 1988;18:243–264.
12. Bruvold WH. A meta-analysis of adolescent smoking prevention programs. *Am. J Public Health.* 1993;83:872–880.
13. Bruvold WH, Rundall TG. A meta-analysis and theoretical review of school-based tobacco and alcohol intervention programs. *Psychol Health.* 1988;2:53–78.
14. Tobler NS. Meta-analysis of 143 adolescent drug prevention programs: quantitative outcome results of program participants compared to a control or comparison group. *J Drug Issues.* 1989;16:537–567.
15. Hansen WB. School-based substance abuse prevention: a review of the state of the art in curriculum, 1980–1990. *Health Educ Res.* 1992;7:403–430.
16. Nyre GF, Rose C. *Drug Abuse Resistance Education (DARE) Longitudinal Evaluation Annual Report, January 1987.* Los Angeles, Calif: Evaluation and Training Institute; 1987. Unpublished report.

17. Nyre GF, Rose C, Bolus RE. *Drug Abuse Resistance Education (DARE) Longitudinal Evaluation Annual Report, August 1987*. Los Angeles, Calif: Evaluation and Training Institute; 1987. Unpublished report.
18. Manos MJ, Kameoka KY, Tanji JH. *Evaluation of Honolulu Police Department's Drug Abuse Resistance Program*. Honolulu, Hawaii: University of Hawaii-Manoa, School of Social Work, Youth Development and Research Center, 1986. Unpublished report.
19. Bauman KE, Dent CW. Influence of an objective measure on the self-reports of behavior. *J Appl Psychol.* 1982;67:623–628.
20. Moskowitz JM. Why reports of outcome evaluations are often biased or uninterpretable. *Eval Progr Plan.* 1993;16:1–9.
21. Murray DM, Hannan PJ. Planning for the appropriate analysis in school-based drug use prevention studies. *J Consult Clin Psychol.* 1990;58:458–468.
22. Biglan A, Ary DV. Methodological issues in research on smoking prevention. In: Bell CS, Battjes R, eds. *Prevention Research: Deterring Drug Abuse Among Children and Adolescents.* Rockville, Md: National Institute on Drug Abuse; 1985:170–195.
23. Clayton RR, Cattarello A, Day LE, Walden KP. Persuasive communication and drug abuse prevention: an evaluation of the DARE program. In Donohew L, Sypher H, Bukoski W, eds. *Persuasive Communication and Drug Abuse Prevention.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc: 1991:295–313.
24. Ringwalt CL, Ennett ST, Holt KD. An outcome evaluation of project DARE (Drug Abuse Resistance Education). *Health Educ Res.* 1991;6:327–337.
25. Clayton RR, Cattarello A, Walden KP. Sensation seeking as a potential mediating variable for school-based prevention intervention: a two-year follow-up of DARE. *Health Commun.* 1991;3:229–239.
26. Ennett ST, Rosenbaum DP, Flewelling RL, Bieler GS, Ringwalt CL, Bailey SL. Long-term evaluation of Drug Abuse Resistance Education. *Addict Behav.* 1994;19:113–125.
27. Ellickson PL, Bell RM, McGuigan K. Preventing adolescent drug use: long-term results of a junior high program. *Am J Public Health.* 1993;83: 856–861.
28. Murray DM, Pirie P, Luepker RV, Pallonen U. Five- and six-year follow-up results from four seventh-grade smoking prevention strategies. *J Behav Med.* 1989;12:207–218.
29. Botvin GJ. Substance abuse prevention: theory, practice, and effectiveness. In: Tonry M, Wilson JQ, eds. *Drugs and Crime.* Chicago, Ill: University of Chicago Press; 1990:461–519.
30. Moskowitz J. Preventing adolescent substance abuse through education. In: Glynn T, Leukefeld CG, Ludford JP, eds. *Preventing Adolescent Drug Abuse: Intervention Strategies.* Rockville, Md: National Institute on Drug Abuse; 1983:233–249.
31. Dryfoo JG. Preventing substance abuse: rethinking strategies. *Am J Public Health.* 1993;83:793–795.

Explanation and Critique

In this article, Ennett and colleagues perform a meta-evaluation of the effectiveness of Project DARE, one of the most widely implemented drug education programs in North America. They focused on the program's core curriculum and its short-term impact (because so few studies reported both short- and long-term effects) on the six outcome measures the program was specifically designed to affect.

The first item Ennett and colleagues report is the nature of their bibliographic search, citing all search sources and listing complete citations for each in appendix A of the articles. Once again, even though they ultimately used only eight of the studies, the reader can evaluate how representative the original sample was by looking at the eight in the context of the original eighteen. Each article was scrutinized for inclusion based on the explicit criteria listed in the methods section: use of a control or comparison group; pretest-posttest design or posttest-only with randomized assignment; use of reliably operationalized quantitative outcome measures. They wanted to include only methodologically rigorous evaluations. An analysis of the characteristics of the included studies is

in table 2. This table is helpful for determining the similarity of the studies. The discussion on the "Characteristics of Evaluations" is clear and very good. They try to assess the reliability and validity of the measures in the eight studies.

The calculation of the effect size is straightforward.

> An effect size is defined as the difference between the intervention and the control group means for each outcome measure, standardized by dividing by the pooled standard deviation [effect size = $mean_I - mean_C/SD$]. If means and standard deviations were not available, we calculated effect sizes using formulas developed to convert other test statistics and percentages to effect sizes. In all cases, we used statistics reflecting covariance-adjusted means, with pretest values as covariates rather than unadjusted means so that any differences between the comparison groups before the intervention would not be reflected in the effect sizes.

They included only one outcome measure per outcome class (e.g., knowledge). When multiple measures for an outcome were reported, an effect size was calculated for each measure and those results were then averaged. These effect sizes are reported in table 3. Note that outside of "Knowledge," the differences between those who received the intervention and those who did not was quite small.

The calculation of weighted scores was done to control for study size (number of students). The idea is that the results of larger studies are more stable than small studies and, therefore, the estimates from the larger studies should have more weight. The weighted mean is calculated by multiplying the effect size by the inverse of its variance. The weighted mean effect sizes and their companion 95 percent confidence intervals are reported in figure 1. We get essentially the same results as we did in table 3. The calculation of the confidence interval, however, allows us to determine which effects were statistically significantly different than no difference at all. Since the confidence interval of the drug use outcome measure contains "0," there is no difference in drug usage between those who received the intervention and those who did not. Although the remaining effect sizes were statistically significant, their magnitude was not great.

"Great" is a relative word. Some impact is evident, but is Project DARE any better than any other drug education program? In order to assess this, Ennett and colleagues creatively compared the effect sizes of DARE outcomes with effect sizes of other drug education programs targeted at the same age group. Using Tobler's meta-analysis of those programs, the comparison between DARE and two classes of drug education programs (interactive and noninteractive) are shown in figure 2. Interactive non-DARE programs were more effective than DARE programs across all four common classes of outcomes. DARE outperformed the noninteractive programs on three outcome measures, but not in the ultimate outcome of drug usage.

In the discussion section of the article an attempt is made to explain the lack of positive findings in light of the number of studies included in the meta-evaluation, the method of presenting the drug education, and the commercial availability of DARE. This article was well conceived and conducted. With the widespread embrace of the DARE program nationally, what do you think the probability is that the results of this meta-evaluation will be utilized?

Note

1. There is specially designed software for meta-analysis, DSTAT, which calculates effect sizes of other summary statistics such as *t*-tests, *F*-values from analysis of variance, correlation coefficients, chi-square, *p*-values (Johnson 1993). The logic to their interpretation is the same as the explanation of mean differences above.

References

Bowen, William, and Chieh-Chen Bowen. 1998."Typologies, Indexing, Content Analysis, Meta-Analysis, and Scaling as Measurement Techniques." In *Handbook of Quantitative Methods in Public Administration,* edited by Marsha Whicker. New York: Marcel-Dekker, 51–86.

Hunter, J.E. and F.L. Schmidt. 1990. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings.* Beverly Hills, Calif.: Sage Publications.

Johnson, B.T. 1993. *DSTAT: Software for the Meta-Analytic Review of Research Literatures.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

PART VII

On Your Own

The presentation of new evaluation designs is now complete. Part VII attempts to aid you in interpreting what you have learned. The next three chapters present articles covering a range of areas of study to illustrate how broad the field of program evaluation is and to give you the opportunity to critique articles on your own.

Chapter 17 is an example of an evaluation of an economic development program, "Are Training Subsidies for Firms Effective? The Michigan Experience." This research evaluated the effects of a state-financed training grant program for manufacturing firms in Michigan. An economic development evaluation is particularly appropriate for this book because most economic development program evaluations have not been very sophisticated or even believable (see Bartick and Bingham 1997).

Chapter 18 looks at the outcome of a court-ordered racial desegregation case in assisted housing in Chicago. The evaluation attempts to find out if low-income minority families are "better off" if they live in predominately white suburbs than they would be if they lived in a minority neighborhood in the city of Chicago. The evaluation, then, attempts to test the concept of "geography of opportunity," which suggests that where individuals live affects their opportunities and life outcomes.

Finally, chapter 19 looks at a state public policy issue: What are the differences in alcohol consumption associated with state versus privately owned liquor stores? In "Changes in Alcohol Consumption Resulting from the Elimination of Retail Wine Monopolies: Results from Five U.S. States," the researchers evaluate alcohol consumption based on public versus private ownership of retail alcohol stores.

You are "on your own." You have the opportunity to critique (and not just criticize) these three articles—to point out strengths as well as weaknesses. But what form should such a critique take? There are many possible forms, but we offer the following:

1. Describe the program.
2. Explain the research design.
3. Determine how the data were collected.
4. Detail the findings.
5. Critique the evaluation.

In describing the program, tell us what the goals and objectives of the program (not the goals and objectives of the evaluation) were. Then explain how the program was carried out. Give the reader a thumbnail sketch of how the program operated.

What was the research design the evaluator used? Was it a simple time-series design? Was it a pretest-posttest comparison group design? Or perhaps it was a patched design of some type?

Next you might critique the data used in the evaluation. In your opinion are the data both reliable and valid (see chapter 3)? If sampling was used, was the sample large enough? What was the response rate? Was the sample valid?

What was the outcome of the program? Did it do what it was supposed to do?

Did the evaluation design adequately handle all of the possible threats to validity? Did the evaluation look for other plausible alternatives that might explain program outcomes? What are the strengths and weaknesses of the evaluation? Given the nature of the problem, would you have done the evaluation any other way?

Good luck, and have fun. Remember, evaluations are really only puzzles waiting to be solved.

Reference

Bartik, Timothy J., and Richard D. Bingham. 1997. "Can Economic Development Programs Be Evaluated?" In *Dilemmas of Urban Economic Development: Issues in Theory and Practice,* edited by Richard D. Bingham and Robert Meer. Thousand Oaks, Calif.: Sage, 246–77.

CHAPTER 17

Are Training Subsidies for Firms Effective?

The Michigan Experience

Harry J. Holzer, Richard N. Block, Marcus Cheatham, and Jack H. Knott*

This paper explores the effects of a state-financed training grant program for manufacturing firms in Michigan. Using a three-year panel of data from a unique survey of firms that applied for these grants, the authors estimate the effects of receipt of a grant on total hours of training in the firm and the product scrap rate. They find that receipt of these grants is associated with a large and significant, though one-time, increase in training hours, and with a more lasting reduction in scrap rates.

The issue of private sector training in the United States has grown in importance in the past few years. Traditionally, most such programs funded by government (such as the Job Training Partnership Act [JTPA] programs of the 1980s and the earlier Comprehensive Employment and Training Act [CETA] programs) have provided training for *individual* unemployed workers. But plant closings and other dislocations, often stemming from growing international competition, have recently generated interest in funding training for employed workers at the *firm* level. The goal of this funding would be to protect the employment of workers at their current firms, and perhaps raise their earnings as well, by increasing their productivity through higher levels of on-the-job training.

What would be the effects of such funding on firm performance and on such labor market outcomes as wages and employment? Would the training subsidies actually raise

* Harry J. Holzer is Professor of Economics, Richard N. Block is Professor and Director of the School of Labor and Industrial Relations, Marcus Cheatham is Survey Director at the Institute for Public Policy and Social Research, and Jack H. Knott is Professor of Political Science and Director of the Institute for Public Policy and Social Research, all at Michigan State University. The authors thank Ann Bartel and Daniel Hamermesh for very helpful comments. This research was part of a larger project on evaluating private sector training in Michigan, and was funded by grants from the Michigan Departments of Commerce and Labor.

The authors will make all data (without employer names and addresses) and programs available to those who request them.

the amount of training that firms provide, or would they merely be windfalls for those already providing such training? The latter question involves the issue of "substitution" of public for private spending, which has been studied extensively in the context of *employment* subsidies for firms but not in the context of *training* subsidies. Yet, this concern is also a major issue in the policy debate over subsidies for private training (Commission on Workforce Quality and Labor Market Efficiency, U.S. Department of Labor, 1989).

In this paper, we explore the effects of state subsidies for private-sector training in a sample of manufacturing firms in Michigan. We consider the effects of training on worker performance (as measured by output scrap rates), and also the extent to which the subsidies actually raise the amount of training provided, as opposed to merely providing windfalls for the firms receiving them.

The data used for this analysis are from a unique survey of firms in Michigan that applied for training grants under the state's MJOB (Michigan Job Opportunity Bank-Upgrade) program during 1988 and 1989. These one-time grants were distributed to qualifying firms on a roughly first-come, first-serve basis. We have data on total hours of training and on other outcomes both for firms that applied successfully for these grants and for firms that applied unsuccessfully.

Given that firms' receipt or nonreceipt of grants was relatively independent of their characteristics, the data come fairly close to representing an experiment on the effects of randomly assigned training subsidies. Because grant recipiency may not have been totally random, however, we employ a variety of control variables in the estimation process. Furthermore, though the survey was only administered once, it contains current and retrospective questions covering a total of three years. The availability of a three-year panel on these firms enables us to control for unobserved characteristics of firms that might be correlated with training and other outcomes. The panel also enables us to test for the exogeneity of training and receipt of a grant with respect to pre-grant outcomes.

Previous Literature

Comparisons of on-the-job training levels between U.S. firms and those in other countries (for example, Osterman 1988; Office of Technological Assessment 1990) generally show lower investments in on-the-job training here. This finding has led to various discussions of market imperfections that might limit these investments, such as excessive employee turnover and liquidity constraints on firms (Parsons 1989). It has also led to several proposals (for example, Commission on Workforce Quality and Labor Market Efficiency, U.S. Department of Labor, 1989) for government subsidies of private sector training in firms to counter these labor market problems.

There is little empirical evidence to date on the question of how training subsidies affect firm behavior and performance.[1] A number of studies (for example, Lillard and Tan 1986; Lynch 1989; Mincer 1989) have used large micro datasets on individuals to determine the distribution of on-the-job training across demographic groups and to assess its effect on wage growth.[2] A few others (for example, Bishop 1988; Barron *et al.* 1989; Holzer 1990) have used data from the Employment Opportunities Pilot Project (EOPP) Survey of Firms in 1980 and 1982 to assess effects of training on worker wage and productivity growth (with the latter measured by performance ratings). These studies have found fairly sizable effects of training on both wage and productivity growth of employees. Bartel (1989), using a Columbia University survey of firms merged with Compustat data, also found positive effects of training on worker output.

Despite this work, there has been little analysis to date of firms' training choices or of

the effects of training on a broader range of worker and firm outcomes (such as output quality, sales, and employment).[3] Furthermore, very little has been done to assess the effects of the many training subsidy programs that have been implemented at the state level to maintain states' industrial bases and attract workers and jobs. "Upgrade" programs to subsidize the training of currently employed workers had been implemented in 46 states as of 1989 (Pelavin Associates 1990), and few of these programs have been formally evaluated.[4]

Firm-Level Data from the MJOB Program in Michigan

The Michigan Job Opportunity Bank-Upgrade program was in effect during the years 1986–90.[5] The program was designed to provide one-time training grants to eligible firms, defined as manufacturing companies with 500 or fewer employees that were implementing some type of new technology and were not past recipients of a grant. The grants were designed to cover the costs of instruction, tuition, travel, and training supplies for employees who were participating—that is, the *direct* costs (as opposed to the opportunity cost of lost employee time) of any formal or structured training program. The program goal was to spur a significant increase in training in each recipient firm at least during the year of the disbursement.

In its five years of operation, the MJOB program administered over four hundred grants to firms in the state, with an average grant size of approximately $16,000. Grants were approved on a first-come, first-serve basis (given available funding) to those who met eligibility requirements, though some funding was specifically set aside for firms that were participating in another state-level initiative known as the Michigan Modernization Service.[6]

In 1990, we mailed a survey questionnaire to all 498 firms that had applied for an MJOB grant during the period 1988–89.[7] One hundred fifty-seven of these firms responded, for a response rate of 32%. Of these firms, 66 had received a grant and 91 had not.[8]

The survey process raises two possible concerns about *sample* selection bias: one based on which firms applied for grants, and the other based on who chose to respond to our survey. The first concern is that *all* firms applying for grants claim to be planning to implement new technology and to desire additional training. If anything, this fact should bias toward zero any estimated effects of training grants on training hours or firm performance, since even non-recipients claim some interest in making these changes. The second concern is that the decisions of firms to respond to our survey might have been correlated with grant recipiency, training, or other observed outcomes. But since we had the actual applications of *all* firms that applied (which contain a few variables such as firm size), as well as the decisions on their grant request, we were able to test for at least a few of these potential correlations. The results of these tests show few correlations overall between survey responses and the observable variables.[9]

We also note that biases in our results could arise from any dishonest reporting of results, if recipients *over*state or nonrecipients *under*state their training changes. But we believe the incentives to do so were small, and that such distortions are unlikely to account for our results.[10]

In our survey, we attempted to gauge the amounts of formal job training (measured by the numbers of employees participating in "training programs" and hours spent per employee) that were provided to employees in each year from 1987 to 1989.[11] Several questions were also asked on the quality of output for each year, using a variety of measures that firms in the state appear to use in their own quality control programs. The quality measure with the highest response rate and with

the clearest interpretation was the scrap rate, which we focus on below in our analysis.[12]

Finally, a number of questions were asked about levels of sales, employment, and wages in each year, as well as about union membership and labor relations practices at the firm. Questions about labor-management practices asked about the presence of incentive pay schemes (such as profit sharing, gain sharing, or "pay for knowledge"); participation in management (through quality circles, labor-management committees, and the like); and a formal grievance procedure. Questions were also asked about the reasons for training in the firm—for example, to meet competitive pressures, incorporate new technologies, or receive a quality rating from customers.[13]

Table 1 presents means, standard deviations, and sample sizes over the three-year period for the characteristics and activities of these firms. Three observations thus appear per firm in most cases. The data are presented for three groups: the entire sample, firms that received a grant in one of the two years, and firms that did not receive a grant.

The table shows that the average firm in our sample has about 60 employees, pays its average employee about $18,000 a year (or $9.00 per hour for full-time workers), and has some type of participatory or incentive pay scheme. These firms often provide training for workers in order to receive quality ratings from customers as well as to facilitate the adoption of new technology.

The results of the table show a few differences between the characteristics of firms that received grants and those that did not. In particular, the firms that received grants were likely to have somewhat larger work forces, be more unionized, and have been assisted by the Michigan Modernization Service (though only the last difference is significant at the .05 level).[14]

Overall, the two sets of firms are quite comparable, and receipt of the grant is uncorrelated with most firm characteristics. This finding supports our contention that the data approximate the results of an experimental study, and that our results are less likely to be contaminated by unobservables and other statistical biases than are those based on traditional survey data.

We also note the relatively low levels of formal training, as measured by training hours per employee, provided by these firms—particularly firms that did not receive a grant, which provided just over a day's worth of training per year to their average employee. But since these data do not distinguish between new and continuing employees, the distribution of this training across workers with different seniority levels remains somewhat unclear.[15]

Finally, note the average grant size of almost $16,000 per firm, or roughly $240 per employee in recipient firms.[16]

Evidence of Effects of Training Grants

Summary Evidence

Before considering more formal evidence from regression equations on the effects of MJOB grants on training and other outcomes, we first consider summary data on these effects. Table 2 presents means and standard deviations for annual hours of training received per employee, cross-tabulated by year and by receipt of a grant. We present data on training in each of the three years for which we have data, as well as by the year in which grants were received, if at all. We can therefore follow firms over time and compare those receiving grants to those not receiving them.

Table 2 shows that receipt of an MJOB grant has a large, one-time effect on training in the firm. In each year in which a grant is received, there is a large positive effect on hours of training. Each of these increases over the previous year is significant.[17]

TABLE 1

Summary Statistics on Training and Firm Characteristics. (Standard Deviations in Parentheses)

Characteristic	*All Firms*	*Firms That*	
Grant		*Received Grant*	*Did Not Receive*
Number of Observations	390	185	205
Annual Hours of Training per Employee	14.97 (25.71)	19.39 (29.98)	10.98 (20.40)
Scrap Rate (Per 100) (as Percent of Sales)	3.84 (6.01)	3.63 (5.02)	4.09 (7.01)
Annual Wages	$18,010.13 (6349.08)	17,918.21 (6237.15)	18,177.63 (6473.74)
Employment	59.32 (74.12)	66.15 (91.15)	54.17 (57.78)
Sales	$5,822,554.3 (7,476,820.4)	5,555,824.1 (7,133,242.2)	5,996,047.0 (7,702,714.6)
Union	.20	.24	.16
Industrial Relations:			
Participation	.75	.80	.73
Incentive Pay	.52	.53	.53
Grievance Procedure	.34	.33	.35
Reason for Training:			
Quality Rating	.62	.65	.59
New Technology	.71	.73	.68
New Product	.29	.27	.30
Michigan Modernization Service	.24	.32	.18
Dollar Value of Training Grants	—	$15,607 (14,397)	—

Source: Author's survey of Michigan firms that applied for MJOB training grants during 1988 and 1989.

Furthermore, for those firms for which we have observations in the year subsequent to receipt of a grant (that is, grant recipients in 1988), we see training drop off almost to its pre-grant level in that subsequent year.[18]

As for the *magnitudes* of these changes, we find that training hours increased at recipient firms by 27.1 hours per employee (above their own trends in the absence of grants) in 1988 and by 38.1 hours per employee in 1989.[19] Given the numbers of employees per recipient firm in each year (69 in 1988 and 73 in 1989, respectively) and the values of grants ($12,160 and 20,077, respectively), we find increases in total training hours of .154 and .139 *per dollar spent* in each year. In other words, each hour of training costs the public sector $6–$7, plus any additional direct costs and foregone output borne by firms or workers.

As noted above, firms may have had some incentives to respond dishonestly to the survey questions on training. It is very unlikely, however, that such upward biases would be large enough to fully account for the magnitude of the observed effects on training (see note 10 above).

Thus, the summary data suggest that the training grants are achieving their purpose of increasing new worker training in recipient

TABLE 2

Annual Hours of Training per Employee, by Year and Receipt of Grant. (Means and Standard Deviations in Parentheses)

Description	*1987*	*1988*	*1989*
Firms Receiving Grants in 1988	7.672 (19.575)	35.978 (36.956)	10.056 (19.475)
Firms Receiving Grants in 1989	6.125 (11.309)	9.316 (17.361)	50.650 (36.218)
Firms Not Receiving Grants	10.703 (18.315)	9.818 (18.627)	12.353 (23.777)

Note: Sample sizes for firms receiving grants in 1988, firms receiving grants in 1989, and those not receiving grants in either year are 35, 28, and 71, respectively.

firms. If some substitution of public for private training is occurring here, it seems far too small to fully counteract the large positive training effects of grant receipt.

Grant Effects on Training: Regression Analysis

We estimate a system of equations below that is consistent with the standard human capital framework (Becker 1975; Mincer 1978), in which training grants would reduce the costs of training to firms and therefore raise their investment in training. This increased training, in turn, should raise employee output per unit cost (measured in terms of quantity or quality), firm sales, profits, and the wages and employment levels of workers in these firms.[20]

To formally test for the effects of receiving a training grant on a variety of firm-level behaviors and outcomes, our estimated equations are of the general forms

$$(1)\ a)\ \Delta \mathrm{TR}_{j,t} = a_0 + a_1\,\mathrm{GRANT}_{j,t} + a_2 \bullet \mathrm{GRANT}_{j,t-1} + a_3 \bullet X_{j,t-1} + u^1_{j,t}$$

$$b)\ \Delta \mathrm{QUAL}_{j,t}\,;\,\Delta \mathrm{OUT}_{j,t} = b_0 + b_1 \bullet \mathrm{GRANT}_{j,t} + b_2 \bullet \mathrm{GRANT}_{j,t-1} + b_3 \bullet X_{j,t-1} + u^2_{j,t}$$

where *j* and *t* denote the firm and time, respectively; GRANT is a dummy variable indicating whether or not a grant was received, TR represents hours of training per employee; QUAL is our measure of product quality; OUT represents various firm outcomes, such as value of sales, employment levels, and wage levels; the *X* variables are a variety of other controls for labor relations and other factors that could influence training or worker quality changes (described more fully in a section above), and the u's are error terms. Changes are defined as occurring between years *t-1* and *t*.

Equations (1a) and (1b) are very reduced-form in nature, with all other outcomes listed as functions of grant receipt. Another set of equations that use the intermediate outcome variable of training as a determinant of the final outcomes can be written as follows:

$$(2)\ \Delta \mathrm{QUAL}_{j,t}\,;\,\Delta \mathrm{OUT}_{j,t} = C_0 + C_1 \bullet \Delta \mathrm{TR}_{j,t} + C_2 \bullet \Delta \mathrm{TR}_{j,t-1} + C_3 \bullet X_{j,t-1} + u^3_{j,t}$$

In order for estimated effects to be unbiased, the error terms across the various equations must be uncorrelated with one another.

Given the retrospective panel nature of the data collected here, we can define virtually all of these variables (except for those *X* variables that are time-invariant) as *changes* (or first differences) rather than levels. This approach purges the estimated equations of unobserved fixed effects that might be correlated with our independent variables. These equations are thus comparable to the "difference in differences" estimates of training effects in the literature on individuals in CETA programs (for example, Bassi 1984).[21] Since our quantitative variables all appear in log form, we can also interpret the differences in logs as percent changes in each of these variables.

Although all variables in first-difference equations normally appear as changes (with time-invariant measures excluded), we also report some estimated equations below in which the *X* variables are included as levels. We do so to capture the possibility that ad-

justments in desired levels of training or in other market outcomes do not occur instantaneously and might be correlated with various characteristics of firms.[22]

It is conceivable, for example, that particular firm characteristics, such as changes in technology, product market competition, or institutional features, could change desired levels of training. Although it is unlikely that all of the relevant underlying variables are directly observable, some might be correlated with characteristics of firms that we do observe. The variables thus included to control for this possibility are union status; assistance from the Michigan Modernization Service; sales, employment, and wage *levels* in the previous year; stated reasons for training; and industrial relations characteristics.[23]

To the extent that we may be "over-controlling" by including level variables in first-difference equations, estimates with and without these controls might provide us with upper and lower bounds to the magnitudes of the true effects. Furthermore, we will still have two cross-sections (changes for 1987–88 and for 1988–89) that can be pooled in our estimation.[24] Pooled equations all contain time dummies, to control for changes over time in overall economic conditions (such as the business cycle). Sales and wage outcomes are adjusted for inflation by the GNP Deflator.

The two cross-sections also enable us to consider one year of *lagged* effects of training grants on outcomes, to determine whether any of the observed effects last beyond the year in which the grant is received. Equations in which lagged effects are estimated will be limited to non-recipient firms and to those that received grants in 1988, since the latter are the only grant recipients for which lagged outcomes (in 1989) can be observed.

Finally, the three-year panel enables us to test for bias due to any non-randomness in the distribution of training grants across firms. Given the first-come, first-serve awarding of grants, the only unobserved characteristics of eligible firms that might have influenced their likelihood of receiving grants were early knowledge of the program and managerial speed and efficiency in response to such knowledge. Furthermore, only if these firm characteristics are correlated with firm training or other *changes* in outcomes is there any likelihood of bias here.

We test for this possibility by seeing whether or not *pre*-grant training and outcomes are correlated with receipt of a grant.[25] Results of these tests performed on 1987 levels and 1987–88 changes or on 1988 levels of training (which are relevant for grants in 1988 and 1989, respectively) generally support the assumption of randomly assigned grants (at least with respect to observable variables).[26] Of course, correlations between grant recipiency and non-fixed unobservable variables could still bias our estimated results.

Training Equations

Table 3 presents the estimated effects of receiving a grant on changes in hours of training provided per employee. Results are presented for both current and lagged effects of training. Estimates are also provided with and without the control variables described above.

The results confirm our earlier finding that receipt of a training grant has a large, positive effect on hours of training per employee. For the full sample (columns 1 and 2), a grant raises the amount of training by a factor of 2 to 3. In other words, training in that particular year more than triples for firms receiving grants. As noted above, if there is any substitution of public for private funding during the period of grant recipiency, it is far too small to eliminate the net positive effect of grant recipiency on training.

Training appears to be reduced the year after the grant is received by an amount almost

TABLE 3

Training Equations: Effects of Grants and Other Factors.
(Standard Errors in Parentheses)

Variable	*1*	*2*	*3*	*4*
Intercept	.048	0.310	.086	.228
	(.120)	(.284)	(.137)	(.280)
Grant	2.398**	2.330**	2.142**	2.093**
	(.185)	(.205)	(.242)	(.245)
Lag Grant	—	—	-2.050**	-2.240**
	—	—	(.241)	(.273)
Time Dummies	Yes	Yes	Yes	Yes
Log (Sales)	—	.0038	—	.0050
		(.092)		(.0034)
Log (Employment)	—	-.0003	—	-.0010
		(.0014)		(.0015)
Log (Wages)	—	.0000	—	.0000
		(.0000)		(.0000)
Union	—	.153	—	.152
		(.265)		(.280)
Worker Participation	—	-.343	—	-.382*
		(.225)		(.224)
Incentive Payments	—	.053	—	-.043
		(.177)		(.183)
Grievance Procedures	—	.037	—	.285
		(.221)		(.226)
Training for:				
Quality Rating	—	.132	—	.124
		(.184)		(.189)
New Product	—	-.116	—	-.063
		(.199)		(.199)
New Technology	—	-.223	—	-.073
		(.196)		(.203)
Michigan Modernization Service	—	.032	—	-.107
		(.203)		(.215)
R^2	.410	.409	.499	.548
N	250	217	194	171

Note: Dependent Variable is Δlog (Annual Hours of Training per Employee). "Grant" and "Lag Grant" are dummy variables. All other independent variables appear in *levels*. Equations 3 and 4 are estimated on a smaller sample that excludes grant recipients in 1989.
* Statistically significant at the .10 level;
** at the .05 level (two-tailed tests).

equal to the original increase (columns 3 and 4). This pattern implies that the effects on training of these one-time grants were indeed one-time, not lasting. Nonetheless, the explicit program goal—to boost training significantly at least during the year in which the grant was disbursed—was unquestionably achieved.

We also note that the control variables have very little effect on the observed effects of grant recipiency on quantities of training. Among these controls, only sales levels seem to have a marginally significant, positive effect on training; and worker participation schemes seem to have marginal negative effects. Whether these relationships are at all causal is, of course, unclear. The controls were not jointly significant in any of the estimated equations.

Scrap Rate Equations

In Table 4 we provide estimates of the effects of grant recipiency and of training in general on our measure of output quality, which is the scrap rate. Negative coefficients below will thus denote positive effects on quality. Columns 1–4 have specifications identical to those of Table 3, and in columns 5–8 the only modification is that the change in the log of hours trained per employee (the dependent variable of Table 3) has now replaced the dummy variable for grant recipiency.

The results show that receipt of a training grant has a negative effect on the scrap rate, indicating improved performance of production workers. The improvement associated with receipt of the grant is approximately 20% without controls and about 13% with them (though only the former is significant, at about the .05 level in a one-tailed test). As remarked, however, the controls are not jointly significant in any equation reported in Table 4.

Given the mean scrap rates listed in Table 1, the reductions in scrap reported here constitute declines of about .5–.7 of a percentage

TABLE 4

Scrap Equations: Effects of Grants and Training. (Standard Errors in Parentheses)

Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Intercept	-.098	.027	.000	.210	-.130	.120	-.076	.221
	(.090)	(.269)	(.125)	(.363)	(.094)	(.281)	(.133)	(.396)
Grant	-.201*	-.134	-.379**	-.306	—	—	—	—
	(.127)	(.140)	(.191)	(.234)				
Lag Grant	—	—	-.082	.109	—	—	—	—
			(.191)	(.247)				
Δ Training	—	—	—	—	-.071**	-.068**	-.090	-.076
					(.037)	(.041)	(.060)	(.074)
Lag Δ Training	—	—	—	—	—	—	.053	.052
							(.054)	(.070)
Other Controls	no	yes	no	yes	no	yes	no	yes
R^2	.031	.149	0.54	.193	.051	.171	.079	.236
N	107	91	87	71	90	85	65	60

Note: Dependent variable is Δlog (Scrap rate). Training and other variables appear as in Table 1. All equations include time dummies.
*Statistically significant at the .10 level; ** at the .05 level (two-tailed tests).

point in the overall scrap rates of firms. If output is valued according to sales, such reductions are worth approximately $30,000–$50,000 per year (though this figure will clearly overstate the gains if, for example, product demand is limited or output is valued at the cost of inputs).

Furthermore, for the subsample for which lagged effects can be observed (columns 3 and 4 in Table 4), this reduction constitutes improvements in scrap rates that are even larger than for the entire sample, and they are not substantially reversed in the subsequent year. The inclusion of control variables again reduces but does not eliminate the observed effects of grant recipiency (since the sum of current and lagged effects of a grant declines from about 46% to 20% with their inclusion). The effects of a grant on the quality of worker output thus appear to be sizable and permanent, despite the one-time nature of the grants.

The estimates of columns 5–8 confirm that, more generally, training has positive effects on the quality of output, though again these effects approach significance only when other controls are not included. A doubling of worker training in any year produces a contemporaneous reduction in the scrap rate of about 7%, though there is some offset of this effect during the subsequent year.

We also note that if training has effects on dimensions of quality *besides* scrap (such as the rework rate), the effects on scrap alone will understate the overall effects of training on quality.[27]

If the fairly sizable positive observed effects of training on quality are accurate, the failure of firms to undertake more of this training in the absence of these grants suggests either high training costs, especially for liquidity-constrained firms, or other factors that limit their ability to realize returns on their investments in such training. Management might reasonably anticipate a weak return to investments in training if, for example, the training is fairly general in nature and substantial numbers of trained em-

ployees can be expected to be be lost through turnover (Becker 1975). Under either of these interpretations, the case for some subsidy of private-sector training is strengthened. Of course, the sensitivity of our estimates to inclusion of control variables raises at least some questions about the exact magnitudes of these effects.

Finally, if training does enhance worker productivity, we would expect this effect to result in higher sales (due to lower costs and prices, as well as higher product quality) and, ultimately, higher wages and employment for workers over the long run. Unfortunately, our three-year panel of data is too short for a rigorous test of these effects. But we do have results for equations identical to those of Table 4 except that the dependent variables are changes in the logs of sales, employment, or wages of workers. These results show that receipt of a training grant had little effect within the first few years on sales and wage changes, but positive and marginally significant effects on short-term employment changes.

In any event, a longer panel of observations on each firm than we have employed here is no doubt needed before the product and labor market effects of subsidized private training can be fully evaluated.

Conclusion

Our results show that receipt of a training grant substantially increased the amount of training observed in the year of receipt, though not beyond that year. The grants thus seem to have achieved the baseline goal of the program—namely, to spur at least a one-time increase in training hours in recipient firms—and did not simply provide a windfall for these firms. Every extra hour of training cost the government roughly $6–$7.

Furthermore, we observed contemporaneous positive effects on the quality of output (as measured by the scrap rate) of grant recipiency and of training more generally, which were not completely lost in subsequent years. This finding indicates that observed effects of training on quality may be lasting ones.

Finally, in results not presented here, we found little effect of grants and of training on sales and wages, though some small positive effects on employment of firms were observed. A longer panel of data on such firms would clearly be needed to estimate such product and labor market effects with any degree of confidence.

Nevertheless, our results suggest that public sector funding for private sector training in firms has the potential to raise the amount of such training substantially without producing large windfalls for these firms, and also to generate other positive product and labor market outcomes. Of course, the case for such funding must ultimately rest on a more thorough evaluation of social costs and benefits, or on some clearer evidence of market imperfections that may prevent firms from achieving optimal investments on their own. But given the evidence presented elsewhere of low training levels overall in the United States and of high returns to workers (and perhaps to firms) from training when it is provided, the results here suggest that public funding of private training might be successful in raising training levels and improving worker productivity overall.

Notes

1. See, for example, Johnson and Tomola (1977), Perloff and Wachter (1979), and Bishop (1981) for studies of *employment* subsidies. This literature considers whether firms will participate in subsidy programs and, if they do, whether the subsidies will mostly provide windfall effects. Both questions are relevant when considering training subsidies, though only the latter can be evaluated with our sampling strategy.
2. These and other studies are summarized in Brown (1989). Comparisons between training in the United States and various other countries have also begun to appear (Blanchflower and Lynch 1991).

3. On the broader issue of how industrial relations and compensation practices affect firm performance, a somewhat wider range of empirical literature is now available. See, for instance, the volumes edited by Kleiner *et al.* (1987) and Ehrenberg (1990).
4. Some discussion (though little formal analysis) of state-financed training programs in Illinois and California appears in Creticos and Sheets (1989).
5. This training program was begun under the administration of Governor James Blanchard and ended shortly after the election of Governor John Engler in November 1990.
6. The Michigan Modernization Service provided technical assistance to companies (primarily in manufacturing) that were seeking to implement new technologies. For more information, see Block *et al.* (1991). Officials from the Governor's Office on Job Training (GOJT) have assured us that no other factors besides these were used to determine grant eligibility and recipiency. Applications had to include detailed accounts of the new technology, which workers would be trained in its use, and expected expenses in such training. Reimbursement for these expenses was made on the basis of receipts provided. No other documentation of worker participation in the training programs was collected by the state.
7. We obtained the names and addresses of these companies from GOJT, which administered the MJOB program. All companies that had applied for grants were mailed a written survey, and phone followups were performed for those that did not return them within the first several weeks.
8. Of the original 498 applicants, 270 received grants in 1988 or 1989. The success rate in our sample (42%) is thus a bit lower than that in the overall applicant pool (54%). There were 45 multiple-year applicants—firms that applied a second time after being rejected once—the full sample (9%) and just 8 in the respondent sample (5%); of these, 15 and 3, respectively, received grants after reapplying.
9. There were no significant differences between respondents and non-respondents in unionization rates, minority ownership, relationship to auto industry, or grant recipiency. Respondents were somewhat more likely than non-respondents to have received assistance from the Michigan Modernization Service (22% versus 15%, respectively) and were also somewhat larger on average (95 versus 77 employees).
10. Although recipient firms might feel some pressure to *over*state their training hours in order to justify their receipt of grants, there was no monetary incentive for them to do so, since they were not eligible to reapply for grants. Non-recipient firms might have had incentives to *under*state their current training in order to be eligible to reapply in the future, but the number of repeated applications is not very large (see note 8). Thus, the magnitudes of reporting biases re likely to be fairly small.
11. These and other questions were all defined for the *plant* level.
12. Other questions in our survey regarding product quality focused on rework rates, excess transportation (visits to customers to rectify product problems), inventory turns, employee hours per unit of output, and process capability kurtosis (a measure of the level and variance of plant output). Besides the scrap rate, only the rework rate was reported by a significant number of firms, and it was highly correlated with the scrap rate (rho = .9) among firms reporting both.
13. Many of these firms are auto parts suppliers, which are often required by major auto companies to meet product quality ratings in order to gain contracts with the companies. The exact number of such auto suppliers in our sample was, however, unclear from our data.
14. Standard errors on these three variables are approximately 6.63, .03, and .034 for grant recipients and 3.65, .025, and .027 for non-recipients, respectively. The standard error on the difference between estimates drawn from two independent samples is the square root of the sum of squared individual standard errors.
15. Since "upgrade" training for using new technology is not limited to new employees, we did not distinguish workers by seniority level in our survey.
16. Smaller firms actually received substantially larger grants per employee than did larger ones. The mean of training money per employee across firms, unweighted by firm size, is roughly $500; the figure in the next implicitly weights by firm size.
17. Standard errors on these means for firms receiving grants in 1988 are 3.31 in 1987 and 6.64 in 1988. For recipients of grants in 1989, the standard errors are 3.28 in 1988 and 6.84 in 1989.
18. There appears to be something of a positive trend in training levels in the absence of the grants, which affects comparisons of pre- and post-grant training levels for 1988 recipients. For instance, a regression of training hours in logs on a time trend for non-recipient firms generates a coefficient (and standard error) of .202 (.123). A similar regression on time dummies generates coefficients of .401 (.247) for non-recipients and .595 (.291) for 1988 recipients on training in 1989, and the two coefficients are not significantly different from each other. The

higher level of training in 1989 relative to 1987 for 1988 grant recipients may therefore be mostly a return to trend rather than a long-term, positive effect of the grant. We also note that non-recipient firms have somewhat higher training levels than do recipient firms in "off" years for the latter. Such differences are removed in our fixed-effects estimates below if they reflect permanent differences between those firms.

19. Own trends in the absence of grants for each set of firms are calculated by interpolating or extrapolating from values in those years when they did not receive grants. These calculations may understate the training increases purchased by grants, since the trends calculated for recipient firms are larger (though not significantly larger) than those observed for firms that never received grants (see note 18).
20. Higher employee output per unit cost will lead to higher sales in perfectly competitive product markets, where product demand is infinitely elastic at the equilibrium wage, and also in imperfect markets, where firms face downward-sloping product demand curves and sales rise in response to lower costs and prices. In the latter case, lower costs can be reflected in higher profits as well as higher sales. If product quality as well as quantity varies, sales may rise even without falling prices as product quality rises. Any increase in demand for the firm's output should shift out the firm's labor demand curve, thereby causing wages and employment to rise as well. Of course, these adjustments may take several years.
21. The "difference-in-differences" approach can also be estimated in levels with a lagged dependent variable as an additional regressor. But such an approach makes more sense for training effects on individuals, whose entry into training might have been based on previous earnings, than for the effects examined here. Furthermore, any serial correlation of errors in any equation would cause serious biases if such an approach were used.
22. These equations bear some resemblance to "partial adjustment" models that are estimated with time-series data.
23. If both recipients and non-recipients proceeded with their stated plans for implementing new technologies, and if these new technologies were (on average) comparable across the two groups of firms, then the direct effects of the technologies on worker output should not bias our estimated results here. Only if some non-recipients had to scrap or delay their plans might the unobserved technological changes result in biased estimates.
24. *F*-tests for differences in estimated effects across years generally showed few significant differences, thereby supporting our decision to pool these data.
25. Bassi (1984) demonstrates the importance of such tests for non-random selection when estimating returns to individual training in CETA (or JTPA) programs. The possibility of such selection is much less serious here because in this case, *unlike the case of individual participation in federal training programs,* grant recipiency is not conditional on low levels of training or otherwise weak performance in the pre-grant year. On the other hand, it is at least possible that the companies applying earliest for the grants—which are most likely to receive them—may be more likely to raise training levels and improve performance in the absence of such a grant.
26. Regressions of log training hours per employee in 1987 on grant status in 1988 generated a coefficient and standard error of -1.667 and 3.456, respectively. Regressions of changes in log training for 1987–88 on grant status in 1989 generated -5.454 (4.128); and, for 1988 training on 1989 grant status, the result was -8.694 (5.645). Comparable tests using other outcome variables also failed to show significant positive relationships between grant recipiency and pre-grant levels or changes in outcomes.
27. Equations comparable to these have been estimated using other quality measures reported by firms, such as the rework rate (see note 12) or a composite of scrap and rework rates. The results are qualitatively similar to, though somewhat weaker than, those for scrap alone (see Block *et al.* 1991). But to the extent that scrap and rework problems *are mutually exclusive categories* for any firm, the overall quality effect would be the *sum* of the two (if both were fully reported), and either alone would understate the true effect.

References

Barron, John, Dan Black, and Mark Lowenstein. 1989. "Job Matching and On-the-Job Training." *Journal of Labor Economics,* Vol. 7, No. 1, pp. 1–19.

Bartel, Ann. 1989. "Formal Employee Training Programs and Their Impact on Labor Productivity: Evidence from a Human Resource Survey." Working Paper, National Bureau of Economic Research.

Bassi, Lauri. 1984. "Estimating the Effect of Training Programs with Non-Random Selection." *Review of Economics and Statistics,* Vol. 66, No. 1, pp. 36–43.

Becker, Gary. 1975. *Human Capital.* New York: National Bureau of Economic Research.

Bishop, John. 1988. "Do Employers Share the Costs

and Benefits of General Training?" Unpublished paper, Cornell University.

____. 1981. "Employment in Construction and Distribution Industries: The Impact of the New Jobs Tax Credit." In S. Rosen, ed., *Studies in Labor Markets.* Chicago: University of Chicago Press, pp. 209–46.

Blanchflower, David, and Lisa Lynch. 1991. "Training at Work: A Comparison of U.S. and British Youth." Unpublished paper, Massachusetts Institute of Technology.

Block, Richard, Jack Knott, Marcus Cheatham, and Harry Holzer. 1991. *An Analysis of the MJOB Upgrade Training Grants Program.* Report to the Michigan Departments of Commerce and Labor.

Brown, Charles. 1990. "Empirical Evidence on Private Training." In Ronald Ehrenberg, ed., *Research in Labor Economics,* Vol. 11. Greenwich, Conn.: JAI Press, pp. 97–114.

Carnevale, Anthony, and Leila Gainer. 1989. *The Learning Enterprise.* Washington, D.C.: American Society for Training and Development.

Commission on Workforce Quality and Labor Market Efficiency. 1989. *Investing in People: A Strategy to Address America's Workforce Crisis.* Washington, D.C.: U.S. Department of Labor.

Creticos, Peter, and Robert Sheets. 1989. *State-Financed, Workplace-Based Retraining Programs.* Washington, D.C.: National Commission for Employment Policy.

Cutcher-Gershenfeld, Joel. 1991. "The Impact on Economic Performance of a Transformation in Workplace Relations." *Industrial and Labor Relations Review,* Vol. 44, No. 1, pp. 241–60.

Ehrenberg, Ronald, ed. 1990. *Employee Compensation and Firm Performance.* Ithaca, N.Y.: ILR Press.

Holzer, Harry. 1990. "The Determinants of Employee Productivity and Earnings." *Industrial Relations,* Vol. 29, No. 3, pp. 403–22.

Johnson, George, and James Tomola. 1977. "The Fiscal Substitution Effect of Alternative Approaches to Public Service Employment Policy." *Journal of Human Resources,* Vol. 12, No. 1, pp. 3–26.

Kleiner, Morris, et al. 1987. *Human Resources and the Performance of the Firm.* Madison, Wis.: Industrial Relations Research Association.

Lillard, Lee, and Hong Tan. 1986. *Private Sector Training: Who Gets It and What Are Its Effects?* Los Angeles: Rand Corporation.

Lynch, Lisa. 1989. "Private Sector Training and Its Impact on the Earnings of Young Workers." Working Paper, National Bureau of Economic Research.

Mincer, Jacob. 1978. *Education, Experience and Earnings.* New York: National Bureau of Economic Research.

____. 1989. "Job Training: Costs, Returns, and Wage Profiles." Working Paper, National Bureau of Economic Research.

Office of Technology Assessment. 1990. *Worker Training: Competing in the New International Economy.* Washington, D.C.: GPO.

Osterman, Paul. 1988. *Employment Futures.* New York: Oxford Press.

Parsons, Donald. 1990. "The Firm's Decision to Train." In Ronald Ehrenberg, ed., *Research in Labor Economics,* Vol. 11. Greenwich, Conn.: JAI Press, pp. 53–76.

Pelavin Associates. 1990. "Upgrade Training for Employed Workers." Report for the Office of Strategic Planning, Employment and Training Administration, U.S. Department of Labor.

Perloff, Jeffrey, and Michael Wachter. 1979. "The New Jobs Tax Credit: An Evaluation of the 1977–78 Wage Subsidy Program." *American Economic Review,* Vol. 67, No. 2, pp. 173–79.

CHAPTER 18

Changing the Geography of Opportunity by Expanding Residential Choice

Lessons from the Gautreaux Program

James E. Rosenbaum
Northwestern University

Abstract

The concept of "geography of opportunity" suggests that where individuals live affects their opportunities. While multivariate analyses cannot control completely for individual self-selection to neighborhoods, this article examines a residential integration program—the Gautreaux program—in which low-income blacks are randomly assigned to middle-income white suburbs or low-income mostly black urban areas.

Compared with urban movers, adult suburban movers experience higher employment but no different wages or hours worked, and suburban mover youth do better on several educational measures and, if not in college, are more likely to have jobs with good pay and benefits. The two groups of youth are equally likely to interact with peers, but suburban movers are much more likely to interact with whites and only slightly less likely to interact with blacks. The article considers how attrition might affect the observations and speculates about the program's strengths and pitfalls.

Introduction

The concept of "geography of opportunity" suggests that where individuals live affects their opportunities and life outcomes. Galster and Killen (1995) propose that individuals' lives can be profoundly changed if they move to environments that offer new opportunities. They suggest that geography influences social networks and normative contexts, and they review studies indicating influences on education, crime, and employment. However, since most geographic moves are chosen by the individual, research on geographic influences cannot completely control for individual effects. For instance, when surveys find a few low-income people in a middle-income neighborhood (or vice versa), one suspects that these are atypical low-income people or else they would not be there. What is needed to test Galster and Killen's propositions is a randomized experiment, but such experiments are unusual. This article describes such an experiment, the Gautreaux program in Chicago.

The Gautreaux program gives low-income blacks housing vouchers to move to many different kinds of communities, including white middle-income suburbs and low-income black city neighborhoods.[1] Because participants are assigned to city or suburban locations in a quasi-random manner, inferences can be made about the effects of residential moves. This article reports the program's impact on adult employment and on youth education, employment, and social integration. It considers how much attrition might affect the observations.

The Gautreaux program has become a model that other cities have sought to follow. Boston, Cincinnati, Dallas, and Hartford have initiated programs, and other cities have considered programs (Feins 1993; Fischer 1991). In addition, the national Moving to Opportunity program will test a version of the Gautreaux program in five cities across the United States.

Yet as we plan to replicate this program in other cities, we must consider what it is that makes this program work. A housing mobility program has many elements, and each may influence its operation and outcomes. The final section of this article examines what elements of the Gautreaux program are crucial to its success and what pitfalls future programs should avoid.

If the speculations presented here are generally on target, they suggest some guidelines for future programs. This is not to say that every program must make the same choices. Programs must base their choices on their own priorities and the ways they value various tradeoffs. However, all programs must consider these details, for they are likely to contribute to the effect of the program in important ways.

The Gautreaux program

The Gautreaux program is a result of a 1976 Supreme Court consent decree in a lawsuit against the U.S. Department of Housing and Urban Development (HUD) on behalf of public housing residents. The suit charged "that these agencies had employed racially discriminatory policies in the administration of the Chicago low-rent public housing program" (Peroff, Davis, and Jones 1979, 4). The Gautreaux program, administered by the nonprofit Leadership Council for Metropolitan Open Communities in Chicago, allows public housing residents, and those who were on the waiting list for public housing in 1981, to receive Section 8 housing certificates and move to private apartments either in mostly white suburbs or in the city of Chicago.[2] The program provides extensive housing services. Two full-time real estate staff find landlords willing to participate in the program. Then placement counselors notify families as apartments become available, counsel them about the advantages and disadvantages of moving, and take them to visit the units and communities. Since 1976, more than 5,000 families have participated, and more than half moved to middle-income white suburbs.

Because of its design, the Gautreaux program presents an unusual opportunity to test the effect of helping low-income people move to better labor markets, better schools, and better neighborhoods. The United States has little experience with economic and racial integration of neighborhoods. Racial and economic homogeneity is the rule in most neighborhoods, so we generally do not know how low-income blacks are affected by living in middle-income white neighborhoods. Moreover, even when exceptions exist, we must suspect that blacks who break the residential barriers and get into white neighborhoods are themselves exceptional people, so their subsequent attainments may reflect more about them than about the effects of neighborhoods. Therefore, when researchers study black employment in suburbs, it is hard to tell whether the suburbs increased black employment or whether the blacks who hap-

pen to live in suburbs are different, perhaps moving to the suburbs *after* getting a job (Jencks and Mayer 1989). Similarly, most studies of black achievement in suburban schools cannot tell whether black children's achievement is due to living in the suburbs or to some unmeasured family assets or values that drew these black families to the suburbs.

Gautreaux participants circumvent the ordinary barriers to living in the suburbs, not by their jobs, personal finances, or values, but by getting into the program. The program gives them rent subsidies that permit them to live in suburban apartments for the same cost as public housing. Moreover, unlike most black suburbanization—working-class blacks living in working-class suburbs—Gautreaux permits low-income blacks to live in middle-income white suburbs (Jencks and Mayer 1989). Participants move to any of more than 115 suburbs throughout the six counties surrounding Chicago. Suburbs with less than 70 percent whites were excluded by the consent decree, and very high rent suburbs were excluded by funding limitations of Section 8 certificates. Yet these constraints eliminate only a small proportion of suburbs. The receiving suburbs range from working class to upper middle class and are 30 to 90 minutes' drive away from participants' former addresses.

The program tries to move more than one family to a neighborhood to provide some social support, but it also avoids moving many families to one neighborhood. While the program mandates did not specify how many families could move to any location, the program tries to avoid sending disproportionate numbers to any one community, and in fact it has succeeded in this goal (Paul Fischer, unpublished tables, 1992). As a result, the program has low visibility and low impact on receiving communities.

Applying for the program is largely a matter of luck and persistent telephoning on registration day, since many more people try to call than can get through. The program also has three selection criteria: To avoid overcrowding, late rent payments, and building damage, it does not admit families with more than four children, large debts, or unacceptable housekeeping. None of these criteria is extremely selective, and all three reduce the eligible pool by less than 30 percent. Although these criteria make those selected an above-average group compared with housing project residents, they are not a "highly creamed" group. All are very low income blacks, are current or former welfare recipients, and have lived most of their lives in impoverished inner-city neighborhoods.

In any case, the program's procedures create a *quasi-experimental design.* While all participants come from the same low-income black city neighborhoods (usually public housing projects), some move to middle-income white suburbs, others to low-income black urban neighborhoods. In principle, participants have choices about where they move, but in practice, they are assigned to city or suburban locations in a quasi-random manner. Apartment availability is determined by housing agents who do not deal with clients and is unrelated to client interest. Counselors offer clients units as they become available according to their position on the waiting list, regardless of clients' locational preference. Although clients can refuse an offer, few do so, since they are unlikely to get another. As a result, participants' preferences for city or suburbs have little to do with where they end up.

Adult and child studies: Methods and sample

The next several sections of this article summarize studies of the Gautreaux program, comparing families moving to white middle-income suburbs ("suburban movers") with families moving to low-income black city neighborhoods ("city movers"). The city movers are a good comparison group for

judging the effects of the suburban move, since both groups meet the same selection criteria and get improved housing. But city movers are a particularly stringent comparison group because they receive better housing and move to better city neighborhoods than they had in the housing projects. We expect that housing-project residents would fare considerably worse than either of the Gautreaux groups. In effect, the suburban effects (relative to city movers) in this study may be considered lower bound estimates of the effects.

My colleagues and I at Northwestern University conducted three studies of this program. To examine adult employment, in the fall of 1988 we surveyed 332 adults and conducted detailed interviews with another 95.[3] The first study of children interviewed one randomly selected school-aged child (ages 8 to 18) from each of 114 families in 1982, and the second study followed up the same children in 1989, when they were adolescents and young adults, and examined their educational and employment outcomes.[4] As implied by the quasi-random assignment procedure, suburban and city movers are highly similar in most attributes in both samples (tables 1 and 2).

TABLE 1

Characteristics of the Adult Study Sample: City-Suburban Comparison

	City (n = 108)	*Suburb (n = 224)*
Years on Gautreaux	5.85	5.37
Age (years)	36.67	35.39
Age of youngest child (years)	9.56	7.85*
Number of children	2.51	2.56
Education premove (years)	11.68	11.91
Education postmove (years)	12.51	12.34
Marital status		
Married now (percent)	8.33	6.25
Never married (percent)	44.4	44.6
Getting AFDC (percent)	53.7	47.8
Long-term AFDC recipient[a] (percent)	68.5	59.8
Second-generation AFDC recipient (percent)	51.9	50.9

Note: Asterisk indicates significance level of difference between city and suburban samples, by chi-square or *t* test.
a. Ever received AFDC for five years of more.
*$p < 0.01$.

TABLE 2

Characteristics of the 1989 Children Sample: City-Suburban Comparison

	City	*Suburb*
Age (years)	18.2	18.8
Male (percent)	45.5	56.8*
Mother not married (percent)	88	86
Mother's education postmove (years)	12.03	12.09
Mother finished high school (percent)	43	47

Note: Asterisk indicates significance level of difference between city and suburban samples, by chi-square or *t* test.
* $p < 0.01$.

Results from the adult study: Will low-income blacks get jobs in the suburbs?

There are several reasons to expect that low-income blacks may not get jobs in the suburbs. After living in low-income environments for many years, these mothers and children may have motivational problems that prevent them from doing well even after their opportunities improve (Lewis 1968; Mead 1986). Discrimination by employers or lack of skills may also prevent low-income blacks from getting jobs. In addition, Gautreaux adults were educated in poor urban schools, and many lack job training or job experience. Furthermore, virtually all the mothers in Gautreaux have received public aid (most for five years or more), many have never had a job, and half grew up in families on public aid.

TABLE 3

Percent of Respondents Employed Postmove by Premove Employment for City and Suburban Movers

	Premove Status					
	City			***Suburb***		
Postmove Status	*Employed*	*Unemployed*	*Total*	*Employed*	*Unemployed*	*Total*
Employed	42 (64.6%)	13 (30.2%)	55	106 (73.6%)	37 (46.2%)	143
Unemployed	23 (35.4%)	30 (69.8%)	53	38 (26.4%)	43 (53.8%)	81
Total	65	43	108	144	80	224

Note: Numbers in parentheses are column percentages.

Premove and postmove employment status of city and suburban movers is compared in table 3. Suburban movers were more likely to have jobs than city movers. Although both groups started from the same baseline (60.2 percent of city movers and 64.3 percent of suburban movers were employed premove), after moving, suburban movers were more than 25 percent more likely to have a job than city movers. While 50.9 percent of city movers had a job after moving, 63.8 percent of suburban movers did.

Among respondents who had been employed before, suburban movers were about 14 percent more likely than city movers to have a job after moving. In contrast, for suburban movers who had never been employed before their move, 46.2 percent found work after moving to the suburbs, while the comparable figure for city movers was only 30.2 percent. For this group of hard-core unemployed, suburban movers were much more likely to have a job after moving than city movers.[5]

City and suburban movers did not differ in hourly wages or number of hours worked per week (table 4). Among those who had a job both before and after moving, both city and suburban movers reported gains in hourly wages and no change in hours worked.[6] The roughly 20 percent gain in wages for both suburban and city movers may represent gains from moving out of housing projects, but since we lack a control group (individuals who are similar to those selected in Gautreaux but remained in housing projects), we have no basis for testing this explanation. Of course, attrition may also contribute to these gains, but attrition is likely to be small, and it can cut both ways: Some attrition comes from people whose wages surpass program income limits.

When asked how the suburban move helped them get jobs, all suburban participants mentioned the greater number of jobs in the suburbs. Improved physical safety was the second most mentioned factor. Adults reported that they did not work in the city because they feared being attacked on the way home from work, or they feared that their children would get hurt or get in trouble with gangs. The suburban move allowed mothers

TABLE 4

City and Suburban Comparison of Wages and Hours Worked

	Premove Mean	*Postmove Mean*	*t*	*p*
City movers, postmove earners (n = 55)				
Hourly wages ($)	5.04	6.20	6.52	0.01
Hours per week	33.27	31.92	-0.60	0.55
Suburban movers, postmove earners (n = 143)				
Hourly wages ($)	4.96	6.00	6.50	0.01
Hours per week	33.62	33.39	-0.60	0.55

to feel free to go out and work. Many adults also mentioned that positive role models and social norms inspired them to work. These comments support Wilson's contention about the importance of role models and social norms (Wilson 1987). Seeing neighbors work, Gautreaux adults reported that they felt that they too could have jobs, and they wanted to try. In the city, few adults saw neighbors working.

In sum, the employment rates of suburban movers surpassed those of city movers, particularly for those who had never before had jobs. Whatever prevented some people from being employed in the past—lack of skills or lack of motivation—was not irreversible, and many took jobs after moving to the suburbs.

Results from the first study of children: Will early disadvantages keep children from benefiting from suburban schools?

Housing moves may affect children even more than adults, since children are at a formative stage and are still acquiring education. Moreover, being less mature, children may have even more difficulty coping with the challenges posed by a suburban move. The obstacles are similar to those for adults. Children's low-income background may make them less prepared or less motivated than middle-income suburban children, they may have attitudes and habits deemed undesirable by suburban teachers and employers, or racial discrimination may deny them full access to suburban resources. For any or all of these reasons, they may have lower achievement than their city counterparts who do not face these barriers. On the other hand, suburban movers will benefit from better educational resources and greater employment prospects in the suburbs, and their fellow suburban students may serve as role models for achievement. Of course, we do not know which process will operate or, if both do, which will win out.

Given the children's initial poor preparation in city schools and their social disadvantage, we wondered how they would do in the suburban schools. In 1982, we studied how the Gautreaux program affected children, comparing Gautreaux children who moved within the city and those who moved to the suburbs (Rosenbaum, Kulieke, and Rubinowitz 1988; Rosenbaum, Rubinowitz, and Kulieke 1986).[7] The two groups were similar in average age, proportion of females, and mothers' education. The families were predominantly female-headed in both the suburban (86 percent) and city (88 percent) groups.

We found that suburban movers initially had difficulties adapting to the higher expectations in the suburban schools, and their grades suffered in the first years in the suburban schools. However, by 1982, after one to six years in the suburbs, their grades and school performance (judged by their mothers) were the same as those of city movers. In addition, suburban movers had smaller classes, higher satisfaction with teachers and courses, and better attitudes about school than city movers. Although the mothers noted instances of teacher racial bias, the suburban movers were also more likely than city movers to say that teachers went out of their way to help their children, and they mentioned many instances of teachers giving extra help in classes and after school.

It is hard to measure academic standards, and the first study had no systematic indicator. Yet the suburban movers clearly felt that the suburban schools had higher academic standards. They reported that the city teachers did not expect children to make up work when they were absent, to do homework, to know multiplication in third grade, or to write in cursive in fourth grade. Passing grades in the city did not indicate achievement at grade level, and even city students on the honor roll were sometimes two years behind grade level.

These mothers were in a good position to notice these differences when their children moved from the city to suburban schools. One mother commented that the suburban school "said it was like he didn't even go to school in Chicago for three years, that's how far behind he was. And he was going every day and he was getting report cards telling me he was doing fine" (Rosenbaum, Kulieke, and Rubinowitz 1988, 32). Indeed, another mother reported an empirical test:

> The move affected my child's education for the better. I even tested it out. … [I] let her go to summer school by my mother's house [in Chicago] for about a month. … She was in fourth grade at that time. … Over in the city, they were doing third grade work; what they were supposed to be doing was fourth grade. The city curriculum seemed to be one to three years behind the suburban schools. (p. 32)

While many suburban movers seemed to be catching up to the higher suburban standards by the time of the interviews, most had been in the suburbs only a few years, and most were still in elementary school, so it was hard to know how they would do later. Many of these children were still struggling to catch up, and it was not clear whether they would succeed. Therefore, we were eager to do a follow-up study to see how things were turning out for these children.

Results from the second study of children: Education and employment

To study later outcomes, we interviewed the same children and their mothers in 1989.[8] By this time, the children had an average age of 18.

However, before turning to those results, I will describe the schools that the youth attended. In 1990, the Illinois Department of Education collected average standardized test scores for all schools in the state. For the schools attended by our sample, the suburban schools' average 11th-grade reading test score (259) was just above the state average (250) but significantly higher than the city schools' average (198). On the ACT examination, the college admissions test most often taken in Illinois, suburban schools' scores (21.5) were close to the state average (20.9), but significantly higher than the city schools' scores (16.1). Moreover, there was almost no overlap between the scores of city and suburban schools these children attended. While less than 6 percent of the city sample attended schools with ACT averages of 20 or better (i.e., roughly the national average), more than 88 percent of the suburban sample attended such schools. Just as the 1982 study suggested higher standards in suburban elementary schools, these results indicate that the higher standards in the suburbs continued in high school.

Of course, higher standards create new challenges as well as new opportunities. The suburban movers face much higher expectations than they have been prepared for in the city schools. The higher levels of achievement in suburban schools may be a barrier to poorly prepared students and may lead to increased dropping out, lower grades, and lower tracks for those still in school and to less college attendance and less employment for those over age 18. The results of this study contradict those expectations (table 5).

Dropout rates

Lower dropout rates for suburban high schools may encourage Gautreaux children who move to the suburbs to stay in school. On the other hand, higher academic standards or discrimination based on race or class may discourage suburban Gautreaux children and make them more likely to drop out. Results from the second study of children supported the first scenario, as more city

TABLE 5
Youth Education and Job Outcomes: City–Suburban Comparison (percent)

Outcome	*City*	*Suburb*
Dropped out of school	20	5*
College track	24	40**
Attend college	21	54***
Attend four-year college	4	27**
Employed full-time (if not in college)	41	75***
Pay under $3.50/hour	43	9***
Pay over $6.50/hour	5	21***
Job benefits	23	55***

Note: Asterisks indicate significance level of difference between city and suburban samples, by chi-square or *t* test.
* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.025$.

movers (20 percent) than suburban movers (5 percent) dropped out of high school.

Grades

Although test scores were not available for individual respondents, grades provide a good indication of how students are achieving relative to their peers and whether teachers judge students' work acceptable. We found that suburban movers had virtually the same grades as city movers (a C-plus average in city and suburbs). Since the national High School and Beyond survey of high school sophomores indicates that suburban students average about half a grade lower than city students with the same achievement test scores, the grade parity of the two samples implies a higher achievement level among suburban movers (Rosenbaum and Kaufman 1991).

College preparatory curricula

Most high schools offer different curricula to college-bound and non-college-bound youth, and these curricula affect college opportunities (Rosenbaum 1976, 1980). Researchers find that blacks are underrepresented in college tracks in racially integrated schools (Coleman 1966; Oakes 1985; Rosenbaum and Presser 1978). Indeed, after being desegregated, the Washington, DC, public schools initiated a tracking system, which a court ruled to be undercutting integration (*Hobson v. Hansen* 1967). Given the higher standards and greater competition in suburban schools, we might expect suburban movers to be less likely than city movers to be in college-track classes. The results showed the opposite: Suburban movers were more often in college tracks than city movers (40 versus 24 percent).

College attendance

Higher suburban standards might be a barrier to these youth's attending college. The results indicate the opposite: Suburban movers had significantly higher college enrollment than city movers (54 versus 21 percent).

Four-year colleges

The type of college is also important. Four-year colleges lead to a bachelor's degree, two-year junior or community colleges lead to an associate's degree, and trade schools lead to a certificate. Moreover, while transfers to four-year colleges are theoretically possible, in fact trade schools almost never lead to four-year colleges, and two-year colleges rarely do. Only 12.5 percent of students in the Chicago city colleges ultimately complete a four-year college degree—less than half the rate of some suburban community colleges in the area (Orfield 1984).

Among the Gautreaux youth attending college, almost 50 percent of the suburban movers were in four-year institutions, whereas only 20 percent of the city movers were. Of those not attending four-year institutions, two-thirds of the suburban movers

were working toward an associate's degree, while just half of the city movers were.

Youth's jobs

For the youth who were not attending college, a significantly higher proportion of the suburban youth had full-time jobs than city youth (75 versus 41 percent). Suburban youth also were four times as likely to earn more than $6.50 per hour than city youth (21 versus 5 percent). The suburban jobs were significantly more likely to offer job benefits than city jobs (55 versus 23 percent).

Results from the second study of children: Harassment and social integration

Will residential integration lead to harassment and rejection of youth?

Blacks are significantly more isolated than either Hispanics or Asians (Massey and Denton 1987). Research also documents extensive antagonism to racial integration. While the majority of whites have become increasingly supportive of racial integration in principle, the majority remain opposed to any government intervention to promote such integration (Schuman and Bobo 1988). Blacks moving into predominantly white areas have faced threats, physical attacks, and property damage (Berry 1979). Throughout the past several decades, black families who moved into white neighborhoods of Chicago were driven from their homes by racial violence (Squires et al. 1987). Yet incidents of harassment, while dramatic, may not reflect the views of all residents, and other neighbors may willingly interact with black newcomers. We can examine the harassment, threats, and fears that blacks face in white schools in which they are a racial and socioeconomic minority.

We expected that the suburban youth would experience more harassment than the city movers. The most common form of harassment was name-calling. In the suburbs, 51.9 percent of the Gautreaux youth reported at least one incident in which they were called names by white students, while only 13.3 percent of the city movers experienced name-calling by whites. Of course, there are few whites in the urban schools to call anyone names. However, 41.9 percent of the city movers experienced name-calling by blacks. As hypothesized, city movers do receive significantly less harassment than suburban movers, but the city movers experience a great deal of name-calling, too.

A second, more serious, indicator of harassment was measured by asking respondents how often they were threatened by other students. As expected, many suburban movers were threatened by whites: 15.4 percent reported being threatened by whites a few times a year or more. However, 19.4 percent of city movers were threatened this often by blacks. Moreover, when we consider those who were threatened at least once a year (by blacks or whites), city movers are as likely to receive a threat as suburban movers (22.7 percent city versus 21.2 percent suburb).

A third indicator of harassment, the most serious, is whether youth were hurt by other students. When asked how often they were actually hurt by others at school, very few members of either group reported such incidents. A similar proportion of both city and suburban movers say they have never been hurt by other students (93.5 versus 94.1 percent).

Social acceptance: Will residential integration lead to social integration?

Given the daily headlines about troubled race relations in American society and schools, social integration might seem hopeless. But daily life is too mundane to make the headlines, and daily life may tell a very different story. The following discussion looks at whether these black youth experience acceptance, friendships, and positive interactions with white

classmates, and it assesses the relative frequency of positive and negative interactions.

School desegregation has been extensively studied (Gerard and Miller 1975; Hawley 1981; Patchen 1982; St. John 1975). However, this form of desegregation has some attributes that may limit its benefits. Because blacks rarely live near whites, many of the school desegregation programs entail special busing efforts, and a busload of students may create high visibility, backlash, and stigma. In addition, as children spend long periods every day riding together on a bus, these commutes reinforce blacks' sense of togetherness and their separateness from those who live near the school. Moreover, the logistics of commuting make after-school activities more difficult. Thus, busing as a method of desegregation creates its own limits on racial interaction.

In contrast, the Gautreaux program creates both residential and school integration. As a result, children live in apartment buildings occupied largely by middle-income whites; they arrive in the suburban schools as community residents, not as outsiders in a busing program; and they come to school in the same buses as their white neighbors. Moreover, the program accomplishes residential integration with little visibility and in small numbers that raise little threat, thus reducing the likelihood of backlash and stigma.

Youth in this program also must face an additional barrier: socioeconomic differences. Researchers know even less about socioeconomic integration than they know about racial integration. Gautreaux children face both kinds of barriers simultaneously. These low-income blacks enter schools and communities that are overwhelmingly white and middle class. Even the blacks they meet are different, since their families are middle class.

Given these barriers, observers have worried that youth in such a program would remain socially isolated (Yinger 1979). Having spent more than six years in all-black urban housing projects, these children have learned habits and tastes different from those of their classmates, they have fewer economic resources than their classmates, and their skin color is different from that of most of their classmates. There is a great risk that these youth will have difficulty being accepted by their suburban, middle-class, white classmates.

Several questions in the survey and interview were designed to measure the children's sense of social acceptance. Both city and suburban movers tended to agree somewhat with the statement, "I feel I am a real part of my school" (on a five-point scale from strong agreement to strong disagreement), and there were no significant differences between the groups (city mean, 3.55; suburban mean, 3.37). On the item "Other students treat me with respect," the suburban movers had more positive responses than the city movers, although the difference was not significant (city mean, 3.93; suburban mean, 4.00). We asked the children how they believed others saw them in a series of questions:

1. "Are you considered part of the in-group?"
2. "Do others see you as popular?"
3. "Do others see you as socially active?"
4. "Do others think you do not fit in?"

For each of these items, no significant differences were found between the answers of city and suburban movers (table 6). Both groups showed positive social integration for all questions.

Contrary to our expectation, the suburban movers were just as accepted by their peers as the city movers. The majority of the children in both groups felt that they fit into their schools socially and were regarded by others as at least somewhat socially active and popular.

Friendships. We expected that the suburban movers might have fewer friends than city

TABLE 6

Frequency of Responses to Questions about Social Integration (percent)

Code	*Suburb*	*City*
Are you considered part of the in-group?		
($t = 0.36$, $df = 67.6$)	($n = 49$)	($n = 31$)
Very much	32.7	32.3
Somewhat	44.9	51.6
Not at all	22.4	16.1
Do other see you as popular?		
($t = 0.95$, $df = 59.52$)	($n = 50$)	($n = 31$)
Very much	36.0	29.0
Somewhat	60.0	61.3
Not at all	4.0	9.7
Do others see you as socially active?		
($t = 0.18$, $df = 63.40$)	($n = 50$)	($n = 31$)
Very much	46.0	48.4
Somewhat	44.0	41.9
Not at all	10.0	9.7
Do others think you do not fit in?		
($t = 1.13$, $df = 52.42$)	($n = 50$)	($n = 30$)
Very much	2.0	3.3
Somewhat	16.0	26.7
Not at all	82.0	70.0

Note: No differences between city and suburban samples are significant at $p < 0.05$.

movers. Given that the suburbs were overwhelmingly white, the suburban movers came in contact with fewer black peers than city movers. However, suburban movers had almost as many black friends as city movers. The mean number of black friends in the suburbs was 8.81, while the mean in the city was 11.06 (difference not significant).

The suburban movers had significantly more white friends than city movers. The mean number of white friends was 7.37 for suburban movers and 2.37 for city movers ($t = 4.71$; $p < 0.0001$). While only 17.3 percent of the suburban youth reported no white friends, 56.3 percent of the city sample did ($t = 3.43$; $p < 0.001$). Only one of the city movers and one of the suburban movers reported having no friends at all.

Interactions. Suburban youth spent significantly more time with white students outside of class than city movers (table 7). Compared with city movers, the suburban movers more often did things outside school with white students, did homework with white students, and visited the homes of white students. When asked how friendly white students were, the suburban movers again were significantly more positive than the city movers ($t = 3.24$; $p < 0.002$). When the same questions were asked about socializing with black students, no significant differences existed between city and suburban movers (table 8).

To get an overview, two index variables were computed from the summed responses to each of the three items for interactions with whites and for interactions with blacks. The findings suggest that the suburban movers divided their time almost equally between blacks and whites, while the city movers spent significantly more time with blacks than with whites (table 9). The experience of the suburban movers seems to reflect a more racially integrated peer network, despite the small numbers of blacks in suburban schools. As one suburban mover reported to us, "We went into a new school and had the opportunity to be with white people, Indian people, just a mix of races and actually get to know people and have people get to know you."

Are harassment and acceptance inversely related?

News accounts of racial harassment are particularly disturbing because the stories often carry the implication that harassment and threats reflect rejection by the entire community. Sometimes that may be true, but it seems possible that it is not true.

Our results indicate that negative behaviors are associated with each other: White name-calling is associated with white threats ($r = 0.53$, $p < 0.01$). Positive behaviors are also

TABLE 7

Frequency of Activities Involving White Students (percent)

Code	*Suburb*	*City*
How often do white students do things with you outside of school?		
($t = 3.65$; $p < 0.001$)	($n = 52$)	($n = 30$)
Almost every day	44.2	6.7
About once a week	13.5	16.7
About once a month	1.9	16.7
A few times a year	23.1	10.0
Never	17.3	50.0
How often do white students do schoolwork with you?		
($t = 2.92$; $p < 0.005$)	($n = 52$)	($n = 30$)
Almost every day	40.4	23.3
About once a week	21.1	16.7
About once a month	21.2	13.3
A few times a year	9.6	0.0
Never	7.7	46.7
How often do white students visit your home or have you to their home?		
($t = 3.75$; $p < 0.0001$)	($n = 52$)	($n = 29$)
Almost every day	28.8	6.9
About once a week	25.0	10.3
About once a month	7.7	13.8
A few times a year	19.2	17.2
Never	19.2	51.7

associated with each other: Doing activities with whites is associated with visiting with whites in their homes ($r = 0.85$, $p < 0.01$).

However, negative behaviors do not predict an absence of positive behaviors. In fact, the experience of the suburban movers indicates that the two are not usually associated, and they are sometimes positively correlated. Suburban movers who report being threatened by whites are slightly (but not significantly) *more* likely to participate in school activities ($r = 0.11$), to do activities with whites after school ($r = 0.05$), or to visit with whites in their homes ($r = 0.09$). Those reporting being called names by whites are also slightly (but not significantly) *more* likely to do activities with whites after school ($r = 0.08$) and to visit with whites in their homes ($r = 0.17$).

These correlations are not statistically significant, but they are substantively very important. They indicate that many of the same individuals who are being threatened and harassed by whites are also being accepted by whites, interacting with whites, going to whites' homes, and participating in school activities with whites. That does not make the threats and name-calling pleasant, but it makes it easier for these youth to feel a part of white suburban schools.

How much does attrition reduce program effects?

One criticism of the Gautreaux studies is the absence of information on the dropouts from the program. If substantial numbers drop out, then the results could be quite different. In the survey of heads of households, we made extensive efforts to locate everyone, yet we could locate only about two-thirds of the sample, and the rate was somewhat lower for suburban movers (60 percent). Of course, most studies have great difficulties in locating low-income people.

Note that people may not be found for either positive or negative reasons. Some may have left the program because they got good jobs and their incomes exceeded the program limit. Indeed, the study located some people who had such successful outcomes. On the other hand, some may have left the program because of negative outcomes: dissatisfaction, poor jobs, or poor children's outcomes. There is no way of knowing what percentage of those we did not find had positive or negative outcomes. We can only conclude that our results may either understate or overstate the program's effects.

Although we do not know exactly how many people left the suburbs, we can still estimate how much attrition could affect our results. Since we located 60 percent of our

TABLE 8

Frequency of Activities Involving Black Students (percent)

Code	*Suburb*	*City*
How often do black students do things with you outside of school?		
(t = 0.70)	(*n* = 52)	(*n* = 31)
Almost every day	59.6	54.8
About once a week	19.2	32.3
About once a month	5.8	6.5
A few times a year	11.5	6.5
Never	3.8	0.0
How often do black students do schoolwork with you?		
(*t* = 1.34)	(*n* = 52)	(*n* = 31)
Almost every day	46.2	64.5
About once a week	25.0	16.1
About once a month	7.7	9.7
A few times a year	11.5	0.0
Never	9.6	9.7
How often do black students visit your home or have you to their home?		
(*t* = 0.43)	(*n* = 52)	(*n* = 31)
Almost every day	50.0	25.8
About once a week	25.0	45.2
About once a month	3.8	22.6
A few times a year	15.4	3.2
Never	5.8	3.2

Note: No differences between city and suburban samples are significant at $p < 0.05$.

TABLE 9

Comparisons of Index Variables Measuring Time Spent with Black Friends versus Time Spent with White Friends (scale of 1 to 15)

	Suburb (n = 60)		***City (n = 38)***	
Index Variable	*Mean*	*SD*	*Mean*	*SD*
Time with black friends	12.02	3.09	12.45	2.17
Time with white friends	10.41	3.64	6.89	3.48
	t = 3.05, p < 0.003		t = 9.04, p < 0.001	

Note: SD = standard deviation.

suburban movers in the suburbs, 40 percent is the upper limit for attrition. It is not likely that all of these 40 percent moved back to the city. But as a mental exercise, we can assume that they all did and that their success rates are the same as the city movers' rates. These assumptions allow us to see how much suburban attrition might reduce our findings.

We can use these hypothetical assumptions to recalculate the adult employment findings. The 224 suburban adults came from an original pool of 373 adults (224/0.60), and the additional 149 individuals we did not find will be assumed to have a city rate of employment of 50.9 percent, so 76 would have been employed. Adding those hypothetical 76 to the actual 143 adults with jobs and dividing by the total of 373 yields an employment rate of 58.7 percent. Thus, while the actual findings for employment were 50.9 percent city, 63.8 percent suburbs, the hypothetical difference is 50.9 percent city, 58.7 percent suburbs—under the extreme (and unlikely) assumption that all unfound people returned to the city and subsequently experienced the same employment rates as those who originally moved to the city. If we use a more realistic assumption, that half of all unfound people returned to the city, the results would be 50.9 percent city, 61.3 percent suburbs.

In sum, even making extreme assumptions yields a substantial suburb-city difference in employment. More realistic assumptions yield a difference that is not far from the original finding. Moreover, most of the children's outcomes show even larger differences, so if these procedures are applied to the children's outcomes, the most extreme adjustments still yield impressive benefits to the suburban movers.

It would be desirable to get better information on the people we did not find. In particular, we would like to know how many returned to the city, their reasons for moving, and the relative distribution of positive and

negative reasons. Indeed, such data could have practical applications in helping new movers cope with the suburban move. However, that information is not likely to alter our conclusions. Even a worst-case scenario yields substantial benefits to suburban movers.

Implementation issues

According to Murphy's Law, anything that can go wrong will go wrong. A corollary of Murphy's Law might suggest that there are few ways to do things right but many ways to do things wrong. The rest of this article presents speculations about why this program has such strong effects and, by extension, how it could be done wrong.

In 1985, at the height of the Reagan years, the results of the first study of the Gautreaux program went out over the news wires and were ignored by all but a few local newspapers. In 1988, as the luster of the Reagan years was subsiding, a journalist "discovered" the Gautreaux program, and many major newspapers carried the story of the 1985 study (later studies were still ongoing). Today, stories about housing mobility appear regularly in every major newspaper. Housing mobility is beginning to be seen as a panacea.

Obviously, I will not say anything to discourage the enthusiasm for this approach. My research findings have persuaded me of its value. However, "housing mobility" is not a single entity, and not all forms of housing mobility will have the same effects. I am voicing these cautions because I do not want the current enthusiasm to lead us to forget the important details that make the program successful. If the details are done badly, then the aphorism may be an apt warning: "The devil is in the details."

While the Gautreaux program had remarkably positive effects, those results probably depend on the program's having done a lot of things right. This is not to say that the Gautreaux program was perfect or did everything it could to increase success. As I will note, the program intentionally made some choices to lower costs and reduce potential benefits.

Alexander Polikoff (the lawyer who took the case to the U.S. Supreme Court and helped design the consent decree) and Kale Williams and his staff at the Leadership Council (who implemented the Gautreaux program) made many big and small decisions about the form the Gautreaux program would take. As Williams reports, some practices emerged without conscious decisions, but they nonetheless became enduring features of the program (personal communication, March 3, 1994). Obviously, given the observed outcomes, they did something right. The following analysis examines the many details that I believe were relevant to these outcomes. Of course, these are speculations; it is not possible to analyze the separate effects of various program components.

A federal program, Moving to Opportunity (MTO), is now being started to test the Gautreaux approach in five cities across the United States. In addition, many localities are adapting aspects of this program. These programs need to make many specific decisions about how to implement the program in their particular settings, and these decisions need not be the same as those made by the Gautreaux administrators. I hope that these speculations may help these programs think through their decisions.

When considering the replicability of the Gautreaux experience, it is important to bear in mind that Chicago is a unique setting. It is one of the most racially segregated metropolitan areas in the nation (Farley and Frey 1992). The city also has a very large population living in extremely poor and distressed neighborhoods (Kasarda 1993). How these extreme conditions affected program outcomes is unknown, so results from a similar program in a different environment may be different.

Central features of the Gautreaux program

A useful way to consider the details of the Gautreaux program is to consider hypothetical alterations of what was done and to speculate about their implications. To do this, one must ask how the Gautreaux program could be done differently. A program that moves people to new places can vary four features: people, places, services, or expectations. Ignoring these possible variations, some people have assumed that any program that moves any people to any suburbs will have results similar to those of the Gautreaux program. This is probably not true. Indeed, the program would probably have worked very differently if it had varied any of these features. The following sections discuss why these features should be important considerations in program design.

People who may not match the program's demands

There is a tradeoff between seeking to move the maximum number of people to better housing and seeking to move only the kind of people who are likely to benefit. The Gautreaux program chose the latter strategy, and it developed clear criteria.

Many kinds of people might have difficulty benefiting from the program: people who cannot handle the rigors of the program, the costs, or the behavioral norms required. The Gautreaux program took steps to help improve the chances of getting people who were appropriate. It selected people who had good rent-payment records and no large debts that would prevent their paying rent, and those who did not cause property damage that would lead to their eviction from suburban apartments. These two criteria eliminated about 12 and 13 percent of otherwise eligible families, respectively. The program also excluded families with four or more children, who would not fit into the available two- to three-bedroom apartments, but this criterion eliminated only about 5 percent of eligible families.

These steps seem reasonable, but not all social programs take such steps. Well-intentioned social programs sometimes seek to help the "most needy." This is certainly a worthy group to serve, but it may be ill suited to a mobility program. If the most needy have large outstanding debts or lack experience at regular budgeting to pay rent, then they are likely to have trouble paying rent regularly. If they do not know how to take care of apartments, or if they have violent family members or visitors, then property damage will be likely. Not only will these circumstances lead to eviction, but even one such eviction can give the program a bad reputation, leading many landlords to avoid taking any more program participants.

Another related problem can arise from taking the most needy. One program took people on the public housing waiting list—people who generally were desperate for any kind of housing. According to informal reports (I know of no systematic information on this program), one consequence was that these people were not very patient about waiting until an appropriate integrated neighborhood was available; they wanted a roof over their heads immediately.

Programs that are concerned about community acceptance might use other criteria. While the Gautreaux program did not do this, some have suggested that such programs should screen out people with felony records. Some protesters against the Baltimore residential mobility program (MTO) were quoted as being concerned about felons moving into their neighborhoods.

The Gautreaux program seeks to integrate low-income people into the private sector housing market. Therefore, the program must select people who can meet the expectations of the private sector: regular rent payment and lack of property destruction. It

must also select people who can wait until appropriate housing becomes available.

One important side effect of this strategy is that the Leadership Council came to be known and trusted by landlords. The Leadership Council's selection procedures became an unofficial warranty of participants' capabilities. Informal reports suggest that even landlords who harbored prejudices against low-income blacks felt they could trust the people selected by the Leadership Council. In effect, the Leadership Council informally certified participants in ways that overcame landlords' prejudices, much as some job training programs serve as warranties of people's job skills and work habits (Kariya and Rosenbaum forthcoming).

Avoiding selecting people who would benefit anyway

On the other hand, the program must not be so selective that it chooses only people who would succeed without it. On the basis of the above-noted criteria that eliminated 12, 13, and 5 percent of applicants, I estimate that the Gautreaux program eliminated about one-third of eligible applicants. The program was selective, but not so selective that it can be called "creaming."

As an aside, people have commented on the Gautreaux families featured on television reports, such as those on *60 Minutes, CNN,* and *ABC World News Tonight.* Those families are articulate, and observers have noted that they do not seem like typical housing project residents. That is true, but they are not typical Gautreaux participants, either. The reason articulate families appear on television is that television producers want to feature articulate families. Average families do not make good TV.

It is noteworthy that the Gautreaux program did not apply more restrictive criteria. For instance, the program did not check on previous neighbor's complaints, children's problems in school, or mothers' work histories. These are all plausible criteria. They would probably have improved the success rate of the program, but the Gautreaux program did not choose to use them.

Such selection criteria were apparently unnecessary. The program had quite good results without them. While the Gautreaux program suggests that some selection criteria may be necessary, it also indicates that some other potential criteria are not necessary for generally successful outcomes.

Places that may not match the program's demands

There is a tradeoff between seeking to move the maximum number of people to better housing and seeking to move people to only the right kinds of places. The Gautreaux program chose the latter strategy, and it developed clear criteria for defining the right kinds of places.

Places can be inappropriate in a number of ways: They can be too expensive, too racist, too vulnerable to white flight, or too vulnerable to flight by the black middle class. The Gautreaux program was sensitive to all of these potential problems.

If housing is too expensive, the program is subject to political criticism. In addition, increased costs per unit limit the number of families who can be helped. The maximum rent limitation is built into the Section 8 program. While some waivers were obtained to include some upper-middle-class communities, some suburbs still had rents that were beyond the scope of Section 8.

In the six-county area surrounding Chicago, there are few places where Gautreaux participants do not live, and most of those places are in the more distant outreaches of this large area, where two-hour driving times, not expense, reduced participation. Fewer than 10 suburbs had no housing units with rents within Section 8 guidelines.

Participants live in more than 115 suburbs surrounding Chicago, and they are widely distributed across these communities. Even upper-middle-class suburbs tend to have pockets of more reasonable rents amid their high-priced housing. While some economic segregation occurred within some suburbs, this was not necessarily class segregation, since many participants reported that they interacted with young highly educated neighbors (apparently people who started their families in apartments with affordable rents).

The Gautreaux program eliminated only two communities because of "intractable racism." Cicero and Berwyn have active Ku Klux Klan groups and a long history of violent attacks on black residents. The program sent no families to those areas.

The Gautreaux program also eliminated communities considered to be too near a "tipping point" (a level of black population above which white residents might feel threatened and flee). While some research has tried to specify the tipping point at around 7 percent black (Farley, Bianchi, and Colasanto 1979), it seems likely that the proportion depends on historical conditions and the rate at which the community has arrived at its current composition. A community that has quickly shifted from 0 to 7 percent is likely to be much more upset and prone to resist integration than one that has shifted over a decade. In many metropolitan areas, there are communities that went from predominantly white to predominantly black in less than a decade. To avoid contributing to such a process, the Leadership Council sent no Gautreaux families to suburbs thought to be near a tipping point.

The program was also careful not to send too many families to any one location in a single year. A town that receives 10 families each year over 10 years will react quite differently from one that receives 100 families in a single year.

As a court-ordered desegregation program, the Gautreaux program was not allowed to send black families to predominantly black suburbs. Other programs, such as MTO, are allowed to do so and thus risk contributing to another kind of tipping point. Black middle-class residents might be alarmed if they felt that MTO was sending too many low-income residents to their suburbs.

It would be all too easy for this to happen. If a program has many families needing to be placed and only a few communities where those families feel comfortable, the easiest way to place more families is to send them to the same locations. This process is similar to real estate "blockbusting," in which an agent instigates neighborhood panic to maintain a high volume of families moving into an area. Of course, we do not expect MTO staff to have such a motivation, but they could inadvertently contribute to such blockbusting. It is easier for a staff person to fall into a habit of sending people to a few well-known places than to discover new ones. And while the program may see its 300 families as a small number relative to the suburb's population of 50,000, the community can see a sudden annual influx of 300 families as a threat, particularly if rental housing is concentrated in one part of town.

Participants' choices also contribute to such a process. Given a choice about where to move, low-income blacks are likely to choose suburbs with substantial black populations where they do not have to worry about racial harassment. Yet this free-choice process is likely to lead to racial and economic resegregation, as whites and middle-class blacks fear the worst and flee. Indeed, the regular Section 8 program relies on free choice, and it sometimes creates these results. Similarly, the national housing voucher experiment found that given free choices about where to move, most recipients moved to areas very similar to the areas they left (Cronin and Rasmussen 1981).

Yet large influxes of low-income blacks can upset middle-class blacks just as they up-

set middle-class whites, leading to fears of increased crime and deteriorating property values. Indeed, middle-class blacks may perceive their middle-class status as more precarious than their white counterparts do. Some Gautreaux participants reported that white neighbors were sometimes more friendly than middle-class black neighbors, who seemed to act "snooty" and "too good to be friendly with us." We suspect that middle-class blacks may worry that these low-income blacks will be stigmatized and that some of the stigma will spill over to them. While we have no systematic information about the magnitude of this concern, we expect that middle-class black suburbs may be concerned about large influxes of low-income blacks.

Unfortunately, MTO programs could inadvertently create such a process. Because MTO legislation seeks to avoid the issue of race, it has no guidelines to prevent large numbers of moves into middle-class black areas. Nor does it have any guidelines to avoid moves into areas that might be near a tipping point. This is a serious shortcoming in the legislation, and we can only hope that the national program will develop program guidelines that prevent such practices and that local programs will be as sensitive to these issues as the Leadership Council has been. It would be a tragedy if this well-intentioned program inadvertently upset the racial or economic balance in a community by panicking residents.

Providing the right amount of help

To help families find housing in areas far from the central city, the Gautreaux program provided extensive help in locating housing. Two people located landlords willing to participate, and four housing counselors took participants to see the housing. Without this help, it was assumed, participants would not be aware of housing opportunities in white suburbs and could not visit these suburbs because of poor public transportation.

While the Gautreaux program provided extensive help in locating housing, it gave little help to families after the move. One staff person served up to 200 new participants for their first six months after moving. This person could not provide extensive help and mainly made referrals to other sources of help in the participants' new communities. Postmove help was limited in order to reduce program costs.

There are numerous nonhousing services that the Gautreaux program does not offer but that could enhance program impacts. For example, transportation was the greatest difficulty facing participants in the suburbs, which had little or no public transportation. If the program could help participants finance the purchase of a car, more people might get jobs, children would have an easier time attending after-school activities, and participants would face fewer frustrations with daily tasks. Child-care assistance would also have been extremely helpful, since suburban movers cannot rely on relatives. Finally, Gautreaux participants might have gotten better jobs if the program had also provided additional education or training.

As a result of the decisions to provide little postmove assistance, the total cost of the Gautreaux program was only about $1,000 per family. This average takes all the Gautreaux program costs and divides them by the number of families served in a year. The number was somewhat more in years when fewer families were moved and somewhat less when more families moved, so there may be some economies of scale.

In setting up the program, there was a concern that if the program provided more help, it would be too costly to be politically feasible in other locations. There was also the opposite concern, that it might provide too little help and thus not succeed. It is note-

worthy that the program had great benefits despite the minimal postmove help it provided. Apparently, most families were able to cope with the difficulties they encountered.

Creating the right expectations

Programs can fail by creating excessive expectations. Expectations are a very real constraint on programs, since unrealistically high expectations will quickly be disappointed, and can lead to the program being abandoned before it has had time to succeed.

This is a great concern for the national MTO program. Founded in part on the *long-term* results of the Gautreaux program, MTO may be expected to show the same benefits *right away,* particularly because policy makers must make extravagant promises to convince Congress and local policy makers to support the program. Moreover, evaluations have been mandated, and their early results will be expected to show immediate benefits.

Thus an oversimple model of change may ignore difficulties that arise in the early years. Such a model might posit that if middle-class areas are beneficial, their benefits will show up immediately in test score improvements. However, the Gautreaux program research (Kaufman and Rosenbaum 1992; Rosenbaum 1993) suggests that low-income suburban movers experience enormous difficulties in their first one to three years after moving. The emotional reactions to these difficulties might lead to test score declines in the first few years, even though increases could occur thereafter. The experience may be similar to moving to a foreign land, where some neighbors are hostile and the school curriculum is years more advanced than in one's former school. Indeed, some children reported that the dialect of white suburban teachers was hard to understand initially. Test score gains may not emerge immediately in such circumstances.

An alternative model may even hypothesize that short-term losses are necessary for long-term gains. This might be termed the "no pain, no gain" model. If children move to areas where the schools are no harder than the city schools they had attended, they may experience fewer difficulties, and their scores will not decline in the first years. Such unchallenging placements might not hurt test scores in the short term, but they might not lead to the long-term gains noted among the suburban Gautreaux children. This potential tradeoff between short term and long term is a speculation, but if true, it indicates that short-term evaluations could give the wrong message.

Other concerns about the Gautreaux program

The Gautreaux program deprives the central city of its leaders

Some have criticized the Gautreaux program for removing the most qualified people from the central city. This would be a serious problem if true. The central cities certainly need good black leaders. Yet the people who are selected for the program are not able to exert strong leadership in their communities; they are mostly single mothers on Aid to Families with Dependent Children (AFDC) struggling to survive and keep their children safe. Being able to go out and get jobs in the suburbs is new for many of them—something they could not do in the housing projects.

The concern about the loss of leaders and talent in the inner city is a serious one. But as Wilson (1987) documented, this loss began two decades ago, when the black middle class began moving to the suburbs. The black middle class could provide leadership, jobs, and positive models to the central city; the Gautreaux mothers cannot do this nearly as well.

The Gautreaux program is too expensive

The program costs $1,000 per family for placement services, plus a Section 8 housing certificate (approximately $6,000 per year) to maintain people in private apartments. Section 8 is already a large national program, so the incremental cost of converting a subset of existing Section 8 certificates to housing mobility is a one-time charge of $1,000 per family. Of course, a major national mobility initiative would require additional Section 8 assistance and would therefore be more expensive.

The other aspect of the cost question is that Gautreaux takes money that could be applied to improving the city. This is wrong on two counts. First, the incremental cost of Gautreaux over the existing Section 8 program is only $1,000 per family—very little money compared with existing urban programs. Second, no one would advocate doing such a program instead of investing in the city. Both should be done.

The Gautreaux program can only be a small program

Many people criticize the Gautreaux program by saying that it can move only a small number of families. This is probably mistaken. While such a program can move only a small number of families into *any single* neighborhood over the course of a few years, it could move large numbers over a decade if it included many scattered neighborhoods in a hundred suburbs surrounding major cities. The key is widely scattered locations.

To estimate the potential magnitude of the program, consider that the Chicago metropolitan area has a population of roughly 7 million, of whom 4 million live in the suburbs. Most suburbs are more than 90 percent white. If a program selected only the "better" half of families in the Chicago Housing Authority buildings—roughly 50,000 families comprising 150,000 individuals—and moved half of them to city apartments and half to suburban apartments, it would have negligible influence on any suburban community. If these 75,000 individuals were evenly spread among 4 million people in the suburbs, that would change the suburban population by less than 2 percent. A 90 percent white suburb would become 88 percent white. This is not the kind of change that panics anyone.

These numbers are larger than anything that is likely to happen. There is no source of funding for a program to move 50,000 families in the Chicago vicinity over the next decade. So these rough estimates are all about a hypothetical program that is beyond the scope of implementation. The point is that suburbs would not be greatly altered by a mobility program much larger than anything that is likely to be implemented, if the program involves a wide variety of areas.

Of course, housing availability is an additional potential constraint. But if the government can create a 10-year housing mobility program, then the housing industry is likely to be able to build new housing to accommodate this new population along with other population growth over the course of a decade. Moreover, as noted, the fact that the program would extend over a decade or more would further reduce its psychological impact.

Serious problems arise only if the program does not involve a wide variety of areas. Indeed, even a program of 300 families a year (smaller than the Gautreaux program) could create panic and rejection if it is narrowly focused on only a few communities.

Although the Gautreaux idea can be expanded to a much larger program, it still seems likely that small programs have stronger benefits than large ones. If three low-income families move to a middle-income neighborhood (of 500 residents), the neighbors and church may provide extensive hospitality. If 30 families move in, neighbors may be friendly, but 30 may be too many to give hospitality. Yet

the children of 30 families will be scattered across many teachers and so still may receive extra individualized help in school. If 300 families move in, conflict and panic may be more likely than hospitality and help.

Numbers also alter the internal dynamics of movers. Children in three new families must interact with neighbors if they are to have friends their age. If children in 30 new families live close together, they can interact with one another and constitute a segregated enclave. The greater ease and comfort of segregated interaction may keep new movers from reaching out to white middle-class neighbors.

This is all speculation. If true, it suggests that numbers do affect the strength of effects of the moves and that as numbers increase, the amount and kinds of help decline.

Conclusions and policy implications

Results of the Gautreaux program show that residential integration can further the aims of improving employment, education, and social integration of low-income blacks. The suburban move greatly improved adults' employment rates, and many adults got jobs for the first time in their lives. The suburban move also improved youth's education and employment prospects. Compared with city movers, the children who moved to the suburbs are more likely to be (1) in school, (2) in college-track classes, (3) in four-year colleges, (4) in jobs, and (5) in jobs with benefits and better pay. The suburban move also led to social integration, friendships, and interaction with white neighbors in the suburbs.

These results provide strong support for the propositions advanced by Galster and Killen (1995), who indicate that segregation discourages educational attainment and employment, and for the effects of moves out of segregated environments. Moreover, unlike the evidence they cite, which relies on multivariate controls, the present findings are based on quasi-random assignment of families to neighborhoods. Since multivariate analyses have difficulty controlling for all the factors that lead to individuals' neighborhood choices, this social experiment provides a useful alternative approach. Galster and Killen also put forward a number of propositions about the mechanisms leading to these outcomes, particularly information, norms, and social networks. We find indications of different social networks in the social interaction findings. While this study lacked systematic indicators of information and norms, these are important issues for future research.

Of course, the social integration was not total, and while harassment declined over time, some prejudice remained. The mothers and youth developed ways of dealing with it, and these unpleasant events were offset by acceptance by many white neighbors and classmates.

Similarly, the children's achievement gains were not immediate. Indeed, virtually all suburban movers experienced great difficulties, and many got lower grades in the first year or two. However, these difficulties were an unavoidable part of adjusting to the higher suburban standards and gaining from the move.

Some critics doubt that housing mobility programs can achieve the integration goals because low-income blacks will not choose to move to middle-income white suburbs. Indeed, a Detroit survey found that few blacks would make all-white neighborhoods their first choice (Farley, Bianchi, and Colasanto 1979; Farley et al. 1993). Moreover, some previous efforts to use tenant-based assistance to encourage racial integration were unsuccessful. The national Experimental Housing Allowance Program "had virtually no impact on the degree of economic and racial concentration experienced by participants" (Cronin

and Rasmussen 1981, 123). Similarly, Project Self-Sufficiency in Cook County, IL, moved very few black participants to white suburbs (Rosenbaum 1988). In both programs, participants were reluctant to make these moves because of strong personal ties to their neighbors, fear of discrimination, and unfamiliarity with the distant suburbs that could have offered them better job prospects.

The results of the Gautreaux program cannot be considered conclusive evidence contradicting these prior studies. Program design features—the lack of real choice about city or suburban locations—that strengthen the research conclusions described above limit any conclusions about what locational choices low-income blacks would make without such constraints. Still, the results suggest that tenant-based assistance can succeed in moving low-income families to suburbs with better schools and better labor markets, and that adults and children will benefit from such moves. The Gautreaux program was able to overcome the reluctance that these families might have felt, in part because the poor quality of life in the city limited the attractiveness of staying there. It is noteworthy that participation is voluntary, and demand for program slots is high.

The Gautreaux program indicates that success is possible but that it requires extensive additional housing services. Real estate staff are needed to locate landlords willing to participate in the program, and placement counselors are needed to inform families about these suburbs, to address their concerns about such moves, and to take them to visit the units and communities. Like participants in other tenant-based assistance programs, Gautreaux participants were reluctant to move to distant suburbs that they had never seen before, and few would have moved without the counselors' encouragement and visits to the suburban apartments. When contrasted with the failures of previous housing voucher programs, the successes of this program indicate the value of having real estate staff and housing counselors.

We have noted a number of pitfalls that such programs must strive to avoid. Programs must select appropriate people, appropriate places, and appropriate services, and they must project appropriate expectations. This does not mean that the program must be very selective and very small, but it must make its operations effective. Nonetheless, the potential benefits make Gautreaux a promising approach, and it is worthwhile to invest more in programs that can lead to these benefits.

Its strategy had the consequence of allowing the Leadership Council to gain the trust of landlords. The Leadership Council's selection procedures became an unofficial warranty of participants' capabilities. Informal reports suggest that even landlords who harbored prejudices against low-income people felt they could trust the people selected by the Leadership Council. In effect, the Leadership Council informally certified participants in ways that overcame landlords' prejudices.

This study also has implications for other housing programs. The results indicate three key factors that helped Gautreaux adults get jobs in the suburbs: personal safety, role models, and access to jobs. If these factors were improved in the city, they might also help city residents. In fact, the Chicago Housing Authority, at the initiative of its director, Vincent Lane, has recently made impressive efforts to improve safety, role models, and job access in public housing projects. To improve the safety of the housing projects, the authority has initiated security measures. To provide positive models, it has initiated a mixed-income housing development, Lake Parc Place, that includes working residents who are positive models for their unemployed neighbors. To improve access to suburban jobs, some

housing projects have provided minibus service to the suburbs. These are the same factors that Gautreaux adults noted as helping them, so they are promising efforts. However, it is not certain how thorough and successful these efforts will be or whether they will result in greater employment. Even improved security may not make the projects as safe as suburbs, and one-hour commutes may limit the attractiveness of taking a minibus to low-paying jobs. It will be some time before we can measure the success of such programs.

The Gautreaux studies support the basic premise of the "geography of opportunity" concept: that where someone lives has an important impact on his or her social and economic prospects. These studies clearly indicate that this housing strategy can lead to great gains in employment, education, and social integration for low-income blacks. Contrary to the pessimistic predictions of "culture of poverty" models, the early experiences of low-income blacks do not prevent them from benefiting from suburban moves. Programs that help people escape areas of concentrated poverty can improve employment and educational opportunities.

Author

James E. Rosenbaum is Professor of Sociology, Education, and Social Policy at Northwestern University.

The Charles Stewart Mott Foundation, the Ford Foundation, the Spencer Foundation, and the Center for Urban Affairs and Policy Research, Northwestern University, provided support for the studies reported here. Nancy Fishman, Julie Kaufman, Marilynn Kulieke, Patricia Meaden, Susan J. Popkin, and Len Rubinowitz made major contributions to these studies. Alexander Polikoff and Kale Williams made important comments to this draft. The ideas expressed here are those of the author and do not necessarily represent those of any other individual or organization.

Notes

1. The program permits moves to black city neighborhoods if they are considered "revitalized." Only Gautreaux families moving to low-income black neighborhoods were included in the city mover sample for the studies reported here.
2. The Section 8 program is a federal program that subsidizes low-income people's rents in private sector apartments, either by giving them Section 8 certificates that allow them to rent apartments on the open market or by moving them into new or rehabilitated buildings whose owners have accepted federal loans that require some units to be set aside for low-income tenants.
3. Our refusal rate on the interviews was less than 7 percent. There are no systematic differences between the interview and survey respondents, but the interview sample is used only for qualitative analysis. Responses to the self-administered questionnaire were consistent with those from the in-person interviews. For a complete description of the sample, instrument, and other analyses, see Rosenbaum and Popkin (1991).
4. Low-income people move often and so are difficult to locate over a seven-year period. We located 59.1 percent, a reasonably large percentage for such a sample. Of course, one must wonder what biases arise from this attrition and whether we were more likely to lose the least successful people (because they were harder to find) or the most successful ones (because they got jobs in distant locations). We suspect both happen, but if one happens more often, then the 1989 sample could be seriously different from the original 1982 sample.
5. The suburban advantage arises because city movers decline in employment. The 15.4 percent decline in employment among the city movers is virtually the same as the 16.3 percent decline found in the Current Population Surveys (CPS) between 1979 and 1989 among poorly educated central-city black adult males, while their non-central-city CPS counterparts have little or no decline (Danziger and Wood 1991, tables 5 and 6). Although selectivity concerns arise in interpreting

differences in the CPS data, the quasi-random assignment makes selectivity less of a threat in our study, which finds the same city-suburban differences as the CPS does. Apparently, the suburban move permitted low-income blacks to escape the declining employment rates in central cities during the 1980s. Moreover, multivariate analyses find that suburban movers are significantly more likely to have a job than city movers, even after controlling for many other factors. These analyses find that other factors also influence employment: previous work experience, years since move, age (inversely), and young children (inversely). Employment is reduced by low internal sense of control and being a long-term Aid to Families with Dependent Children (AFDC) recipient (five years or more), but not by being a second-generation AFDC recipient. Employment is barely influenced by education, and it is not at all affected by postmove general equivalency diploma or college. For details of these analyses, see Rosenbaum and Popkin (1991).

6. Multivariate analyses on postmove hourly wages and on hours worked per week—controlling for the same variables, plus months of employment and the premove measure of the dependent variable (wages or hours, respectively)—confirm the above findings: Living in the suburbs has no effect on either dependent variable. Job tenure, premove pay, and the two "culture of poverty" variables (internal control and long-term AFDC) have significant effects on postmove wages. Job tenure, premove hours worked, and postmove higher education have significant effects on postmove hours worked. None of the other factors had significant effects. For details of these analyses, see Rosenbaum and Popkin (1991).
7. For a complete description of the sample, instrument, and other analyses, see Rosenbaum, Kulieke, and Rubinowitz (1988).
8. For a complete description of the sample, instrument, and other analyses, see Rosenbaum and Kaufman (1991).

References

Berry, Brian J. L. 1979. *The Open Housing Question: Race and Housing in Chicago, 1966–1976.* Cambridge, MA: Ballinger.

Coleman, James S. 1966. *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Cronin, Frank J., and David W. Rasmussen. 1981. Mobility. In *Housing Vouchers for the Poor: Lessons from a National Experiment,* ed. Raymond J. Struyk and Marc Bendick, Jr., 107–28. Washington, DC: The Urban Institute Press.

Danziger, Sheldon, and Robert G. Wood. 1991. Black Male Joblessness in the 1980s. Unpublished paper. Ann Arbor: University of Michigan.

Farley, Reynolds, Suzanne Bianchi, and Diane Colasanto. 1979. Barriers to the Racial Integration of Neighborhoods: The Detroit Case. *Annals of the American Academy of Political and Social Science* 411(January):97–113.

Farley, Reynolds, and William H. Frey. 1992. Changes in the Segregation of Whites from Blacks during the 1980s: Small Steps toward a More Racially Integrated Society. Research Report 92–257. University of Michigan Population Studies Center.

Farley, Reynolds, Charlotte Steeh, Tara Jackson, Maria Krysan, and Keith Reeves. 1993. Continued Racial Residential Segregation in Detroit: "Chocolate City, Vanilla Suburbs" Revisited. *Journal of Housing Research* 4(1):1–38.

Feins, Judith D. 1993. Provide Implementation Assistance and Evaluation for the Moving to Opportunity Demonstration. Research design plan. Washington, DC: Abt Associates.

Fischer, Paul B. 1991. *Is Housing Mobility an Effective Anti-Poverty Strategy?* Cincinnati: Stephen H. Wilder Foundation.

Galster, George C., and Sean P. Killen. 1995. The Geography of Metropolitan Opportunity: A Reconnaissance and Conceptual Framework. *Housing Policy Debate* 6(1):7–43.

Gerard, Harold B., and Norman Miller. 1975. *School Desegregation: A Long-Term Study.* New York: Plenum.

Hawley, Willis D., ed. 1981. *Effective School Desegregation: Equity, Quality, and Feasibility.* Beverly Hills, CA: Sage.

Hobson v. Hansen. 1967. 269 F. Supp. 401 (D.D.C.).

Jencks, Christopher, and Susan E. Mayer. 1989. Residential Segregation, Job Proximity, and Black Job Opportunities: The Empirical Status of the Spatial Mismatch Hypothesis. Center for Urban Affairs Working Paper. Northwestern University.

Kariya, Takehiko, and James E. Rosenbaum. Forthcoming. Institutional Linkages between Education and Work. In *Research in Social Stratification and Mobility.* Greenwich, CT: JAI Press.

Kasarda, John D. 1993. Inner-City Concentrated Poverty and Neighborhood Distress: 1970 to 1990. *Housing Policy Debate* 4(3):253–302.

Kaufman, Julie E., and James E. Rosenbaum. 1992. The Education and Employment of Low-Income

Black Youth in White Suburbs. *Educational Evaluation and Policy Analysis* 14(3):229–40.

Lewis, Oscar. 1968. The Culture of Poverty. In *On Understanding Poverty: Perspectives from the Social Sciences,* ed. Daniel P. Moynihan, 187–200. New York: Basic Books.

Massey, Douglas S., and Nancy A. Denton. 1987. Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970–1980. *American Sociological Review* 52:802–25.

Mead, Lawrence M. 1986. *Beyond Entitlement: The Social Obligations of Citizenship.* New York: Free Press.

Oakes, Jeannie. 1985. *Keeping Track: How Schools Structure Inequality.* New Haven, CT: Yale University Press.

Orfield, Gary. 1984. *Chicago Study of Access and Choice in Higher Education.* Chicago: University of Chicago, Department of Political Science.

Patchen, Martin. 1982. *Black-White Contact in Schools: Its Social and Academic Effects.* West Lafayette, IN: Purdue University Press.

Peroff, Kathleen A., Cloteal L. Davis, and Ronald Jones. 1979. *Gautreaux Housing Demonstration: An Evaluation of Its Impact on Participating Households.* Washington, DC: Division of Policy Studies, Office of Policy Development and Research, U.S. Department of Housing and Urban Development.

Rosenbaum, James E. 1976. *Making Inequality: The Hidden Curriculum of High School Tracking.* New York: Wiley.

Rosenbaum, James E. 1980. Social Implications of Educational Grouping. *Review of Research in Education* 8:361–401.

Rosenbaum, James E. 1988. An Evaluation of Project Self-Sufficiency in Cook County. Unpublished paper. Center for Urban Affairs and Policy Research, Northwestern University.

Rosenbaum, James E. 1993. Closing the Gap. In *Affordable Housing and Public Policy,* ed. Lawrence B. Joseph. Chicago: University of Chicago Press.

Rosenbaum, James E., and Julie E. Kaufman. 1991. Educational and Occupational Achievements of Low-Income Black Youth in White Suburbs. Paper read at the Annual Meeting of the American Sociological Association, August, Cincinnati.

Rosenbaum, James E., Marilynn J. Kulieke, and Leonard S. Rubinowitz. 1988. White Suburban Schools' Responses to Low-Income Black Children: Sources of Successes and Problems. *Urban Review* 20(1):28–41.

Rosenbaum, James E., and Susan J. Popkin. 1991. Employment and Earnings of Low-Income Blacks Who Move to Middle-Class Suburbs. In *The Urban Underclass,* ed. Christopher Jencks and Paul E. Peterson, 342–56. Washington, DC: The Brookings Institution.

Rosenbaum, James E., and Stefan Presser. 1978. Voluntary Racial Integration in a Magnet School. *School Review* 86(2):156–86.

Rosenbaum, James E., Leonard S. Rubinowitz, and Marilyn J. Kulieke. 1986. Low-Income Black Children in White Suburban Schools. Report to the Center for Urban Affairs and Policy Research, Northwestern University.

Schuman, Howard, and Lawrence Bobo. 1988. Survey-Based Experiments on White Racial Attitudes toward Residential Integration. *American Journal of Sociology* 94(2):273–99.

Squires, Gregory D., Larry Bennett, Kathleen McCourt, and Philip Nyden. 1987. *Chicago: Race, Class, and the Responses to the Urban Decline.* Philadelphia: Temple University Press.

St. John, Nancy H. 1975. *School Desegregation: Outcomes for Children.* New York: Wiley.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago: University of Chicago Press.

Yinger, John. 1979. Prejudice and Discrimination in the Urban Housing Market. In *Current Issues in Urban Economics,* ed. Peter Mieszkowski and Mahlon Straszheim, 430–68. Baltimore: Johns Hopkins.

CHAPTER 19

Changes in Alcohol Consumption Resulting from the Elimination of Retail Wine Monopolies

Results from Five U.S. States[*]

Alexander C. Wagenaar, *Ph.D. and* Harold D. Holder, *Ph.D.*[**]

Abstract

Objective: The role of publicly owned and operated retail alcohol monopolies is currently under debate in many Western and former Eastern-bloc countries. We studied the effects of privatizing wine sales in five U.S. states. *Method:* Data on monthly sales of alcoholic beverages were collected for each of the five states, for all states bordering each of the five states and for the U.S. as a whole for the period from 1968 through 1991. Beer and spirits sales data were collected for comparison with wine sales. A quasiexperimental interrupted time-series design was used, including comparison groups consisting of border states, all other U.S. states and beer and spirits sales within the focal state. Box-Jenkins time-series statistical modeling was used to control for intra- and cross-series dependencies and to estimate the net effect of privatization on wine sales.
Results: After controlling for both nationwide and state-specific trends, we found significant increases in wine sales after privatization of 42% in Alabama, 150% in Idaho, 137% in Maine, 75% in Montana and 15% in New Hampshire. The increases in liters of pure ethanol per year in the form of wine were 621,000 in Alabama, 432,000 in Idaho, 364,000 in Maine, 363,000 in Montana and 171,000 in New Hampshire. *Conclusions:* The structure of the retail alcohol distribution system has a significant effect on alcohol sales. We recommend that the social costs associated with increased alcohol use be carefully considered before such major policy changes are contemplated. (*J. Stud. Alcohol* 56: 566–572, 1995)

Received: June 20, 1994. Revision: October 24, 1994.

* Research and preparation of this article were supported in part by National Institute on Alcohol Abuse and Alcoholism grant AA06282 to the Prevention Research Center, Pacific Institute for Research and Evaluation, and by grant AA90142 to the University of Minnesota, School of Public Health, Division of Epidemiology.

† Harold D. Holder is with the Prevention Research Center, Berkeley, Calif.

STATE-OWNED MONOPOLIES for retail sales of alcoholic beverages have been established in a number of countries. These monopolies are often a reflection of a national, state or provincial public policy to restrict alcohol availability with the purpose of reducing alcohol consumption and associated health and social problems. The role of alcohol monopolies in minimizing the health and social costs of drinking is currently under debate in many Western and former Eastern-bloc countries (Ferris et al., 1994).

Effects of retail monopolies on the consumption of beer, wine or distilled spirits have been examined previously. Early investigations used cross-sectional research designs, looking at differences in alcohol retail control systems between states, regions or nations, and relationships between alcohol control systems and alcohol consumption (e.g., Colon, 1981, 1982; Smart, 1977; Swidler, 1986).

Cross-sectional studies are not sufficient for determining the effects of retail monopolies on alcohol consumption. Differences across jurisdictions may be due to deeper differences in countries or states/provinces that are associated with cultural traditions of alcohol use, income levels, economic conditions and other factors that are independent of the retail sales system for alcohol. As a result, longitudinal designs are preferred. With longitudinal designs, researchers take advantage of changes in existing retail monopoly systems, comparing patterns of alcohol sales and consumption before the change with the patterns after the elimination of, or a change in, the alcohol retail monopoly.

Over the past two decades several retail alcohol monopolies have been partially or fully eliminated, providing opportunities to examine effects on drinking. MacDonald (1986) analyzed wine, beer and spirits sales in four U.S. states (Idaho, Maine, Washington and Virginia) that implemented a partial or complete elimination of the retail monopoly for wine over the period 1961 through 1978. Based on ordinary least squares regression methods, MacDonald found statistically significant increases in wine consumption of 190% in Idaho, 305% in Maine and 26% in Washington. No evidence of substitution from distilled spirits or beer was found, since total absolute ethanol consumption also increased, and the increases in consumption were maintained over time. The legal change in Virginia involved the privatization of fortified wine only; table wine had been sold in grocery stores for decades. As a result, no significant increase in overall wine consumption was observed in Virginia.

Smart (1986) examined wine sales and total alcohol sales in Quebec Province, Canada, both before and after retail sales for domestically produced or bottled table wine began in grocery stores in 1978 (eliminating the provincial monopoly for domestic wine). Only domestically produced wines and wines bottled by the provincial monopoly could be sold in private stores. Imported wine continued to be sold by the retail monopoly in Quebec until 1984, when all table wines became available in private stores. Smart compared Quebec wine and total alcohol sales patterns over the period from 1967 to 1983 with similar data from Ontario. Based on results from ordinary least squares regression models, Smart concluded that introduction of domestically produced wine in grocery stores created no short- or medium-term increase in wine sales or total per capita alcohol consumption. Smart observed that table wine was not the most popular alcoholic beverage and that because imported wines, often preferred in French-speaking Quebec, remained limited to provincial monopoly stores, the results were not surprising.

Adrian and associates (1994) analyzed the Quebec experience using alcohol sales data from 1953–1990. Based on multiple regression and time series analyses forecasting

the expected level of wine consumption after the 1978 change, the authors found an increase in wine consumption soon after privatization, but conclude that it was a temporary effect only and not representative of a long-term change in wine consumption.

Mulford and Fitzgerald (1988) investigated the 1985 privatization of table wine in Iowa. Using survey data before and after the policy change, they found a statistically significant increase in self-reported (last 30-day) purchase of wine, a significant decrease in spirits consumption and no change in self-reported beer consumption. Mulford et al. (1992) analyzed Iowa wine sales data using interrupted time-series analyses and concluded that there were no changes in wine sales or overall alcohol consumption associated with the end of the Iowa wine monopoly.

Wagenaar and Holder (1991) conducted interrupted time series analyses of the elimination of state retail wine monopolies in West Virginia and Iowa. They found a 48% increase in wine sales in West Virginia and a 92% increase in wine sales in Iowa associated with the end of the retail wine monopolies. Controlling for effects of cross-beverage substitution, national alcohol sales patterns and border-state alcohol sales, Wagenaar and Holder found a net increase in overall alcohol consumption associated with the policy changes in both West Virginia and Iowa.

Finally, Wagenaar and Langley (in press) studied the 1990 introduction of wine into grocery stores in New Zealand. Based on an interrupted time-series design with nationwide quarterly alcohol sales data from 1983 to 1993, they found a 17% increase in wine sales associated with the new policy. Increased sales were limited to the specific category of alcoholic beverage permitted in grocery stores—table wine. Sales of fortified wine, distilled spirits and beer did not increase.

Most previous studies have found that privatizing wine sales, introducing sales into grocery and other stores in jurisdictions that had previously limited sales to state-owned monopoly outlets, results in an increase in wine sales and consumption. However, this conclusion is not uniform, with some studies finding a temporary effect that dissipates over time and some authors suggesting that increased wine sales might be partially compensated by decreases in sales of beer or spirits. The purpose of the current study is to replicate analyses of the effects of wine privatization in additional jurisdictions, including long-term follow-up of privatization effects, and including analyses of all types of alcoholic beverages for possible compensatory or spill-over effects.

Study states

We identified five U.S. states, in addition to Iowa and West Virginia, that privatized wine sales between 1968 and 1991—Alabama, Idaho, Maine, Montana and New Hampshire. The experience in these states has either not previously been studied at all or has been analyzed using only ordinary least squares linear regression. Because alcohol sales data are serially correlated, the preferred analytic strategy is Box-Jenkins interrupted time series analyses (i.e., using a combination of arima and transfer function models; see McCleary and Hay, 1980). Thus, the current study replicates in five additional states the time series design and analytic strategy Wagenaar and Holder (1991) used to evaluate the effects of wine privatization in Iowa and West Virginia.

Alabama eliminated its monopoly on table wine sales in two phases. In October, 1973, three of the most populated counties (Jefferson, Tuscaloosa and Mobile) were permitted to have private outlets sell table wines for off-premise consumption. Also, sale of table wine for on-premise consumption only was permitted in Montgomery County (con-

taining the state capital). The three counties that permitted private off-premise sales of wine in 1973 represented 31% of the total Alabama population. Privatized table wine sales were permitted in all Alabama counties beginning in October, 1980.

Idaho eliminated the public monopoly of table wines in July, 1971. According to representatives of the Idaho State Liquor Dispensary, this change caused a dramatic increase in promotion and sales of wines in the state (private conversation, September 4, 1993).

Maine eliminated its monopoly on retail sales of table wine as of January 1, 1971. Maine continued to sell fortified wine (14% or more alcohol by volume) in state stores.

Montana ended the state retail monopoly of table wine (14% or less alcohol by volume) in October, 1979, and raised this level to 16% alcohol by volume in 1985.

New Hampshire eliminated the public retail monopoly for table wine in August, 1978. Fortified wines continued to be sold in state retail monopoly stores.

Three states (Washington, Oregon and Virginia) were not included in the current study because they implemented only partial changes in their wine retail monopolies. In 1969, Washington permitted the sale of fortified wine in private stores and permitted the purchase of wine from private wholesale sources. Previously, the Washington monopoly sold all wholesale wine. Similarly, Oregon in May 1974 permitted fortified wines to be sold in private retail outlets. Table wines were already sold privately. Virginia always permitted the sale of all wines in both private retail outlets and the state monopoly. Thus a dual public and private wine retail market existed. In 1986, Virginia eliminated the sale of wines by the state monopoly, except for wines produced in Virginia. This policy change actually lowered the number of retail outlets for wine in Virginia, reportedly associated with a reduction in wine sales over the subsequent 4 years (Department of Alcoholic Beverage Control, Commonwealth of Virginia). Because the policy changes in these three states were quite different from the other changes in privatized wine sales, we did not include them in the current study.

Method

Design

The preferred research design for evaluating effects of elimination of retail monopolies on wine sales is an interrupted time series design (Cook and Campbell, 1979). The design can be shown as:

$$O_1\ O_2 \ldots O_n X O_{n+1}\ O_{n+2} \ldots O_{n+m}$$

$$O_1\ O_2 \ldots O_n\ \ O_{n+1}\ O_{n+2} \ldots O_{n+m}$$

where each O_i represents the sales of spirits, wine or beer per month expressed as volume of absolute ethanol, X represents the policy change ending a state monopoly, n represents the number of observations before the end of the monopoly and m is the number of observations after the end of the monopoly. In the current study, the date of the policy change varies from state to state, which further strengthens the research design by eliminating contemporaneous history as a threat to the internal validity of the design. For example, Mulford et al. (1992) argued that observed increases in wine sales in Iowa were due to an increase in the popularity of wine coolers, not the privatization of wine sales that was implemented at approximately the same time. In the current study the five focal states implemented their changes at widely different times over the past quarter century. As a result, particular historical events such as the rapid popularization of wine coolers in the mid-1980s are eliminated as a threat to the internal validity of the research design.

The second line in the design diagram reflects comparison time series not influenced by the specific policy change being evaluated. In this study the first comparison group con-

sists of the other 47 contiguous states that did not implement any policy changes regarding wine sales at the time the focal state did. As an additional comparison, we examined wine sales in all the states bordering the focal state that eliminated its wine monopoly to identify whether observed sales effects represent shifts in sales from adjacent states or additional wine consumption.

Data

Outcome measures used are quantities of absolute ethanol sold in the form of wine, beer or spirits measured as shipments from wholesalers to retail sellers. Wholesale shipments are used to provide a consistent measure of alcohol sales. While retail sales data for state monopoly stores are available before the end of wine monopolies, such data are not available from individual retail outlets after privatization. Data on wholesale shipments, however, are consistently available throughout the 24-year period studied. Because data are based on wholesale shipments, we controlled statistically for stocking effects—sudden (but temporary) bursts of wine shipments immediately at the time of privatization to stock new retail channels.

Source of spirits data is monthly reports of total volume of spirits sold to licensed retail establishments (state stores before the end of state monopoly and private licenses after) obtained from the Distilled Spirits Council of the United States. Source of wine data is the Wine Institute, based on reports from the Bureau of Alcohol, Tobacco, and Firearms, U.S. Department of the Treasury. Beer data were obtained from annual volumes of *Brewers' Almanac,* published by the Beer Institute. Monthly shipments of alcoholic beverage volume into retail stores were converted to absolute alcohol using the following ethanol proportions: beer, 0.045 (all years); wine, 0.16 (1968–71), 0.145 (1972–76) and 0.129 (1977–89); spirits, 0.45 (1968–71), 0.43 (1972–76) and 0.411 (1977–89), based on Laforge et al. (1987). Wine ethanol figures were adjusted beginning in 1984 to take into account the growing market share of wine coolers, which have a lower alcohol concentration than other wine. The volume of juice in wine coolers was removed from the wine figures before calculating ethanol volumes (Beverage Industry Annual Manual, Cleveland, Ohio, 1989).

Statistical analyses

Box-Jenkins intervention-analysis models were developed for each outcome measure, beginning with identification or specification of a parsimonious ARIMA model. The ARIMA model isolates all aspects of the stochastic autocorrelation structure of the series and provides a benchmark for assessment of intervention effects. The ARIMA model accounts for variability in the dependent series that is due to identifiable trend, seasonal and other autocorrelation patterns in the data. The residual white-noise, or random-error variance then permits a sensitive test of the statistical significance of intervention effects. Because alcohol sales vary substantially by season of the year, the general multiplicative seasonal model was considered for each series. The general seasonal ARIMA model for outcome variable y_t is:

$$Y_t = \frac{(1-\Theta_1 B^s-\ldots\Theta_Q B^{sQ})(1-\theta_1 B-\ldots\theta_q B^q)u_t + \alpha}{(1-\Phi_1 B^s-\ldots\Phi_P B^{sP})(1-\phi_1 B-\ldots\phi_P B^P)(1)(1-B)^D}$$

where p is the order of the auto-regressive process, d is the degree of nonseasonal differencing, q is the order of the moving-average process, P is the order of the seasonal autoregressive process, D is the degree of seasonal differencing, Q is the order of the seasonal moving-average process, s is the seasonal span, Φ_1 to Φ_P are seasonal autoregressive parameters, ϕ_1 to ϕ_p are regular autoregressive parameters, Θ_1 to Θ_Q are seasonal moving-average parameters, θ_1 to θ_q are regular

moving-average parameters, u_t is a random (white-noise) error component, α is a constant and B is the backshift operator such that $B(Z_t)$ equals Z_{t-1}. To reduce heteroscedasticity, we transformed each series taking the natural logarithms.

Theoretical autocorrelation and partial-autocorrelation functions corresponding to various ARIMA models have been described by Box and Jenkins (1976) (also see Harvey, 1990; Wei, 1990). In the present study, a preliminary ARIMA $(p,d,q)(P,D,Q)_s$ model was identified for each series based on an examination of the estimated autocorrelations and partial autocorrelations, assessing the degree to which the actual autocorrelations fit one of the theoretically expected patterns. The simplest model that could plausibly account for the behavior of the series was selected.

Transfer functions representing hypothesized effects of the intervention were then added to the ARIMA model. The general form of the transfer function is:

$$Y_t = \frac{(\omega_o - \omega_1 B - \ldots \omega_s B^s)}{(1 - \delta_1 B - \ldots \delta_r B^r)} (X_{t-b}) + Z_t \beta_i \qquad (2)$$

where ω_o to ω_s and δ_1 to δ_r specify the manner in which the input, or policy intervention variable, X_t, influences the output, or dependent variable, y_t; B is the backshift operator such that $B(Z_t)$ equals z_{t-1}; X_t is either a step function with the value zero before the intervention and one thereafter, or a pulse function with the value one for the month in which the intervention begins and zero otherwise; b is a delay parameter indicating the length or lag, or dead time, between the intervention and the initial effects of the intervention; and β_i is a vector of covariates. Many specific forms of the general transfer function are possible, depending on whether the hypothesized effect pattern is immediate or delayed, sudden or gradual, temporary or permanent. In the present case, the long-term effects of privatization were best modeled using the formula $\omega_o X_t$, while temporary store-stocking effects were best modeled using the formula $(\omega_o/1-\delta)X_t$. All models also incorporated measures of nationwide alcohol sales as covariates in order to control for national trends when assessing intervention effects in specific states; wine sales trends clearly changed over the quarter-century examined (Figure 1). The national covariate for each state consisted of wine sales in all other states other than the focal state. All models were estimated using the Gauss-Marquardt backcasting algorithm implemented in BMDP2T.

For those unfamiliar with these analytical methods, a few points are worth noting. The statistical models control for several factors before estimating the change in alcohol sales attributable to the privatization policy changes. First, long-term trends and regular cycles *within* each state are taken into account. For example, if a policy change was implemented in the mid-1970s, a period during which most states were experiencing an upward trend in sales, one could not attribute the increased sales to the privatization policy, unless the increase was clearly larger than the expected upward trend. We estimate the effects of the privatization policy only after taking into account such trends. Second, trends in alcohol sales *across states* are controlled. For example, if a given state shows a sudden increase in sales after privatization, but a similar increase occurred in other states that did not implement a privatization policy, one could not attribute the increase to the new policy. Only the degree to which a change in sales in the given state is above and beyond any changes observed in the comparison states represents an effect of the policy change. Third, sudden increases in wine sales immediately at the time of privatization represent the effect of stocking the new private outlets. These temporary stocking effects were controlled before estimating the long-term effects of the policy. For all these reasons, simple before and after comparisons of alcohol sales are not an accurate or valid way to assess policy ef-

fects. In summary, the estimated effects of privatization policies shown in Table 1 are net effects attributable to that policy.

Results

All five states experienced statistically significant increases in wine sales after privatization, after controlling for nationwide wine sales trends, state-specific trends and short-term stocking effects with Box-Jenkins time series models (Table 1 and Figure 1). In four of the five states, the increases in wine sales were dramatic. Wine sales in Alabama increased 42% when privatization was extended from three counties to all counties in 1980. The 1973 privatization of sales in three populous Alabama counties was associated with a 16%

TABLE 1

Effects of Privatization Policies on Wine Sales: Results from Box-Jenkins Time Series Models

State	Model	Adjusted R^2	Estimate	Percent change	95% Confidence Interval	
					Low	High
Alabama: 1973 change						
Wine	ARIMA $(0,1,8)\ (0,1,1)_{12}$	0.88	0.144	15.5	-6.2	42.4
Border states wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.96	-0.024	-2.4	-9.9	5.8
Beer	ARIMA $(0,1,3)\ (0,1,1)_{12}$	0.90	0.003	0.4	-7.6	9.0
Spirits	ARIMA $(0,1,9)\ (0,1,1)_{12}$	0.94	-0.017	-1.7	-6.2	3.1
Alabama: 1980 change						
Wine	ARIMA $(0,1,8)\ (0,1,1)_{12}$	0.88	0.350	42.0*	13.4	77.7
Border states wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.96	0.020	2.0	-5.9	10.5
Beer	ARIMA $(0,1,3)\ (0,1,1)_{12}$	0.90	-0.076	-7.4	-14.7	0.6
Spirits	ARIMA $(0,1,9)\ (0,1,1)_{12}$	0.94	-0.051	-5.0	-9.7	0.1
Idaho						
Wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.95	0.917	150.1*	129.2	172.9
Border states wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.94	-0.034	-3.3	-10.7	4.7
Beer	ARIMA $(0,1,4)\ (0,1,1)_{12}$	0.93	0.090	9.5	-7.0	28.8
Spirits	ARIMA $(0,0,4)\ (0,1,1)_{12}$	0.84	0.065	6.8	-0.3	14.4
Maine						
Wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.91	0.862	136.7*	112.6	163.5
Border states wine	ARIMA $(0,1,3)\ (0,1,1)_{12}$	0.94	0.102	10.8	-2.6	26.0
Beer	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.85	0.031	3.2	-5.5	12.7
Spirits	ARIMA $(0,0,6)\ (0,1,1)_{12}$	0.87	-0.019	-1.9	-7.3	4.0
Montana						
Wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.97	0.561	75.3*	-15.7	96.0
Border states wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.91	-0.014	-1.4	-15.7	15.2
Beer	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.87	-0.050	-4.9	-13.0	4.0
Spirits	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.64	-0.044	-4.4	-14.0	6.4
New Hampshire						
Wine	ARIMA $(0,1,10)\ (0,1,1)_{12}$	0.95	0.122	13.0*	1.2	26.2
Border states wine	ARIMA $(0,1,1)\ (0,1,1)_{12}$	0.92	0.006	0.6	-6.6	8.3
Beer	ARIMA $(0,1,2)\ (0,1,1)_{12}$	0.91	-0.023	-2.2	-10.5	6.8
Spirits	ARIMA $(0,1,2)\ (0,1,1)_{12}$	0.92	-0.012	-1.2	-7.6	5.7

*Statistically significant at $p < .05$.

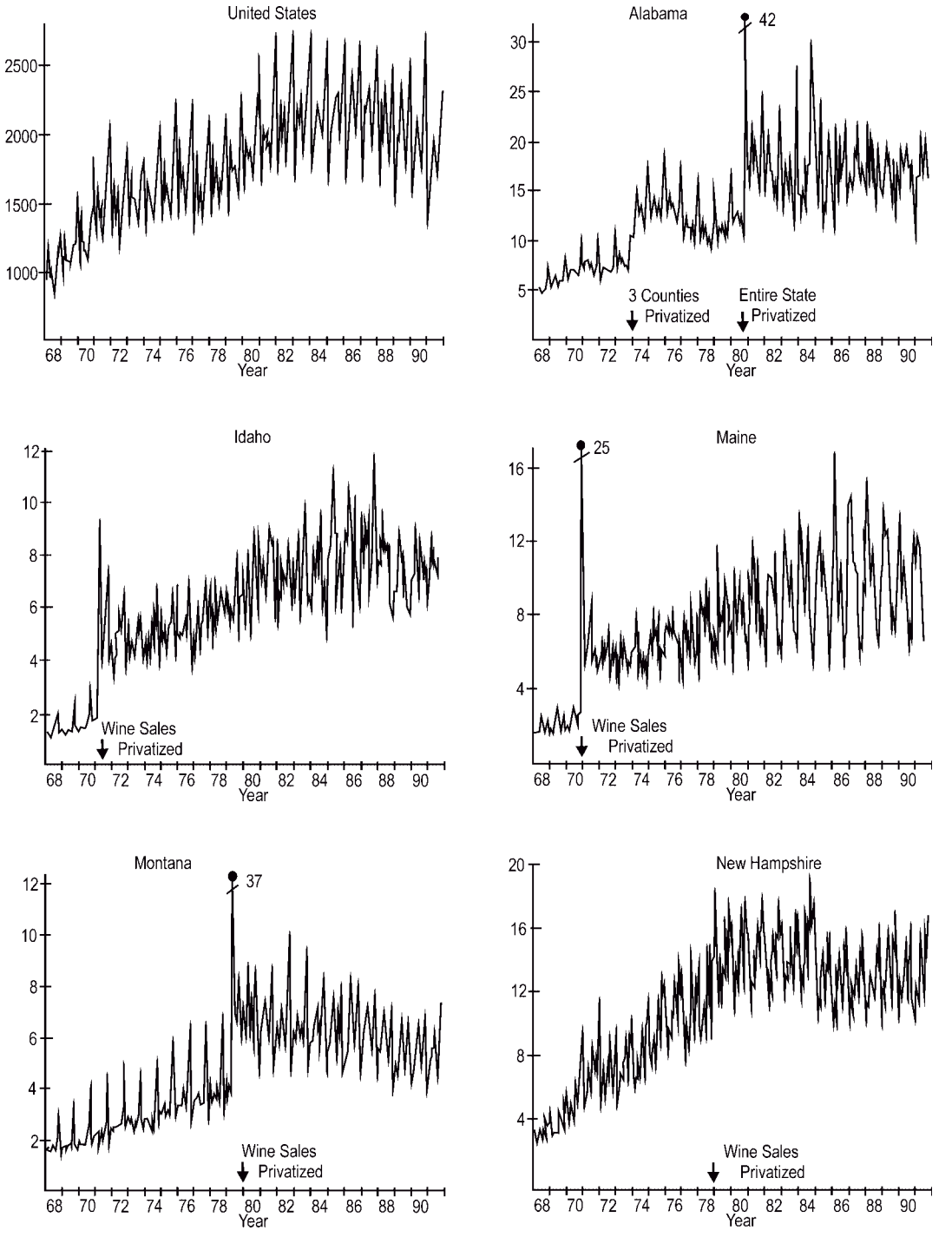


FIGURE 1
Wine sales in liters of pure ethanol per month (x 10,000)

increase in statewide wine sales. In Idaho, wine sales increased 150% after the state monopoly was replaced with licensed private outlets. Maine's 1971 privatization policy was followed by a 137% increase in wine sales. Montana experienced a 75% increase in wine sales following its 1979 introduction of wine into private stores. The smallest effect was seen in New Hampshire, where the 1978 privatization policy resulted in a 13% wine sales increase.

Despite the fact that wine represents a relatively small portion of the overall market for alcoholic beverages, the sales increases after privatization are substantial. Alabama experienced an increase in consumption of 621,000 liters of pure ethanol per year after privatization. Comparable figures for the other states are 364,000 in Maine, 363,000 in Montana, 432,000 in Idaho and 171,000 in New Hampshire.

Four of the five states exhibited additional large increases in wine sales immediately after the privatization policy took effect, above the long-term increases noted in Table 1. In Idaho, Maine and Montana the short-term increases exceeded 400%, and in Alabama the increase exceeded 100%. These large short-term increases in wine sales measured at the wholesale level represent temporary stocking effects as new wine outlets purchased inventory. Such short-term distribution system effects were controlled before estimating the long-term effect of privatization on wine consumption.

Analyses of states bordering these five focal states revealed no evidence that the wine sales increases observed after privatization were due to a shift of sales from adjacent states to the newly privatized states (Table 1). In each of the five cases, there was no significant decline in wine sales in states adjacent to the focal state. In short, increased wine sales after privatization appears to reflect substantially increased wine consumption in states changing their wine distribution policies.

We also analyzed beer and spirits sales separately in each of the five states, using the same type of time series models that we used for wine sales (Table 1). In none of the five states were there statistically significant changes in beer or spirits sales at the time of the significant increases in wine sales.[1] Thus, the increased wine sales cannot be attributed to other state-specific changes that could have influenced the demand for alcoholic beverages across beverages categories. There is similarly little evidence of substitution or spillover effects; that is, the increased wine sales were not associated with either reductions (substitution) or increases (spillover) in sales of beer and spirits.

Conclusions

Results from five additional U.S. states confirm our previous finding from Iowa and West Virginia (Wagenaar and Holder, 1991). Elimination of public monopolies on retail sale of wine is followed by significantly increased wine sales. The magnitude of the effects observed in the current study are substantial, ranging from 13% in New Hampshire to 150% in Idaho. Results also suggest that the sales increases reflect increased consumption of wine, because there is no evidence of shifts of sales from border states or shifts from beer or spirits to wine within each state.

The pattern of significantly increased wine sales following privatization has now been found in 9 jurisdictions (Table 2). The replications of this "natural experiment" (Cook and Campbell, 1979) with wine privatization occurred at different points in time over the last two decades, further strengthening the conclusion that observed effects were not due to particular historical events that occurred at the time of privatization. The only such historical confound that has been suggested in the literature is the expansion of the market for wine coolers in the mid-1980s (Mulford et al., 1992; see also Wine Trends, 1989). Replications in the 1970s, however, before wine coolers were a significant

TABLE 2
Summary of Results from Studies of the Effects of Elimination of Public Retail Table Wine Monopolies*

Jurisdiction	*Date of Policy Change*	*Effect on Wine Sales (%)*	*Study*
Alabama	October 1980	+42	Present article
Idaho	July 1971	+190	MacDonald, 1986
Iowa	July 1985	Temporary	Mulford and Fitzgerald, 1988 Mulfiord et al., 1992
		+92	Wagenaar and Holder, 1991
Maine	January 1971	+305	MacDonald, 1986
Montana	July 1979	+75	Present article
New Hampshire	August 1978	+13	Present article
New Zealand	April 1990	+17	Wagenaar and Langley, in press
Quebec	June 1978	None Temporary	Smart, 1986 Adrian, 1994
Washington	1969	+26	MacDonald, 1986
West Virginia	July 1981	+ 48	Present article

*Cross-sectional studies noted in text that did not analyze before/after data are not included here.

market factor, found magnitudes of effect that were similar to those found in the 1980s.

The now considerable number of replications of analyses of effects of wine privatization policies confirm what has been found in a much broader literature on the effects of alcohol availability on drinking: when alcoholic beverages are made more accessible, consumption typically increases. This pattern has been found whether the increase in availability is a result of lowering the legal age for drinking (Wagenaar, 1993), a reduction in taxes and thus the retail price of alcohol (Toomey et al., 1993) or an increase in the density of alcohol outlets (Gruenewald et al., 1993). Policymakers concerned about the many social and health problems exacerbated by high drinking levels in society should approach with caution changes in the alcohol distribution system that increase the availability of alcohol.

Note

1. One reviewer suggested that the negative point estimates for beer and spirits in several states indicate that wine privatization may result in increased wine sales at the expense of other alcoholic beverages, with no effect on overall alcohol consumption. However, only in Alabama at the time of 1980 expansion of wine privatization do the declines in beer and spirits sales approach statistical significance, suggesting a compensatory decline in beer and spirits sales (Table 1). In contrast, the point estimates for Idaho suggest the opposite hypothesis—a spillover effect—with beer and spirits sales *increasing* at the time of the wine sales increase, although again the effects are not significant. Because none of these estimates are statistically significant, we conclude the data do not support either the substitution or the spillover hypothesis.

References

Adrian, M., Perguson, B.S. and Her, M.H. *Wine in Quebec,* Toronto: Addiction Research Foundation, 1994.

Beverage Industry Annual Manual, Cleveland, Ohio, 1989.

Box, G.E.P. and Jenkins, G.M. *Time Series Analysis: Forecasting and Control (Revised Edition)*, Oakland, Calif.: Holden-Day, Inc., 1976.

Colon, I. Alcohol availability and cirrhosis mortality rates by gender and race. *Amer. J. Publ. Hlth* **71:** 1325–1328, 1981.

Colon, I. The influence of state monopoly of alcohol distribution and the frequency of package stores on single motor vehicle fatalities. *Amer. J. Drug Alcohol Abuse* **9:** 325–331, 1982.

Cook, T.D. and Campbell, D.T. *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Co., 1979.

Ferris, F., Dunning, R., Giesbrecht, N. and West, P. (Eds.) *International Symposium on Alcohol Monopolies and Social and Health Issues*, Toronto: Addiction Research Foundation, 1994.

Gruenewald, P.J., Ponicki, W.R. and Holder, H.D. The relationship of outlet densities to alcohol consumption: A time series cross-sectional analysis. *Alcsm Clin. Exp. Res.* **17:** 38–47, 1993.

Harvey, A. *The Economic Analysis of Time Series,* 2d Edition, Cambridge, Mass.: MIT Press, 1990.

Laforge, R., Stinson, F., Fveel, C. and Williams, G. *Apparent Per Capita Alcohol Consumption: National, State and Regional Trends, 1971–1985. Surveillance Report #7*, Bethesda, Md.: National Institute on Alcohol Abuse and Alcoholism, 1987.

McCleary, R. and Hay, R.A., Jr. *Applied Time Series Analysis for the Social Sciences*, Newbury Park, Calif.: Sage Pubns., Inc., 1980.

MacDonald, S. The impact of increased availability of wine in grocery stores on consumption: Four case histories. *Brit. J. Addict.* **81:** 381–387, 1986.

Mulford, H.A. and Fitzgerald, J.L. Consequences of increasing off-premise wine outlets in Iowa. *Brit. J. Addict.* **83:** 1271–1279, 1988.

Mulford, H.A., Ledolter, J. and Fitzgerald, J.L. Alcohol availability and consumption: Iowa sales data revisited. *J. Stud. Alcohol* **53:** 487–494, 1992.

Smart, R.G. The relationship of availability of alcoholic beverages to per capita consumption and alcoholism rates. *J. Stud. Alcohol* **38:** 891–896, 1977.

Smart, R.G. The impact on consumption of selling wine in grocery stores. *Alcohol Alcsm* **21:** 233–236, 1986.

Swidler, S. A reexamination of liquor price and consumption differences between public and private ownership states: Comment. *Southern Econ. J.* July: 259–264, 1986.

Toomey, T.L., Jones-Webb, R. and Wagenaar, A.C. Recent research on alcohol beverage control policy: A review of the literature. *Ann. Rev. Addict. Res. Treat.* **3:** 279–292, 1993.

Wagenaar, A.C. Minimum drinking age and alcohol availability to youth: Issues and research needs. In: Hilton, M.E. and Bloss, G. (Eds.) *Economics and the Prevention of Alcohol-Related Problems.* NIAAA Research Monograph No. 25, NIH Publication No. 93–3513, Washington: Government Printing Office, 1993, pp. 175–200.

Wagenaar, A.C. and Holder, H.D. A change from public to private sale of wine: Results from natural experiments in Iowa and West Virginia. *J. Stud. Alcohol* **52:** 162–173, 1991.

Wagenaar, A.C. and Langley, J.D. Alcohol licensing system changes and alcohol consumption: Introduction of wine into New Zealand grocery stores, *Addiction,* in press.

Wei, W.W.S. *Time Series Analysis: Univariate and Multivariate Methods*, Reading, Mass.: Addison-Wesley Publishing Co., Inc., 1990.

Wine trends: Sales of imports; Coolers Fade. Beverage Industry Annual Manual, Cleveland, Ohio, 1989, pp. 58–62.

PART VIII

Dilemmas of Evaluation

We close the book with two articles and a critique. They are two of the most interesting articles to come down the pike in a long time, and they illustrate one of the real dilemmas of evaluation. As evaluators, we must always be sure that we are asking the right question, and this is the dilemma presented here. Two different research groups did two separate evaluations of the school choice program in Milwaukee, Wisconsin, *using the same data set* and came to different conclusions. One evaluation concluded that the program was ineffective—that it did not lead to improved academic achievement for participants. The other evaluation looked for the same outcomes—improved academic achievement—but in a slightly different way. They came to the conclusion that the program did indeed produce improved academic achievement for participants.

Read these evaluations and see what you think. Then turn to the critique. We have asked one of our colleagues, research methodologist Chieh-Chen Bowen, what she thinks.

CHAPTER 20

Fifth-Year Report

Milwaukee Parental Choice Program

John F. Witte
Troy D. Sterr
Christopher A. Thorn

Department of Political Science and
The Robert La Follette Institute of Public Affairs
University of Wisconsin-Madison

Executive Summary

This report consists of four sections: (1) a description of the Milwaukee Parental Choice Program and the data being collected; (2) a description of the choice families and students; (3) a five-year report on outcomes; and (4) a brief response to some of the criticisms of our previous evaluations. For the most part, this report updates the *Fourth Year Report* (December 1994) and should be read in conjunction with that report. Most of the findings are consistent with that very detailed report.

The Original Program. The Milwaukee Parental Choice Program, enacted in spring 1990, provides an opportunity for students meeting specific criteria to attend private, nonsectarian schools in Milwaukee. A payment from public funds equivalent to the MPS per member state aid ($3,209 in 1994–95) is paid to the private schools in lieu of tuition and fees for the student. Students must come from families with incomes not exceeding 1.75 times the national poverty line. New choice students must not have been in private schools in the prior year or in public schools in districts other than MPS. The total number of choice students in any year was limited to 1% of the MPS membership in the first four years, but was increased to 1.5% beginning with the 1994–95 school year.

Schools initially had to limit choice students to 49% of their total enrollment. The legislature increased that to 65% beginning in 1994–95. Schools must also admit choice students without discrimination (as specified in s. 118.13, Wisconsin Stats.). Both the statute and administrative rules specify that pupils must be accepted "on a random basis." This has been

interpreted to mean that if a school was oversubscribed in a grade, random selection is required in that grade. In addition, in situations in which one child from a family was admitted to the program, a sibling is exempt from random selection even if random selection is required in the child's grade.

The New Program. The legislation was amended as part of the biennial state budget in June 1995. The principal changes were: (1) to allow religious schools to enter the program: (2) to allow students in grades kindergarten through three who were already attending private schools to be eligible for the program: (3) to eliminate all funding for data collection and evaluations (the Audit Bureau is required simply to file a report by the year 2000). Thus unless the legislation changes, data of the type collected for this and previous reports will not be available for the report to be submitted in the year 2000.

Choice Families and Students. Enrollment in the Choice Program has increased steadily but slowly, never reaching the maximum number allowed by the law. September enrollments have been 341, 521, 620, 742, and 830 from 1990–91 to 1994–95. The number of schools participating was: 7 in 1990–91, 6 in 1991–92, 11 in 1992–93, and 12 in the last two years. The leading reasons given for participating in the Choice Program [were] the perceived *educational quality* and the *teaching approach and style* in the private schools. That is followed by the *disciplinary environment and the general atmosphere* that parents associate with those schools. Frustration with prior public schools, although not unimportant, was not as important a reason for applying to the Choice Program as the attributes of the private schools.

The Choice Program was established, and the statute written, explicitly to provide an opportunity for relatively poor families to attend private schools. The program has clearly accomplished this goal. In terms of reported family income (Table 5a), the average income was $11,630 in the first five years. Incomes of 1994 choice families increased considerably to an average of $14,210. In racial terms, the program has had the greatest impact on African-American students, who comprised 74% of those applying to choice schools and 72% of those enrolled in the first five years of the program (Table 5b). Hispanics accounted for 19% of the choice applicants and 21% of those enrolled. In terms of marital status (Table 5c), choice families were much more likely to be headed by a non-married parent (75%) than the average MPS family (49%), and somewhat more likely than the low-income MPS parent (65%). The percentage was almost identical for the five separate years. A unique characteristic of choice parents was that despite their economic status, they reported higher education levels than either low-income or average MPS parents (Table 5e). Over half of the choice mothers reported some college education (56%), compared with 40% for the entire MPS sample and 30% of the low-income MPS respondents. Consistent with the education levels of parents, educational expectations for their children were also somewhat higher for choice parents than MPS and low-income MPS parents.

Finally, the experiences of choice families with prior MPS schools differed from non-choice, MPS families in three important ways: (1) students enrolling in the choice program were not achieving as well as the average MPS student; (2) choice parents were considerably less satisfied than other MPS parents; and (3) parental involvement in their prior schools and in educational activities at home was greater for choice parents.

Outcomes. Outcomes after five years of the Choice Program remain mixed. Achievement change scores have varied considerably in the first five years of the program. Choice stu-

dents' reading scores increased the first year, fell substantially in the second year, and have remained approximately the same in the next three years. Because the sample size was very small in the first year, the gain in reading was not statistically significant, but the decline in the second year was. In math, choice students were essentially the same in the first two years, recorded a significant increase in the third year, and then significantly declined this last year.

MPS students as a whole gained in reading in the first two years, with a relatively small gain in the first year being statistically significant. There were small and not significant declines in the last two years. Low-income MPS students followed approximately the same pattern, with none of the changes approaching significance. Math scores for MPS students were extremely varied. In the first year there were significant gains for both the total MPS group and the low-income subgroup. In the second year, the scores were essentially flat, but in the third year they declined significantly. Again, in the fourth year there was essentially no change in either the total MPS or low-income MPS groups.

Regression results, using a wide range of modeling approaches, including yearly models and a combined four-year model, generally indicated that choice and public school students were not much different. If there was a difference, MPS students did somewhat better in reading.

Parental attitudes toward choice schools, opinions of the Choice Program, and parental involvement were very positive over the first five years. Parents' attitudes toward choice schools and the education of their children were much more positive than their evaluations of their prior public schools. This shift occurred in every category (teachers, principals, instruction, discipline, etc.) for each of the five years. Similarly, parental involvement, which was more frequent than for the average MPS parent in prior schools, was even greater for most activities in the choice schools. In all years, parents expressed approval of the program and overwhelmingly believed the program should continue.

Attrition (not counting students in alternative choice schools) has been 44%, 32%, 28%, 23%, and 24% in the five years of the program. Estimates of attrition in MPS are uncertain, but in the last two years, attrition from the Choice Program was comparable to the range of mobility between schools in MPS. The reasons given for leaving included complaints about the Choice Program, especially application and fee problems, transportation difficulties and the limitation on religious instruction. They also included complaints about staff, general educational quality and the lack of specialized programs in the private schools.

Conclusions. We ended the *Fourth Year Report* by summarizing the positive and negative consequences of the program. We then wrote:

> Honorable people can disagree on the importance of each of these factors. One way to think about it is whether the majority of the students and families involved are better off because of this program. The answer of the parents involved, at least those who respond to our surveys, was clearly yes. This is despite the fact that achievement, as measured by standardized tests, was no different than the achievement of MPS students. Obviously the attrition rate and the factors affecting attrition indicate that not all students will succeed in these schools, but the majority remain and applaud the program.

Although achievement test results may be somewhat more bleak for choice in this analysis, the differences are not very large in terms of their impact, and the negative esti-

mates are based on less stable models and smaller sample sizes. In addition test scores are only one indication of educational achievement. Thus we see no reason to change last year's conclusion.

Acknowledgments

This research was funded initially by a grant from the Robert La Follette Institute of Public Affairs of the University of Wisconsin-Madison and subsequently by a substantial grant from the Spencer Foundation in Chicago. We gratefully acknowledge this support.

Availability of Data, Codebooks, Reports and Papers on the Internet

A World Wide Web site has been created for choice data and some reports and papers. Included are all data modules used for all reports, with electronic codebooks; the 1994 report; this report; and a paper by John Witte and Chris Thorn comparing the characteristics of public and private school families in the State of Wisconsin and in the City of Milwaukee. Subsequent analyses and publications will be added to the site.

The Internet address of the site is:
http://dpls.dacc.wisc.edu/choice/choice_index.html

I. Introduction

This report consists of four sections: (1) a description of the Milwaukee Parental Choice Program and the data being collected; (2) a description of the choice families and students; (3) a five-year report on outcomes; and (4) a brief response to some of the criticisms of our previous evaluations.

This report is an abbreviated version of earlier reports, for three reasons. First, adequate resources to do this research have not been available for the past two years. Second, program legislation has been changed to include parochial schools and allow students already in those schools to enter the program. The new program will not be comparable and will not be evaluated except for an audit in the year 2000. Third, comparable Milwaukee Public School (MPS) system achievement test score data were not available for the fifth year because the MPS has essentially dropped the Iowa Test of Basic Skills (ITBS) as state mandated tests have been introduced. The private schools continue to use the ITBS and are not required to take the state tests. This report does, however, more than fulfill the statutory requirements (section 119.23 (5) (d)). This report does not contain any information on the 1995–96 school year.

For the most part, this report updates the *Fourth Year Report* (December 1994) and should be read in conjunction with that report. Most of the findings are consistent with that very detailed report. This report includes three new items: (1) a brief description of the new program legislation; (2) new multivariate estimates of achievement differences between choice and MPS schools, combining four years of data; and (3) a brief response to some of the criticisms of previous reports.[1] Some of these critiques were inaccurate in the extreme and should not have been used as policy-making tools. The Peterson critique, for example, circulated when the legislature was considering major changes in the legislation, and dramatic changes did occur—particularly an increase in the scope of the program and appropriations for students to attend religious schools.

II. The Milwaukee Parental Choice Program

The Original Program. The Milwaukee Parental Choice Program, enacted in spring 1990, provides an opportunity for students

meeting specific criteria to attend private, nonsectarian schools in Milwaukee. A payment from public funds equivalent to the MPS per-member state aid ($3,209 in 1994–95) is paid to the private schools in lieu of tuition and fees for the student. Students must come from families with incomes not exceeding 1.75 times the national poverty line. New choice students must not have been in private schools in the prior year or in public schools in districts other than MPS. The total number of choice students in any year was limited to 1% of the MPS membership in the first four years, but was increased to 1.5% beginning with the 1994–95 school year.

Schools initially had to limit choice students to 49% of their total enrollment. The legislature increased that to 65% beginning in 1994–95. Schools must also admit choice students without discrimination (as specified in s. 118.13, Wisconsin Stats.). Both the statute and administrative rules specify that pupils must be accepted "on a random basis." This has been interpreted to mean that if a school was oversubscribed in a grade, random selection is required in that grade. In addition, in situations in which one child from a family was admitted to the program, a sibling is exempt from random selection even if random selection is required in the child's grade.

The New Program. The legislation was amended as part of the biennial state budget in June 1995. The principal changes were: (1) to allow religious schools to enter the program; (2) to allow students in grades kindergarten through grade three who were already attending private schools to be eligible for the program;[2] (3) to eliminate all funding for data collection and evaluations (the Audit Bureau is required simply to file a report by the year 2000). Thus unless the legislation changes, data of the type collected for this and previous reports will not be available for the report to be submitted in the year 2000.

Research and Data. The study on which this report is based employs a number of methodological approaches. Surveys were mailed in the fall of each year from 1990 to 1994 to all parents who applied for enrollment in one of the choice schools. Similar surveys were sent in May and June of 1991 to a random sample of 5,475 parents of students in the Milwaukee Public Schools. Among other purposes, the surveys were intended to assess parent knowledge of and evaluation of the Choice Program, educational experiences in prior public schools, the extent of parental involvement in prior MPS schools, and the importance of education and the expectations parents hold for their children. We also obtained demographic information on family members. A follow-up survey of choice parents assessing attitudes relating to their year in private schools was mailed in June of each year.

In addition, detailed case studies were completed in April 1991 in the four private schools that enrolled the majority of the choice students. An additional study was completed in 1992, and six more case studies in the spring of 1993. Case studies of the K-8 schools involved approximately 30 person-days in the schools, including 56 hours of classroom observation and interviews with nearly all of the teachers and administrators in the schools. Smaller schools required less time. Researchers also attended and observed parent and community group meetings, and Board of Director meetings for several schools. Results of these case studies were included in the December 1994 report.

Finally, beginning in the fall of 1992 and continuing through the fall of 1994, brief mail and phone surveys were completed with as many parents as we could find who chose

not to have their children continue in the program.

In accordance with normal research protocol, and with agreement of the private schools, to maintain student confidentiality, reported results are aggregated and schools are not individually identified. Thus these findings should not be construed as an audit or an assessment of the effectiveness of the educational environment in any specific school.

Readers of earlier reports may notice slight differences in a few of the statistics reported for earlier years. These differences are attributable to improvements in collecting data (such as search for student names in MPS records), errors in reporting, survey results that come in after the reports are prepared, and data entry and analysis errors that are subsequently corrected. Unless noted in the text, these differences are minor.

III. Choice Families and Students

Over the years we have reported on a number of important questions concerning the students and families who applied to the choice program. How many families apply? How many seats are available in the private schools? How do families learn about the program? Why do they want to participate? What are the demographic characteristics of students and families? What are parental attitudes toward education? What are the prior experiences of applicants in public schools? How well were the students doing in their prior public schools?

Findings are very consistent over the five years of the program. For economy, and because five years of data provide a better picture than a single year, most tables contain combined data from 1990 to 1994. The most appropriate comparison group to the choice families, on most measures, is the low-income MPS sample. That group, which includes about two-thirds of Milwaukee students, is defined as qualifying for free or reduced-priced lunch. The income level for reduced-priced lunch is 1.85 times the poverty line, free lunch is 1.35 times the poverty line. Almost all low-income students qualify for full free lunch.

Enrollment in the Choice Program. Enrollment statistics for the Choice Program are provided in Table 1. Enrollment in the Choice Program has increased steadily but slowly, never reaching the maximum number allowed by the law. September enrollments have been 341, 521, 620, 742, and 830 from 1990–91 to 1994–95. The number of schools participating was: 7 in 1990–91, 6 in 1991–92, 11 in 1992–93, and 12 in the last two years. The number of applications has also increased, with again the largest increase in 1992–93. In the last two years, however, applications have leveled off at a little over 1,000 per year. Applications exceeded the number of available seats (as determined by the private schools) by 171, 143, 307, 238 and 64 from 1990–91 through 1994–95. Some of these students eventually fill seats of students who are accepted but do not actually enroll. The number of seats available exceed the number of students enrolled because of a mismatch between applicant grades and seats available by grade. It is difficult to determine how many more applications would be made if more schools participated and more seats were available. In 1992–93, when the number of participating schools increased from 6 to 11, applications rose by 45%. In the last two years, however, seats available increased by 22% and 21%, but applications increased only by 5% from 1992–93 to 1993–94 and declined this past year.

Learning About Choice and the Adequacy of Information on Choice. Table 2 indicates how survey respondents learned about the Choice Program. The results are fairly similar

for the first four years. The most prevalent source of information on choice remains friends and relatives, which essentially means word-of-mouth information. That informal communication is almost double the frequency of other sources. It also increased in 1994.

Parental satisfaction with the amount and accuracy of that information, and with the assistance they received, is presented in Table 3. Satisfaction with the amount of information on the program in general is high in all years and even higher in 1994. There is, however, a difference in satisfaction of parents selected and not selected into the choice schools. Looking ahead to Table 6, when we create additive scales for these questions measuring satisfaction with administration of the Choice Program, we see a large difference between those applicants who enroll in the Choice Program and those not selected.[3] As indicated in the rows for Scale T3 (top panel, Table 6), choice enrollees for 1990–94 have a mean dissatisfaction of 11.6, while non-selected parents have a much higher average dissatisfaction score of 14.4.

Why Choice Parents Participated in the Choice Program. Table 4 provides the responses to survey questions rating the importance of various factors in parents' decisions to participate in the Choice Program. The results are consistent across the years. The leading reasons given for participating in the Choice Program [were] the perceived *educational quality* and the *teaching approach and style* in the private schools. That is followed by the *disciplinary environment and the general atmosphere* that parents associate with those schools. Frustration with prior public schools, although not unimportant, was not as important a reason for applying to the Choice Program as the attributes of the private schools. At the bottom of the list are siblings in the school and the location of the school.

Demographic Characteristics of Choice Families and Students. The Choice Program was established, and the statute written, explicitly to provide an opportunity for relatively poor families to attend private schools. The program has clearly accomplished this goal. Relevant demographic statistics are presented in Table 5, which, unless otherwise noted, are based on our surveys. For comparison purposes seven groups are identified, including applicants in the most recent year (1994), and the combined year samples for choice applicants, choice students actually enrolled, students not selected, students who left the program (and did not graduate), and the MPS control groups.

In terms of reported family income (Table 5a), the average income was $11,630 in the first five years. Incomes of 1994 choice families increased considerably to an average of $14,210. Low-income MPS parents reported a slightly higher family income, which averaged $12,100. The average in the full MPS control group was $22,000.

In racial terms, the program has had the greatest impact on African-American students, who comprised 74% of those applying to choice schools and 72% of those enrolled in the first five years of the program (Table 5b). Hispanics accounted for 19% of the choice applicants and 21% of those enrolled. Both of these groups were disproportionately higher than the MPS sample. Compared with the low-income MPS sample, however, Hispanics were the most overrepresented, with Asians and White students the most underrepresented. The overrepresentation of Hispanics was due to a new building and considerable expansion in capacity of Bruce-Guadalupe Bilingual School, which has an emphasis on Spanish culture and is located in a neighborhood with an increasing Hispanic population.

In terms of marital status (Table 5c), choice families were much more likely to be headed by a non-married parent (75%) than

the average MPS family (49%), and somewhat more likely than the low-income MPS parent (65%). The percentage was almost identical for the five separate years.

One important difference between MPS and choice applicants was in family size (Table 5d). For the combined years, only 42% of the choice families reported having more than two children. The average number of children in families applying to the choice program was 2.54. This compared with 54% of the MPS families having more than 2 children (2.95 children per family) and 65% of the low-income MPS families (3.24 on average).

A unique characteristic of choice parents was that despite their economic status, they reported higher education levels than either low-income or average MPS parents (Table 5e). Over half of the choice mothers reported some college education (56%), compared with 40% for the entire MPS sample and 30% of the low-income MPS respondents. That number was even higher in 1994, with 64% of the mothers new to the program reporting some college. That was consistent with the higher incomes reported for the last year (Table 5a, column 2). The biggest difference in education appears in the category titled "some college." Although fathers more closely match the MPS control groups, they were also somewhat more educated.

The Importance of Education and Educational Expectations. Based on our measures, choice and MPS parents were similar in terms of the importance they place on education. We measured the importance of education relative to other important family values. Scale descriptive statistics are in Table 6.

Educational expectations were high for all groups, with choice parents somewhat higher than MPS and low-income MPS parents. Eighty-six percent of choice parents in the first four years indicated that they expected their child to go to college or do post-graduate work. This compared with 76% of the MPS parents and 72% of the low-income MPS parents. Because sample sizes were large, these proportions were significantly different.

Experience of Choice Parents in Prior Public Schools. A more complete picture of choice parents includes the level of parental involvement, attitudes toward, and student success in their child's prior public school. Our surveys measured the degree of parental involvement in the school, the amount of parental help for children at home, and parent satisfaction with prior schools. The results are presented in Table 6. Prior achievement test results are provided in Table 8.

Based on five years of highly consistent data, the conclusions we draw are as follows: (1) that choice parents were significantly more involved in the prior education of their children than MPS parents (Table 6, Scales A3–A5); (2) that they expressed much higher levels of dissatisfaction with their prior schools than MPS parents (Table 6, Scale A6); and, (3) that their children had lower prior achievement test results than the average MPS student (Table 8).[4]

Choice students had prior test scores at or, in some years below, the low-income MPS students. The test scores in Table 8 are Iowa Test of Basic Skills (ITBS) which are given in grades 1–8. The tests used are the Reading Comprehension Test and the Math Total Test. The latter consists of three subtests: Math Concepts, Math Problem Solving, and Math Computations.[5] The results reported are for the last test taken while the student was in MPS. The reason for this is that we are trying to get as complete a picture as possible of a relatively small number of choice students. The majority of those tests were taken in the spring of the year of application.[6]

The prior test scores for the last year, 1994, were the lowest of any cohort, especially in math. This may have been due to the small number of students who completed all portions of the math test (because of changes in MPS testing requirements). In any event, the pattern remained consistent.

The absolute level of the scores indicates the difficulty these students were having prior to entering the program. The median national percentile for choice students ranged from 25.5 to 31, compared with the national median of 50. The Normal Curve Equivalent (NCE), which is standardized to a national mean of 50, ranged from 35.5 to 39.8, which is about two-thirds of a standard deviation below the national average. *In short, the choice students were already very low in terms of academic achievement when they entered this program.*

IV. Outcomes

We discuss five outcome measures in this section: (1) achievement test results; (2) attendance data; (3) parent attitudes; (4) parental involvement; and 5) attrition from the program. The legislation specified that suspension, expulsion, and dropping out also be monitored. Those measures, however, would be meaningful only at the high-school level.[7] The high-school-level choice schools consisted of alternative education programs for seriously at-risk students. Therefore it is unclear what the relevant MPS comparisons would have been.

Achievement Test Results

Cohort Test Results. Table 9 provides the aggregate test results for 1991 to 1995 for choice students and for 1991 to 1994 for students taking tests in MPS in the respective years. Tests were administered in April or May of each year. The results may be compared only crudely with those in Table 8, which indicated prior test scores for students accepted into the Choice Program. The prior test data in Table 8 were mostly from the previous spring, but were based on the last prior test in the student's file. As stated above, those data indicated the choice applicants were clearly behind the average MPS student, and also behind a large random sample of low-income MPS students.

As noted in the introduction, test scores from the MPS sample were not available for 1995. MPS is only requiring the ITBS of fifth-grade students due to state-mandated tests, which have been substituted for the ITBS. In addition, the ITBS version given to fifth graders is a new and very different test which is not substantively comparable to the older versions used in the choice school. Choice schools were not required to take the state tests. Thus we report on only choice 1995 tests, and change scores are analyzed using only the first four years of data.

In reading, Table 9 indicates that the choice students tested in 1991 did better than the MPS low-income group, but in math they did somewhat worse. Tests taken in the choice schools in the second year—spring 1992—were considerably lower. They were lower than the scores in both the full MPS sample and among the low-income MPS students. Comparing the more relevant low-income MPS and choice students, the choice students mean math NCE was more than five points lower; the reading score was two points lower.

For 1993, the scores of choice students were approximately the same in reading as in the prior year, but were considerably higher in mathematics. Although the reading scores were lower than the low-income MPS sample, the math scores had the same median National Percentile Ranking, but were 2.3 NCE points higher.

The 1994 spring scores for all choice stu-

dents were slightly higher than the 1993 choice cohort, and very similar to the MPS low-income group on both reading and mathematics. In contrast to the first two years, for both 1993 and 1994, choice students did better in math than in reading (which is the pattern for MPS students in all years).

Choice scores in 1995 were very similar to the scores in 1994. In terms of NCEs, they were almost identical. In terms of median National Percentile Rankings, the 1995 scores of choice students were 2% lower than in 1994 for both reading and math.

Because the cohort scores do not report on the same students from year to year, and because we hope that schools are able to add to the educational achievement of students, the most accurate measure of achievement are "value-added," change scores.

Change Scores. Analysis of year-to-year change scores for individual students produced somewhat different results.[8] By comparing students' scores with national samples, we can estimate how individual students changed relative to the national norms over the year. Descriptive change scores are depicted in Tables 10a to 10e for the respective years. We caution the reader, as we have in previous reports, that sample sizes in some years are quite small for choice students. The results in the table are based on differences in NCEs, subtracting the first-year score from the second.[9]

In the first year, choice students gained in reading, but math scores stayed essentially the same. MPS students improved considerably in math, with smaller gains in reading. Both low-income and non-low-income MPS students gained in math. Three of these gains were statistically significant (due in part to larger sample sizes), while the gains for choice students were not.

There were quite different effects in the second year (Table 10b). Change scores for choice students in math, and for MPS students in both reading and math, were not appreciable. None of these differences approached standard levels of statistical significance. In contrast to the first year, however, reading scores dropped for choice students. The decline was 3.9 NCE points for all students between 1991 and 1992. Because NCE scores are based on a national norm, this means that choice students scored considerably below the national sample in the prior year. The decline was statistically significant at the .001 level.

The results shifted again in the third year. Choice students declined slightly in reading, which was not significant. On the other hand, for the first time, math scores for choice students improved. The mean math NCE went from 38.3 to 42.7 for a 4.4 NCE gain, which was statistically significant. Scores for MPS students, on the other hand, declined for both tests and for both groups. Because of relatively large sample sizes for the MPS control group, the decline in math scores was significant and estimated to be 1.2 NCEs for both the total MPS control group and the low-income sample.[10]

The results for 1994 again shifted for both groups. For all groups, reading scores effectively did not change. The same was true of math scores for the MPS groups, although the low-income MPS scores decline[d] slightly more than the non-low-income group. For the choice students, after a large math increase in 1993, there was a decline of 2 NCE points in 1994.

In 1995, reading scores for choice students essentially remained the same, but again there was a 1.5 drop in NCE math scores.

Regression Analysis. Because test score differences could be based on a number of factors, and the factors could be distributed differen-

tially between choice and MPS students, it is necessary to control statistically for factors other than if students were in MPS or choice schools. Those controls are provided by multivariate regression analysis of the combined MPS and choice samples. There are a number of ways to model achievement gains. The most straightforward is to estimate the second-year test score, controlling for prior achievement and background characteristics, and then include an indicator (dichotomous) variable to measure the effect of being in a choice or MPS school. A series of dichotomous variables indicating the number of years in the choice program can be substituted for the single choice indicator variable.[11]

In previous year reports regressions were provided for each yearly change. The conclusion of those analyses, which were comprehensively described in the *Fourth Year Report* (Tables 12–16), was that there was no consistent difference between choice and MPS students, controlling for a wide range of variables.

For annual reporting purposes, and to test for consistency of results between years, previous reports included yearly regressions. For those analyses, small sample sizes precluded including survey variables. In this report, we "stack" four years of change scores and thus we have included survey variables in the regressions reported in Tables 12a and 12b. Critics of earlier reports surmised that exclusion of these variables biased the results against choice. Unfortunately, as Witte predicted in response to critics, *the results of including these variables do not favor the choice schools.*[12]

The *B* columns in Tables 11 and 12 contain the coefficients that determine the effects of the variables on predicting the dependent variable (the relevant test). In both tables, a critical coefficient will be the "Choice" (dichotomous) indicator variable. Dichotomous variables have a value of 1 if the student has the characteristic, and 0 if they do not. The effects of the coefficient can be read as a straight difference in the NCE score on the respective test being estimated. Thus, for example, in Table 11a, being a low-income student has an estimated negative effect of 2.510 NCE points on the reading score of students taking two tests a year apart from 1990 to 1994. Similarly, the effect of being in a private school in the Choice Program (as compared to being in MPS) produced an estimated decrease in Reading of .568 Normal Curve Equivalent points. Because of the size of the coefficient relative to its standard error, the negative impact of being in a low-income family is considered statistically significant, but not the effect of being in a choice school. These predicted effects occurred after simultaneously controlling, or in essence keeping constant, all other variables in the model.

The variables in the models that are not dichotomous variables are also easy to interpret. They can be read as the effect that a *one point difference in the variable* has on the predicted test. Thus, for reading in Table 11a, if a student is one NCE point better on the prior reading test than another student, we would predict he or she would do .495 points better on post-test. For "Test Grade" for each additional grade students are in, we would estimate a lower reading score of .176 NCE points—lower because the sign of the coefficient is negative.

The probabilities indicated in the tables are a statistical measure of how likely the coefficients are to differ from 0. Traditionally, those coefficients that have a probability of .05 or less are considered "significant."

Table 11 differs from Table 12 in that the latter includes a full set of survey variables, with a more accurate measure of income than free-lunch, mother's education, marital status, and educational expectations for their children. There are also four scales of paren-

tal involvement, paralleling scales A2–A5 in Table 6. Also notice that the sample sizes decline from 4716 and 4653 (in Table 11) to 1385 and 1372 (in Table 12).

There are several consistent patterns across all the models. Prior tests were always good predictors of post-tests, and the coefficients were relatively stable no matter how the results were modeled. Income was also always significant, measured either as a low-income dichotomous variable (qualifying for free lunch) or as family income in thousands of dollars. Higher income was always associated with higher achievement gains. Gender was significantly related to greater achievement in three of the models, with girls scoring higher than boys. Test grade was only significant on math models (tables 11b and 12b). The higher the grade of the student, the lower they scored.

Racial effects varied, although African American students always did less well than white students. For Hispanics and other minority students, the coefficients relative to whites were usually negative, but only significantly so for reading as shown in Table 11a.

As we anticipated in an earlier response to our critics, most of the other survey variables included in Tables 12a and 12b were not statistically significant. Mother's education was significant for reading, and two parental involvement variables were barely significant for math. The parental involvement variables, however, went in the wrong direction: greater involvement predicted *lower* test scores. The reason these variables were generally insignificant, and probably the reason the sign of the parental involvement coefficients were negative, was that these were correlated with other variables in the equations. This correlation, and the small sample sizes, produce unstable results and large standard errors for the estimates of the Bs.

The effects of being in a choice school were somewhat more discouraging. With the large samples, excluding the survey variables, none of the choice variables were even close to significant. In model 1, including the simple indicator variable of being in a choice school or not (Tables 11a and 11b), the coefficients were very close to zero, with large standard errors. The same was generally true when, rather than simply indicating whether a student was in a choice school, we included a set of indicator variables for the years they had been in choice (thus the "Choice Year 2" variable is 1 for a second-year choice student and 0 for everyone else). The negative 1.2 on reading for two-year students was the only one of these variables that was close to significant (Table 11a, model 2).

The one exception to the result that there were no differences between public school students and choice students was for reading scores including the survey variables (Table 12a). In that model, choice students' reading scores were estimated to be 2.15 NCE points *lower* than MPS students—controlling for all the other variables. The estimate was significant at the .01 level of probability. When analyzed one year at a time in the Choice Program, the effect was mostly due to second-year students. This result may well be an artifact of the poor performance of choice students in the second year of the program (1991–92). In earlier reports it was estimated that the difference in reading in that year was -3.35 points (significant at the .001 level). It was also previously reported that after that year, many of the poorest students left the choice schools (see "*Second Year Milwaukee Parental Choice Report,*" pp. 16–17, Tables 20 and 21).

The reason choice estimates were more negative in this model than in other models in Table 11a and 11b, was probably due to the inclusion of mother's education. The variable was positively related to higher achievement and was significant in the model. Choice mothers were more educated than MPS mothers, which

should have produced higher scores for choice students, but did not. Thus when we controlled for education of mothers, the estimated impact of choice schools was more negative than when we did not. As noted above, this result was produced with a much smaller sample size, and with high standard errors on a number of variables. Thus it should be interpreted with caution. What clearly cannot be concluded, however, is that the inclusion of survey variables improves the results for choice students relative to public school students. At best they did as well as MPS students, and they may have done worse.[13]

Attendance

Attendance is not a very discriminating measure of educational performance at this level because there is little school-to-school variation. For example, in the last three years, average attendance in MPS elementary schools has been 92% in each year. Middle school attendance for the same years averaged 89, 88 and 89%. Attendance of choice students in the private schools (excluding alternative schools) averaged 94% in 1990–91 and 92% in 1991–92, which puts them slightly above MPS, but the differences were obviously slight. In 1992–93, excluding SEA Jobs and Learning Enterprises, attendance at the other schools was 92.5%. For 1993–94, attendance in the nonalternative schools (thus excluding Exito and Learning Enterprises) was 93%. For 1994–95 attendance dropped in the choice schools to 92%. It can be concluded that overall attendance was satisfactory and on average not a problem in choice schools.

Parental Involvement

Parental involvement was stressed in most of the choice schools and, in fact, was required in the contracts signed by parents in several of the schools. Involvement took several forms: (1) organized activities that range from working on committees and boards to helping with teas and fund raising events; (2) involvement in educational activities such as chaperoning field trips, and helping out in the classroom or with special events.

Table 6 provides five-year data on parental involvement scales for choice parents both in their prior public schools and in the private schools their students attended under the Choice Program. Table 6 provides the means and standard deviations of the scales for the applicants from 1990 to 1994 and the comparable scale scores for choice parents from 1991 to 1995 June surveys.

Of the four types of parental involvement we measured, already very high levels of involvement significantly increased in the private choice schools in three areas (Table 6, Scales A2, A3, and A4). These include parents contacting schools, schools contacting parents, and parental involvement in organized school activities. The increases were significant at the .001 level. The one exception was parent involvement in educational activities at home (Scale A5). For that scale, involvement increased but the increase was not statistically significant. These results have been confirmed independently in each of the five years of the program.

Parental Attitudes

Parental Satisfaction with the Choice Schools. In all five years, parental satisfaction with choice schools increased significantly over satisfaction with prior public schools. School satisfaction is indicated by scale A6 in Table 6. As described above, satisfaction with their prior schools was significantly lower than satisfaction of the average or low-income MPS parent. Reported satisfaction with the choice schools is indicated in these tables as "Choice Private School 1991–95" and was measured on the spring surveys in each year.

Another indication of parent satisfaction was the grade parents gave for their children's school. On the follow-up survey in June of each year, we asked parents to grade the private school their children attended. The grades, which indicate substantial differences with the grades they gave their prior MPS schools, are given in Table 7. For the five-year period, the average prior grade (on a scale where an A is 4.0) improved from 2.4 for prior MPS schools to 3.0 for current private schools. The overall grades were relatively consistent for each year and were always above the grades given to prior MPS schools.

Attitudes Toward the Choice Program. Follow-up surveys in June of each year asked parents of choice students if they wanted the Choice Program to continue. Respondents almost unanimously agreed the program should continue. Over the five years, 98% of the respondents felt the program should continue.

Attrition From the Program

Attrition Rates. One of the issues concerning the Choice Program is the rate of attrition from the program. Attrition rates, calculated with and without alternative high school programs are presented in the last two rows of Table 1. Attrition from the program was not inconsequential, although there appears to be a downward trend that leveled off in the last year. Overall attrition, defined as the percentage of students who did not graduate and who could have returned to the choice schools, was 46%, 35%, 31%, 27%, and 28% over the five years. Excluding students in alternative programs, the rates were 44%, 32%, 28%, 23%, and 24%. Whether these attrition rates are high or not has been discussed at length in prior reports. A Legislative Audit Bureau report on the program last year concluded it was a problem. We interpreted it as a problem for both public and private schools in that the public school student mobility rate was similar in magnitude to the attrition rate from choice schools.

Why students did not return to the choice schools? Because analysis of the causes of attrition was not part of the original study design, by the time we realized how many students were not returning after the first year, we were unable to follow up with non-returning families. That was rectified in the following years by using very brief mail surveys or telephone interviews. The surveys and interviews simply asked where the students were currently enrolled in school (if they were), and (open ended) why they left the choice school. The response rate to our inquiries has been 39%. This rate of return was slightly lower than for the other surveys. The normal problems of mailed surveys were compounded by the fact that we did not know who would return until enrollment actually occurred in September. Thus in many cases addresses and phones numbers were not accurate. The largest bias in our responses was undoubtedly families who moved out of the Milwaukee area and did not leave forwarding addresses. Telephone searches were impossible for that group. Our results should thus be treated with some caution.

The reasons parents gave for leaving are presented in Table 13. Approximately 18% o the responses indicated child or family specific reasons—including moving. We suspect that this category is underestimated since we undoubtedly were not able to locate as high a proportion of families who moved out of the area.

Almost all of the remaining respondents were critical of some aspect of the Choice Program or the private schools. The leading problems with the program were transportation problems, difficulties in reapplying to the program, problems with extra fees charged by some schools, and the lack of religious train-

ing, which was not allowed by the statute. Within-school problems most often cited were unhappiness with the staff, usually teachers, dissatisfaction with the general quality of education, and perceptions that discipline was too strict. The lack of special programs, which might have been available elsewhere, was also cited in 6% of the responses.

Outcome Summary

Outcomes after five years of the Choice Program remain mixed. Achievement change scores have varied considerably in the first five years of the program. Choice students' reading scores increased the first year; fell substantially in the second year, and have remained approximately the same in the next three years. Because the sample size was very small in the first year, the gain in reading was not statistically significant, but the decline in the second year was. In math, choice students were essentially the same in the first two years, recorded a significant increase in the third year, and then significantly declined this last year.

MPS students as a whole gained in reading in the first two years, with a relatively small gain in the first year being statistically significant. There were small and not significant declines in the last two years. Low-income MPS students followed approximately the same pattern, with none of the changes approaching significance. Math scores for MPS students were extremely varied. In the first year there were significant gains for both the total MPS group and the low-income subgroup. In the second year, the scores were essentially flat, but in the third year they declined significantly. Again, in the fourth year there was essentially no change in either the total MPS or low-income MPS groups.

Regression results, using a wide range of modeling approaches, including yearly models and a combined four-year model, generally indicated that choice and public school students were not much different. If there was a difference, MPS students did somewhat better in reading.

Parental attitudes toward choice schools, opinions of the Choice Program, and parental involvement were very positive over the first five years. Attitudes toward choice schools and the education of their children were much more positive than their evaluations of their prior public schools. This shift occurred in every category (teachers, principals, instruction, discipline, etc.) for each of the five years. Similarly, parental involvement, which was more frequent than for the average MPS parent in prior schools, was even greater for most activities in the choice schools. In all years, parents expressed approval of the program and overwhelmingly believed the program should continue.

Attrition (not counting students in alternative choice schools) has been 44%, 32%, 28%, 23%, and 24% in the five years of the program. Estimates of attrition in MPS are uncertain, but in the last two years, attrition from the Choice Program was comparable to the range of mobility between schools in MPS. The reasons given for leaving included complaints about the Choice Program, especially application and fee problems, transportation difficulties and the limitation on religious instruction. They also included complaints about staff, general educational quality and the lack of specialized programs in the private schools.

V. A Brief Response to Some of Our Critics

Several choice supporters criticized our previous reports in terms of outcomes they interpreted as negative for choice. Positive results were usually lauded and accepted as gospel.[14] They focused on two sets of results: high attrition from the program, and test

score results. Our attrition numbers are straightforward and accurate. We counted students who could have returned to choice schools, not subject to random selection, but chose not to return. These statistics were highlighted in an independent report by the Wisconsin Legislative Audit Bureau. Many students did not stay in these schools. It is that simple.

In terms of test scores, three points require a response. The first is the belief that inclusion of the survey variables would aid the results for choice students, and thus omitting them from the analysis biased the results against choice. In general, the excluded variables were correlated with variables already in the earlier equations and thus did not have significant effects. For the most part that is what happened in the regressions depicted in Tables 12a and 12b.

If the survey variables would have an effect, however, the *a priori* prediction would be that inclusion would favor the public schools, not the reverse. The reason is that on most of the omitted variables choice families were higher, and we would expect the variables to be positively correlated with greater achievement. For example, we would predict that the greater a mother's education, the more her involvement in her child's education, and the higher her educational expectation for her child, the higher her child's achievement.[15] *Because those variables were higher for choice families, holding those factors constant should lower, not increase, the relative scores for choice families. And that is precisely what happened on reading scores (Table 12a).*

Second, several commentators and newspaper reports have picked up on claims that because choice students started with low test scores, they should be expected to do worse than other students. This is undoubtedly true in absolute terms, but not likely in terms of *change scores.* That is why, in both regressions and descriptive statistics, we emphasized the changes in test scores, not the cohort scores. In fact, because of regression to the mean effects, those scoring near the bottom of a distribution on a prior test are likely to improve more than those scoring initially higher.[16] That means that in terms of change scores, choice students would be advantaged, not disadvantaged by initially lower cores.

Finally, and the most absurd of all, is the charge that we did not control for poor English speakers, and that biased the results against choice because of the disproportionate number of Hispanics in the program. Unfortunately, there is no measure of this variable available in any of the data sets. Hispanic origin (which has always been included in the analyses) probably is a good proxy, but certainly not a perfect one.

However, again, when considering *changes scores*, the exact opposite result is likely. Regression to the mean will also occur for these students and thus favor students with poor English on a second test. These students, however, will also be getting a year of instruction in English between the tests. Think of a child coming into a lower grade knowing little English. The student will obviously not do well on either the initial reading or math tests (because math tests also require reading). The student is then immersed in English for the first time. Relative to English speakers, what will the *change score* look alike after one year?

The analysis of the critics is so blatantly erroneous that it seems as though either they have little statistical or research training, or they simply do not care and are motivated by other factors, or both.

VI. Conclusions and Recommendations

We ended the *Fourth Year Report* by summarizing the positive and negative consequences of the program. We then wrote:

TABLE 1

Participation and Attrition from the Choice Program, 1990–1995

	1990–91	*1991–92*	*1992–93*	*1993–94*	*1994–95*
Number of students allowed in the Choice Program (limited to 1% of MPS enrollment)/ 1.5% 1994–95	931	946	950	968	1450
Number of private non-sectarian schools in Milwaukee	22	22	23	23	23
Number of schools participating	7	6	11	12	12
Number of applicants	577	689	998	1049	1046
Number of available seats	406	546	691	811	982
Number of students participating					
September count	341	521	620	742	830
January count	259	526	586	701	—
June count	249	477	570	671	—
Graduating students	8	28	32	42	45
Number of returning Choice students	NA	178	323	405	491
Attrition rate[1]	.46	.35	.31	.27	.28
Attrition rate without alternative schools	.44[2]	.32	.28	.23	.24

1. The attrition rate for year *t* is defined as: 1.0 – [the number of returning students in year *t* + 1/ (the September count in year *t* - graduating students in year *t*)]

2. If Juanita Virgil Academy is excluded, the attrition rate is .29.

> Honorable people can disagree on the importance of each of these factors. One way to think about it is whether the majority of the students and families involved are better off because of this program. The answer of the parents involved, at least those who respond to our surveys, was clearly yes. This is despite the fact that achievement, as measured by standardized tests, was no different than the achievement of MPS students. Obviously the attrition rate and the factors affecting attrition indicate that not all students will succeed in these schools, but the majority remain and applaud the program.

Although achievement test results may be somewhat more bleak for choice in this analysis, the differences are not very large in terms of their impact, and the negative estimates are based on less stable models and smaller sample size. In addition test scores are only one indication of educational achievement. Thus we see no reason to change last year's conclusion.

Our recommendations are no different from those offered in previous years. For those interested, they are stated in their entirety in the *Fourth Year Report.*

TABLE 2

How Choice Applicants Learn about the Program, 1990–1994
(Percent indicating source used)

	1990–94	*1994*
Friends or relatives	50.9	58.3
Television or radio	21.4	8.6
Newspapers	24.2	5
Private schools	17.7	20.1
Churches	3.7	2.9
Community centers	6.7	10.1
(N)	(1,060)	(139)

Question: "How did you learn about the Private School Choice Program?"

TABLE 3

Satisfaction with Information and Assistance on the Choice Program, Applicants 1990–94 (Percent satisfied or very satisfied)

	1990–94	*1994*
Amount of information on the Choice Program	76.4	79.4
Accuracy of information on the Choice Program	75.3	80.8
Amount of information on the Private Schools	66.3	65.9
Accuracy of information on the Private Schools	70.9	76.1
Assistance from school you applied to	76.8	75.2
Assistance from Dept. of Public Instruction in Madison	64.8	78.2
(N)	(1,060)	(139)

Question: "How satisfied were you with the following?"

TABLE 4

Factors Affecting Decisions to Participate in Choice Program, Applicants 1990–94 (Percentages)

	1990–1994				*1994*			
	Very Import.	*Import.*	*Some Import.*	*Not Import.*	*Very Import.*	*Import.*	*Some Import.*	*Not Import.*
Educational Quality in Chosen School	88.6	10.5	0.7	0.30	90.4	9.6	0.00	0.00
Teaching Approach or Style[3]	85.7	13.2	0.7	0.4	86.8	12.5	0.7	0
Discipline in Chosen School	74.6	22.1	2.9	0.40	68.4	26.5	4.4	0.7
General Atmosphere in Chosen School	74.3	22.5	2.7	0.50	73.4	21.9	4.4	0.00
Class Size[3]	72	22.4	4.9	0.7	66.7	25.2	6.7	1.5
Financial Considerations	69.8	23.3	4.8	2.1	74.6	18.1	3.6	3.6
Special Programs in Chosen School	68.4	25.7	3.6	2.3	68.1	23.7	6.7	1.5
Location of Chosen School	60.5	22.2	11.7	5.6	64.2	24.1	7.3	4.4
Frustration with Public Schools	59.5	23.9	10.7	5.9	61.3	20.4	14.6	3.6
Other Children in Chosen School	36.6	29.2	14.8	19.4	36.1	30.1	3.5	20.3
(N)	(964)				(139)			

Question: "Please rate all of the following issues and their importance in your decision to participate in the Choice program."

3. This question was not asked in the first two years of the study. The N for these responses is 581 in the 1990–94 responses.

TABLE 5 DEMOGRAPHICS

TABLE 5A

Household Income (Percentages)

Thousands	***1*** *Choice Applied 1990–94*	***2*** *New Choice 1994*	***3*** *Choice Enrolled 1990–94*	***4*** *Choice Non-Select 1990–94*	***5*** *Attrition 1990–93*	***6*** *Low-Inc MPS 1991*	***7*** *MPS Control 1991*
$0–$5	17.7	12.8	16	19.4	18	19	13
$5–$10	37.2	33	39.6	33.2	40	34	23
$10–$20	28.7	24.1	29.6	28.9	27	29	21
$20–$35	15.8	29.3	14.7	16.9	15	14	24
$35–$50	.6	0	0.20	1.2	1	3	13
$50 & over	0.10	0.80	0.00	0.30	0	0	8
(N)	(1,020)	(133)	(627)	(325)	(293)	(880)	(1,513)
Mean Income	11.63	14.21	11.34	12.24	11.77	12.13	22.00

TABLE 5B

Race[4] (Percentages)

	1 *Choice Applied 1990–94*	***2*** *New Choice 1994*	***3*** *Choice Enrolled 1990–94*	***4*** *Choice Non-Select 1990–94*	***5*** *Attrition 1990–93*	***6*** *Low-Inc MPS 1991*	***7*** *MPS Control 1991*
African American	74.5	71.9	73.2	78.0	75	67	55
Asian	0.40	0	0.20	0.20	0	5	4
Hispanic	18.7	19.4	21.3	14.2	16	11	10
Native American	0.4	0.00	0.1	0.7	1	1	1
White	4.9	7.9	4.9	4.3	7	15	29
Other	1	0.70	0.3	2.6	1	1	1
(N)	(2,673)	(139)	(1,490)	(886)	(676)	(3,179)	(5,365)

[4] Based on MPS Student Record Data Base (SRDB). In cases where Choice Students could not be found in the SRDB, race data were taken from the school enrollment sheets.

TABLE 5C

Parent Marital Status (Percentages)

	1 *Choice Applied 1990–94*	***2*** *New Choice 1994*	***3*** *Choice Enrolled 1990–94*	***4*** *Choice Non-Select 1990–94*	***5*** *Attrition 1990–93*	***6*** *Low-Inc MPS 1991*	***7*** *MPS Control 1991*
Married	24.9	23	23.5	29.6	24	35	51
Single	38.9	42.2	39.8	35	42	32	22
Separated	12.4	17.8	13	11.5	9	11	8
Divorced	15.1	9.6	15.6	13.9	15	14	13
Widowed	3.2	.7	3.5	2.7	5	2	2
Living Together	5.5	6.7	4.6	7.3	5	6	4
(N)	(1,032)	(135)	(633)	(331)	(294)	(924)	(1,637)

TABLE 5D

Family Size (Percentages)

Children per Family	***1*** *Choice Applied 1990–94*	***2*** *New Choice 1994*	***3*** *Choice Enrolled 1990–94*	***4*** *Choice Non-Select 1990–94*	***5*** *Attrition 1990–93*	***6*** *Low-Inc MPS 1991*	***7*** *MPS Control 1991*
1	27.8	31.7	25.9	29.7	19	9	13
2	30.2	30.9	31.9	28	31	26	33
3	19.2	18.7	22.3	14	24	27	26
4	12.5	8.6	11.5	13.4	16	18	151
5 or more	10.3	10	8.3	14.9	10	20	1,415
(N)	(1,060)	(139)	(645)	(343)	(287)	(908)	(1,611)
Avg. number of children per family	2.54	2.4	2.5	2.66	2.72	3.24	2.95

TABLE 5E

Parents' Education (Percentages)

	8th Grade	***Some HS***	***GED***	***H.S.***	***Some College***	***College Grad***	***Some Post Grad***	***(N)***
Choice Applied 1990–94								
Mother/Female Guardian	4.3	11.6	9.2	20.3	46.4	6.5	1.8	(1,027)
Father/Male Guardian	10.4	20.4	6.9	25.8	28.8	5	2.7	(624)
New Choice 1994								
Mother/Female Guardian	.8	9.8	9.8	15.9	53	6.8	3.8	(132)
Father/Male Guardian	6.2	16	6.2	23.5	33.3	4.9	9.9	(81)
Choice Enrolled 1990–94								
Mother/Female Guardian	3	12	9.3	21.2	45.5	7.5	1.4	(626)
Father/Male Guardian	7.6	19.3	7	28.2	28.5	6.3	3.1	(383)
Choice Non-Select 1990–94								
Mother/Female Guardian	6.9	10.8	9	19.2	46.2	5.7	2.1	(333)
Father/Male Guardian	15.6	22.1	6.5	20.6	30.2	3.5	1.5	(199)
Attrition 1990–93								
Mother/Female Guardian	4	9	11	19	48	8	1	(292)
Father/Male Guardian	6	18	8	32	30	4	2	(175)
Low-Inc MPS 1991								
Mother/Female Guardian	12	25	9	25	26	3	1	(881)
Father/Male Guardian	15	22	9	25	21	6	2	(535)
MPS Control 1991								
Mother/Female Guardian	8	18	7	28	29	6	5	(1,525)
Father/Male Guardian	9	16	8	26	27	9	6	(1,127)

TABLE 6

Scale Data — Means, Standard Deviations, Reliability, and Sample Size

Source of Scale, Range, and Direction by Group	*Mean*	*Standard Deviation*	*Alpha*	*(N)*
Dissatisfaction with the administration of the choice application process				
T3 – DisChAdm – Range = 6–24 (High = More Dissatisfied)				
Choice Applied 1990–94 (Fall)	12.6	4.4	.91	609
New Choice 1994 (Fall)	12.0	3.8	.88	86
Choice Enrolled 1990–94 (Fall)	11.6	3.6	.88	364
Non-selected Choice 1990–94 (Fall)	14.4	5.0	.92	206
Importance of education compared to other goals				
A1 – EdImport – Range = 7–15 (High = More Important)				
Choice Applied 1990–94 (Fall)	11.6	1.9	.72	996
New Choice 1994 (Fall)	11.4	1.8	.72	129
Choice Enrolled 1990–94 (Fall)	11.7	1.8	.70	627
Non-Selected Choice 1990–94 (Fall)	11.4	1.9	.73	301
Choice Private School 1991–95 (Spring)	11.5	1.9	.77	732
Non-Selected Choice 1991–95 (Spring)	11.3	1.9	.76	210
Low-Income MPS 1991	11.8	2.0	.74	811
MPS Control 1991	11.7	2.0	.71	1,554
Frequency of parent contacting school				
A2 – PiParScl – Range = 0–21 (High = More)				
Choice Applied 1990–94 (Fall)	8.8	4.9	.79	775
New Choice 1994 (Fall)	9.3	5.1	.78	88
Choice Enrolled 1990–94 (Fall)	8.6	5.1	.81	466
Non-Selected Choice 1990–94 (Fall)	9.0	4.6	.74	269
Choice Private School 1991–95 (Spring)	9.4	4.8	.78	722
Non-Selected Choice 1991–95 (Spring)	8.0	5.0	.83	212
Low-Income MPS 1991	5.8	4.4	.79	807
MPS Control 1991	6.0	4.3	.78	1,529
Frequency of school contacting parent				
A3 – PiSclPar – Range = 0–12 (High = More)				
Choice Applied 1990–94 (Fall)	3.6	2.9	.67	811
New Choice 1994 (Fall)	4.1	2.9	.67	94
Choice Enrolled 1990–94 (Fall)	3.7	2.9	.67	495
Non-Selected Choice 1990–94 (Fall)	3.5	2.7	.63	276
Choice Private School 1991–95 (Spring)	4.4	2.9	.70	740
Non-Selected Choice 1991–95 (Spring)	3.8	3.0	.71	223
Low-Income MPS 1991	2.7	2.5	.67	834
MPS Control 1991	2.7	2.5	.65	1,594
Parental involvement in school organizations				
A4 – PiSclOrg – Range = 0–5 (High = More)				
Choice Applied 1990–94 (Fall)	2.4	1.5	.71	788
New Choice 1994 (Fall)	2.6	1.4	.71	90
Choice Enrolled 1990–94 (Fall)	2.4	1.5	.72	481
Non-Selected Choice 1990–94 (Fall)	2.4	1.4	.69	268
Choice Private School 1991–95 (Spring)	3.0	1.3	.54	731
Non-Selected Choice 1991–95 (Spring)	2.3	1.4	.66	218
Low-Income MPS 1991	1.7	1.3	.67	831
MPS Control 1991	1.9	1.4	.67	1,586

TABLE 6 *(CONTINUED)*

Source of Scale, Range, and Direction by Group	*Mean*	*Standard Deviation*	*Alpha*	*(N)*
Parental involvement in educational activities with child				
A5 – PiChild – Range = 0–15 (High = More)				
Choice Applied 1990–94 (Fall)	8.7	3.5	.76	987
New Choice 1994 (Fall)	8.7	3.5	.76	129
Choice Enrolled 1990–94 (Fall)	8.8	3.5	.76	619
Non-Selected Choice 1990–94 (Fall)	8.7	3.5	.78	302
Choice Private School 1991–95 (Spring)	8.9	3.8	.81	737
Non-Selected Choice 1991–95 (Spring)	8.8	3.9	.82	218
Low-Income MPS 1991	7.5	4.3	.85	833
MPS Control 1991	6.9	4.2	.83	1,575
Dissatisfaction with prior school				
A6 – DisPrScl – Range = 8–32 (High = More Dissatisfied)				
Choice Applied 1990–94 (Fall)	16.5	5.5	.89	646
New Choice 1994 (Fall)	16.0	5.2	.88	69
Choice Enrolled 1990–94 (Fall)	16.4	5.7	.89	406
Non-Selected Choice 1990–94 (Fall)	16.7	5.1	.88	209
Choice Private School 1991–95 (Spring)	13.6	4.9	.90	604
Non-Selected Choice 1991–95 (Spring)	15.4	5.6	.92	178
Low-Income MPS 1991	14.4	4.2	.85	636
MPS Control 1991	14.5	4.2	.85	1,224

TABLE 7

Grade Given to Choice School Experience 1991–95 (Percentages)

	1 *Choice Private 1991*	*2* *Choice Private 1992*	*3* *Choice Private 1993*	*4* *Choice Private 1994*	*5* *Choice Private 1995*
A	40.1	33.0	29.5	42.0	39.7
B	37.4	39.4	35.9	32.7	37.5
C	14.3	22.0	21.1	17.5	18.0
D	2.7	3.2	9.7	6.3	4.1
F	4.5	2.3	3.8	1.5	0.7
GPA	3.04	2.89	2.78	3.07	3.11
(N)	(147)	(218)	(237)	(269)	(461)

TABLE 8
Prior Test Scores

	Applied Choice		*Low-Income MPS*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1990						
% At or Above 50% of NPR	23.3	31.1	27.2	.36.2	34.8	42.8
Median NPR	29	31	32	37	37	42
Mean NCE	39.1	39.7	40.1	42	43.6	45.8
Std. Dev. Of NCE	15.9	18.9	17	19.2	18.5	20.2
(N)	(262)	(257)	(2,136)	(117)	(3,157)	(3,130)
1991						
% At or Above 50% of NPR	27.1	22.6	28.2	36.2	36.1	43.4
Median NPR	26	30	32	38	38	43
Mean NCE	37.5	37.9	40.2	42.9	43.7	46.3
Std. Dev. Of NCE	16.8	17.7	17	19	18.6	20.2
(N)	(199)	(204)	(2,470)	(2,447)	(3,668)	(3,643)
1992						
% At or Above 50% of NPR	28.2	31.4	28.2	35.3	36.6	43.2
Median NPR	29	33	32	38	38	43
Mean NCE	40.0	40.3	40.2	42.4	43.9	46.0
Std. Dev. Of NCE	17.6	19.5	17.7	19.5	19.0	20.7
(N)	(234)	(226)	(2,839)	(2,801)	(4,024)	(3,991)
1993						
% At or Above 50% of NPR	25.7	26.7	28.2	34.7	37.5	42.1
Median NPR	29	28	32	37	37	37.5
Mean NCE	38.0	40.0	40.2	42.0	43.3	45.2
Std. Dev. Of NCE	19.1	18.6	17.7	19.3	19.0	20.6
(N)	(179)	(175)	(3,069)	(3,049)	(3,980)	(3,962)
1994						
% At or Above 50% of NPR	27.4	23.3	25.9	34.4	36.3	43.8
Median NPR	26.0	25.5	32.0	36.0	40.0	44.0
Mean NCE	37.3	35.8	38.7	41.4	42.9	46.0
Std. Dev. Of NCE	17.4	19.0	17.5	20.0	19.0	20.9
(N)	(146)	(86)	(1,940)	(1,208)	(4,127)	(3,204)

R = Reading Scores
M = Math Scores
NPR = National Percentile Ranking
NCE = Normal Curve Equivalent

TABLE 9

Achievement Test Scores, 1991–95

	Enrolled Choice		*MPS Low-Income*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1991						
% At or Above 50% of NPR	29.3	29.2	24.9	33.4	32.2	40.0
Median NPR	34	32	31	35	35	40
Mean NCE	42.6	40.2	39.1	42.0	42.2	45.2
Std. Dev. Of NCE	16.4	19.9	16.1	18.2	17.9	20.0
(N)	(177)	(185)	(1,433)	(1,419)	(1,697)	(1,957)
1992						
% At or Above 50% of NPR	21.2	22.8	25.2	33.5	32.4	39.5
Median NPR	27	28	29	35	34	40
Mean NCE	37.3	36.7	39.3	41.8	42.3	44.6
Std. Dev. Of NCE	16.2	18.0	17.2	19.1	18.8	20.5
(N)	(349)	(369)	(1,397)	(1,338)	(1,919)	(1,896)
1993						
% At or Above 50% of NPR	16.7	28.3	24.9	29.5	29.9	35.0
Median NPR	26	32	30	32	32	36
Mean NCE	37.2	42.2	38.8	39.9	40.9	42.9
Std. Dev. Of NCE	15.6	17.6	16.9	18.9	18.0	19.6
(N)	(398)	(395)	(1,212)	(1,189)	(1,443)	(1,370)
1994						
% At or Above 50% of NPR	28.8	31.3	25.7	42.4	30.5	48.4
Median NPR	30	38	30	39	32	47
Mean NCE	38.2	42.2	38.4	42.0	40.5	44.0
Std. Dev. Of NCE	15.4	18.8	17.0	19.1	18.1	20.2
(N)	(440)	(471)	(1,019)	(996)	(1,168)	(1,141)
1995[6]						
% At or Above 50% of NPR	14.6	26.1				
Median NPR	28.0	36.0				
Mean NCE	38.8	42.8				
Std. Dev. Of NCE	16.3	19.3				
(N)	(421)	(462)				

R = Reading Scores
M = Math Scores
NPR = National Percentile Ranking
NCE Normal Curve Equivalent

6. Only one MPS grade took the ITBS in 1995. Since it was a new version of the ITBS, comparable data do not exist for 1995.

TABLE 10A

Achievement Test Change Scores (NCEs), 1990 to 1991[7]

	Enrolled Choice		*MPS Low-Income*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1990 Mean NCE	40.0	39.2	37.5	39.5	39.5	41.6
Std. Dev.	13.9	19.6	15.2	17.8	16.6	18.7
1991 Mean NCE	41.8	39.1	38.2	42.2	40.5	44.2
Std. Dev.	14.4	15.3	15.3	17.7	16.6	18.6
Mean Difference	+1.8	-0.1	+0.7	+2.7	+1.0	+2.6
Std. Dev.	13.1	16.0	14.7	14.7	14.3	14.4
Probability Mean Difference = 0	.193	.935	.144	.000	.022	.000
(N)	(84)	(88)	(812)	(792)	(1,048)	(1,029)

7. Only students who were tested in the Spring of 1990 and the Spring of 1991 are included. For Choice students only those who completed a full year in the Choice schools are included. Sample sizes are small relative to the total students in the Choice Program.

TABLE 10B

Achievement Test Change Scores (NCEs), 1991 to 1992

	Enrolled Choice		*MPS Low-Income*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1991 Mean NCE	39.8	39.0	38.0	41.5	40.0	43.4
Std. Dev.	17.0	18.6	15.1	17.3	16.6	18.4
1992 Mean NCE	35.9	38.4	38.4	41.3	40.5	43.1
Std. Dev.	14.4	15.3	15.3	17.7	16.6	18.6
Mean Difference	-3.9	-0.6	+0.4	-0.3	+0.5	-0.3
Std. Dev.	16.0	16.3	14.6	15.3	14.2	14.6
Probability Mean Difference = 0	.001	.586	.484	.605	.286	.574
(N)	(192)	(198)	(911)	(895)	(1,173)	(1,148)

TABLE 10C

Achievement Test Change Scores (NCEs), 1992 to 1993

	Enrolled Choice		*MPS Low-Income*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1992 Mean NCE	38.7	38.3	39.5	42.2	40.8	43.2
Std. Dev.	16.2	19.0	17.3	19.0	17.9	19.4
1993 Mean NCE	38.3	42.7	38.8	41.0	40.1	42.0
Std. Dev.	14.2	17.2	16.7	18.1	17.4	18.7
Mean Difference	-0.4	+4.4	-0.7	-1.2	-0.8	-1.2
Std. Dev.	14.6	16.8	14.0	15.5	14.0	15.1
Probability Mean Difference = 0	.928	.000	.131	.002	.091	.019
(N)	(282)	(288)	(873)	(842)	(973)	(938)

TABLE 10D

Achievement Test Change Scores (NCEs), 1993 to 1994

	Enrolled Choice		*MPS Low-Income*		*MPS Control*	
	R	*M*	*R*	*M*	*R*	*M*
1993 Mean NCE	37.6	44.0	28.8	41.8	40.1	42.
Std. Dev.	15.7	17.1	16.7	18.58	17.6	19.4
1994 Mean NCE	37.5	42.0	35.6	41.5	39.9	42.7
Std. Dev.	15.8	17.8	16.9	19.1	18.6	19.8
Mean Difference	-0.1	-2.0	-0.2	-0.3	-0.2	-0.1
Std. Dev.	14.7	14.7	14.0	16.5	14.0	16.3
Probable Mean Difference = 0	.928	.002	.564	.631	.620	.904
(N)	(289)	(281)	(688)	(678)	(766)	(755)

TABLE 11A

Estimated Reading Iowa Test of Basic Skills, 1990–94, with Choice Indicator and Yearly Choice Variables

	Model 1 *Choice Indicator*		***Model 2*** *Choice Years*	
Variable	*B*	*SE B*	*B*	*SE B*
Constant	18.131***	.971	18.123***	.971
Prior Reading	.495***	.014	.495***	.014
Prior Math	.170***	.012	.170***	.013
Test Grade	-.176	.091	-.175	.091
Gender (1 = female)	1.559***	.366	1.562***	.366
African American	-3.911***	.546	-3.915***	.546
Hispanic	-1.599*	.725	-1.617*	.726
Other Minority	-2.435*	1.047	-2.439*	1.047
Low Income	-2.510***	.577	-2.506***	.577
Choice Indicator	-.568	.484	—	—
Choice Year 1	—	—	.100	.818
Choice Year 2	—	—	-1.200	.734
Choice Year 3	—	—	-.093	.875
Choice Year 4	—	—	-1.394	1.411
Adj. R^2	.46		.46	
F Statistic	446.32		441.23	
Probability F = 0	.000		.000	
Dependent Mean	40.09		40.09	
Dependent St. Dev.	16.98		16.98	
(N)	(4,716)		(4,716)	

*** probability that B=0 < .001
** probability that B=0 < .01
* probability that B=0 < .05

TABLE 10E

Achievement Test Change Scores (NCEs), 1994 to 1995[8]

	Enrolled Choice	
	R	*M*
1994 Mean NCE	38.4	43.1
Std. Dev.	15.1	18.8
1995 Mean NCE	38.6	41.7
Std. Dev.	16.0	17.6
Mean Difference	0.2	-1.5
Std. Dev.	14.8	15.8
Probable Mean Difference = 0	.783	.102
(N)	(285)	(306)

8. See footnote 6.

TABLE 11B

Estimated Math Iowa Test of Basic Skills, 1990–94, with Choice Indicator and Yearly Choice Variables

	Model 1 *Choice Indicator*		***Model 2*** *Choice Years*	
Variable	*B*	*SE B*	*B*	*SE B*
Constant	22.260***	1.060	22.249***	1.061
Prior Math	.546***	.014	.546***	.014
Prior Reading	.170***	.015	.170***	.015
Test Grade	-1.089***	.100	-1.086***	.100
Gender (1 = female)	-.065	.400	-.065	.400
African American	-4.182***	.599	-4.182***	.599
Hispanic	-1.546	.795	-1.568*	.796
Other Minority	.288	1.151	.284	1.152
Low Income	-1.431	.633	-1.429*	.633
Choice Indicator	-.109	.527	—	—
Choice Year 1	—	—	-.682	.894
Choice Year 2	—	—	.059	.801
Choice Year 3	—	—	.564	.946
Choice Year 4	—	—	-.823	1.541
Adj. R^2	.46		.49	
F Statistic	498.60		487.23	
Probability F = 0	.000		.000	
Dependent Mean	42.48		42.48	
Dependent St. Dev.	18.83		18.83	
(N)	(4,653)		(4,653)	

*** probability that B=0 < .001
** probability that B=0 < .01
* probability that B=0 < .05

TABLE 12A

Estimated Reading Iowa Test of Basic Skills, 1990–94, With Choice Indicator, Yearly Choice, and Survey Variables

	Model 1 *Choice Indicator*		***Model 2*** *Choice Years*	
Variable	*B*	*SE B*	*B*	*SE B*
Constant	14.947***	2.168	14.968***	2.168
Prior Reading	.433***	.025	.436***	.025
Prior Math	.177***	.021	.174***	.021
Test Grade	-.020	.167	-.036	.168
Gender (1 = female)	2.508***	.657	2.517***	.657
African American	-3.737***	.930	-3.762***	.929
Hispanic	-1.144	1.173	-1.029	1.174
Other Minority	-3.216	2.343	-3.194	2.340
Income ($1,000)	.092**	.032	.092**	.032
Mother's Education	.491**	.260	.499*	.260
Married (1-=yes)	-.271	.792	-.304	.792
PI-Schl. Cont.	-.115	.090	-.112	.090
PI-Par. Cont.	.081	.154	.080	.154
PI-Schl. Organ.	.477	.267	.469	.267
PI-Child	-.093	.097	-.094	.097
Edu. Expectations	.167	.349	.180	.349
Choice Indicator	-2.154**	.749	—	—
Choice Year 1	—	—	-1.382	1.092
Choice Year 2	—	—	-3.849***	1.024
Choice Year 3	—	—	-1.256	1.225
Choice Year 4	—	—	-.410	1.852
Adj. R^2	.50		.50	
F Statistic	86.15		73.03	
Probability F = 0	.000		.000	
Dependent Mean	42.95		42.95	
Dependent St. Dev.	16.47		16.47	
(N)	(1,385)		(1,385)	

*** probability that B=0 < .001

** probability that B=0 < .01

* probability that B=0 < .05

TABLE 12B

Estimated Math Iowa Test of Basic Skills, 1990–94, with Choice Indicator, Yearly Choice, and Survey Variables

	Model 1 *Choice Indicator*		***Model 2*** *Choice Years*	
Variable	*B*	*SE B*	*B*	*SE B*
Constant	22.558***	2.501	22.574***	2.509
Prior Math	.535***	.024	.533***	.024
Prior Reading	.150***	.028	.151***	.029
Test Grade	-1.285***	.192	-1.296***	.194
Gender (1 = female)	1.486*	.755	1.488*	.756
African American	-5.112***	1.081	-5.13***	1.083
Hispanic	-2.449	1.357	-2.389	1.361
Other Minority	-3.384	2.691	-3.373	2.693
Income ($1,000)	-.112**	.037	.112**	.037
Mother's Education	.007	.301	.012	.302
Married (1 = yes)	-.520	.911	-.532	.913
PI-Schl. Cont.	-.217*	.103	-.215*	.103
PI-Par. Cont.	.304	.179	.303	.179
PI-Schl. Organ.	-.202	.309	-.204	.310
PI-Child	-.264*	.112	-.265*	.112
Edu. Expectations	.740	.403	.749	.404
Choice Indicator	.093	.864	—	—
Choice Year 1	—	—	.282	1.261
Choice Year 2	—	—	-.464	1.187
Choice Year 3	—	—	.190	1.410
Choice Year 4	—	—	1.297	2.131
Adj. R^2	.53		.53	
F Statistic	97.02		81.60	
Probability F = 0	.000		.000	
Dependent Mean	45.37		45.37	
Dependent St. Dev.	19.53		19.53	
(N)	(1,372)		(1,372)	

*** probability that B=0 < .001
** probability that B=0 < .01
* probability that B=0 < .05

TABLE 13

Why Choice Students Left the Choice Program, 1991–1994[9]

Responses	*(N)*	*%*
Quality of Program	68	30.5
Lack of religious training	13	5.8
Lack of transportation	17	7.6
Income	9	4.0
Application problems	14	6.3
Fee changes	13	5.8
Selection problems	2	0.9
Quality of the Choice School	96	43.0
Poor education	20	9.0
Too disciplinarian	13	5.8
Unhappy with staff	31	13.9
Lack of programs for special needs students	14	6.3
Lack of programs for talented students	2	0.9
Too segregated	1	0.4
Child terminated	3	1.3
Lack of teaching materials	7	3.1
Child/Family Specific	43	19.3
Transportation – too far away	12	5.4
Moved	15	6.7
Pregnancy	3	1.3
Quit school	4	1.8
Child custody change	2	0.9
Miscellaneous	9	4.0
Total	223	

Question: "What were the major reasons your child did not continue in last year's school?" (Respondents could give up to three answers)

[9] Response rate for years 1992 to 1994 was 38%.

TABLE 14

Parents Participation In Educational Activities, 1991, Including Parents of Students No Longer In Choice Schools (Percentages)

	Times/Week				
	0	*1–2*	*3–4*	*5 or more*	*(N)*
Help with child's homework	3	17	31	48	(137)
Read with or to your child	9	25	29	37	(137)
Work on arithmetic or math	8	26	27	39	(135)
Work on penmanship or writing	16	31	24	30	(135)
Watch educational program on T.V. with your child	15	46	24	15	(136)
Participate together in sports activities	23	43	15	18	(136)

Question: "How many times in a normal week did you participate in the following activities with your child?"

Notes

1. Daniel McGroarty, "School Choice Slandered," *The Public Interest* (No. 17), Fall, 1994; Paul E. Peterson, "A Critique of the Witte Evaluation of Milwaukee's School Choice Program," Center for American Political Studies, Harvard University, February 1995, unpublished occasional paper.
2. This change is extremely important because most students have been admitted to the Choice Program in those grades. Private schools in general prefer to limit lateral entry at higher grades and therefore most have a grade structure which has many more students in the lower grades than the upper grades.
3. Table 6 contains information on "scales" which are sets of questions measuring an underlying concept. The exact questions for the scales appear in the *Fourth Year Report* (Tables A1–A6). We created simple scale scores by adding together responses to each item. Table 6 contains statistics on the scales, defines the scale direction, and reports the Cronbach Alpha statistic for the scale. Cronbach Alpha is a measure of how well the items form a scale. It has a maximum value of 1.0 when all items are perfectly correlated with each other.
4. Test scores were not available for all students in either group because tests were not given every year in MPS. Therefore, there were no tests for 4- and 5-year old kindergarten, and few for first-grade students. Lateral entry into higher grades could also have missed some students because primary testing was in grades 2, 5, 7, and 10. For the few high school students in the Choice Program, the 10th grade test was excluded because very few of these students were tested and because students were entering alternative schools (schools for students contemplating dropping out of school or pregnant teenage students).
5. A number of the tests taken in MPS were dictated by rules for the federal Chapter I program which required testing in every grade in reading and math using a standardized achievement test. In 1993, the federal regulations changed from requiring total math, consisting of three subtests, to just "problem solving." With that change, MPS dropped Chapter I testing using all three subtests for some students. Fortunately, the correlation between the Problem Solving Component and the Total Math score is .88. We were able to use an estimated regression model with Problem Solving estimating Total Math for students taking just the Problem Solving portion. The details of this procedure were described in Technical Appendix F of the *Fourth Year Report.*
6. Because sample sizes were relatively small for choice students, the most reliable statistic in these tables is the mean Normal Curve Equivalent (NCE). Median and percent at or above 50% National Percentile Rankings are included because these statistics are routinely reported by MPS. Because a number of students may be bunched together with small samples, both of these numbers are volatile.
7. There is almost no dropping out at the elementary level. Drop-out rates are also extremely low in middle schools. In MPS suspensions are also rare in these grades and the policies and reporting vary considerably from school to school. For example, student fighting, which leads to a suspension for up to three days in most of the private schools, may result in a student being sent home in MPS. Whether that becomes an official suspension or not may depend on the principal and the reactions of the child or parents. The numbers of official expulsions are even smaller than dropouts or suspensions. See John F. Witte, "Metropolitan Milwaukee Dropout Report," Report of the Commission on the Quality and Equity of the Milwaukee Metropolitan Public Schools, 1985.
8. Please note that the cohort population described in the last section is not identical with students for whom we have change data from one year to the next. Thus tables 9 and 10 are not directly comparable.
9. Normal Curve Equivalents are used because National Percentile Rankings are not interval level data. One of the problems with the transformation from NPRs to NCEs is that the very lowest and highest ends of the distribution are compressed. This tends to inflate very low-end scores and deflate very high-end scores. The lower end inflation may affect this population, which has quite a few test scores below the 10th National Percentile. For later analysis, if sample sizes become large enough, we will also analyze scores by grade using the Iowa Test Standard Scores, which are interval level but do not have this compression effect. NCEs are, however, the national standard for reporting results across populations and grades for Chapter I and other programs.
10. All MPS results in this and last year's report use the estimated math score. In the third year we reported the estimated scores in footnotes and appendices. We changed because we feel the more accurate method is to include the estimated scores to prevent an income bias associated with Chapter I eligibility.
11. These regressions are the equivalent of a change score with prior tests weighted proportionate to their B values (rather than constrained to 1.0). This can be determined by simply rewriting the regression equations subtracting the prior tests from each side of the equation.

12. See John F. Witte, "A Reply to Paul Peterson," unpublished paper, February 1995.
13. A final set of regressions, correcting for selection into and out of the choice and MPS samples, is currently being done. Those corrections require quite robust statistical assumptions and will be interpreted with caution. However, preliminary results indicate that taking into account primarily students not taking tests in a given year (because they left the respective systems or were not tested if present), both MPS and choice scores are corrected upward, but MPS scores are corrected upward considerably more. The reason is that non-Chapter I, MPS students were not as likely to be tested each year and White, non-low-income students were more likely to leave the system.
14. The McGroarty article may be the most notable for this. After spending pages demonstrating how choice opponents use our reports to support their positions, and criticizing test scores and attrition along the lines indicated below, he ends with an unqualified litany of the positive results we "demonstrated" (positive parental attitudes, parental involvement, etc.). See McGroarty, 1994.
15. The one exception to this is the income variable. Choice families have lower income than MPS families and we assume income is related to higher achievement. That variable, however, was already in the earlier equations as an indicator variable defined as qualifying for free lunch. the survey variable may add accuracy, but essentially that control was already in place.
16. Regression to the mean, taught in first semester statistics courses, occurs because of measurement error in tests. The easiest way to think about this is to consider that a person's true score varies randomly around the score the person actually records. For the person near the bottom of the distribution of test scores, the distribution of true scores is limited by not being able to go below 0, and thus the probability is that they will go up, not down. The same is true for people near the top of the distribution, because they cannot go above 100. If they are at 98 on the first test, the likelihood is that they will go down on the second try. Thus those near the bottom are likely to move upward, toward the mean, and those at the top the reverse.

CHAPTER 21

School Choice in Milwaukee

A Randomized Experiment

Jay P. Greene, Paul E. Peterson, and Jiangtao Du

School choice or voucher plans in which parents could use public funds to select the public or private schools that their children would attend have been receiving serious consideration as a means of improving the quality and efficiency of educational services. Although there are theoretical reasons to believe that choice and competition in schooling might be beneficial, evidence from a randomized experiment has not been available to substantiate or refute those theories. The evidence presented in this chapter from the school choice program in Milwaukee provides an opportunity to learn more about the effects of voucher programs from experimental data.

The Milwaukee experiment is unique in that it is the first publicly funded voucher program in the country and the only one with several years of results. The Milwaukee experiment is also unique in that vouchers were assigned by lottery when classes were oversubscribed. Analysis of data from a randomized experiment avoids the selection bias of comparing choosers to nonchoosers that has plagued other studies of choice in education. The Milwaukee experiment is of further interest in that it offers a hard test of choice theories because of the numerous constraints under which the program has operated.

Scholars have suggested that privatization may enhance efficiency in education in three different ways. First, competition among providers may reduce the cost and improve the quality of services.[1] Second, government-financed services may more closely match consumer preferences if the latter are given opportunities to sort themselves among an array of options.[2] Third, private producers may more easily enlist the participation of consumers in the co-production of the services, thereby enhancing service quality and effectiveness.[3]

If school choice could significantly improve the quality of education, the political and social benefits would be more than trivial. Apart from cash-transfer services, education is the largest part of the gross national product (GNP) of any publicly provided service.[4] In 1990 the cost of publicly financed education services constituted $305.6 billion, or 5.6 percent of the GNP.[5] Yet public confidence in public schools remains very low. In 1993 only

19 percent of the population was willing to give schools a grade of A or B, a fall of 8 percentage points since a decade earlier.[6]

Weak confidence in our public schools may be due to their failure to keep pace with rising public expectations. Estimated real costs within the educational sector, adjusted for inflation, rose by 29 percent or at an annual rate of 1.5 percent between 1974 and 1991.[7] Meanwhile, students' performance as measured by test scores, an important educational outcome, remained fairly constant.[8] Between 1970 and 1992 elementary and secondary students averaged no more than a gain of .1 of a standard deviation in mathematics and reading on the National Assessment of Educational Progress, generally thought to be the best available measure of student achievement. Meanwhile, their scores in science fell by .24 of standard deviation.[9] Increasing costs with at best slight gains in student achievement suggest that the public school system has become less efficient.

Opportunities for efficiency gains are particularly large in central cities. Whereas competition among small school districts exists in suburban parts of many metropolitan areas,[10] most city schools are governed by a single school board that does not ordinarily allow schools to compete for students.[11] Schools in rural areas often function as community institutions, facilitating co-production of educational services, but city schools have more limited ties to their immediate neighborhoods. Perhaps for these reasons, educational outcomes in the city lag those outside the central city.[12]

It has been argued that any efficiency gains are unlikely to result in higher levels of student achievement, because cognitive skills are either inherited or set in place at an early age, making them hardly susceptible to manipulation by educational processes.[13] But the weight of the evidence is in the opposite direction; numerous studies have found that school characteristics affect student achievement.[14] If these findings are correct, it may be hypothesized that if government grants are made available to families so they can purchase educational services for their children, efficiency gains accompanying privatization will result in enhanced student achievement.[15] Under such arrangements, competition among producers increases. Inasmuch as consumers' educational preferences vary and entry into the educational market is not prohibitively large, many producers will attempt to meet a demand for a range and variety of services. Co-production by consumers and providers (families and the schools) is more likely if families have a choice of schools.[16]

Yet efficiency gains that facilitate academic achievement may not be as great as these considerations suggest. Consumers may not have the information necessary to discern schools' academic quality.[17] Or consumers may choose schools on the basis of the schools' nonacademic characteristics, such as proximity, religiosity, sports facilities, or social segregation.[18]

Potential gains in student achievement as a result of privatization are much disputed, in part because empirical research has left the issue unresolved. Although two different research traditions have sought to estimate the comparative efficiency of private and public schools, neither has provided a definitive answer. The first research tradition has relied on data from national samples (High School and Beyond, the National Longitudinal Study of Youth, and the National Education Longitudinal Study) to estimate the achievement effects of attending public and private schools. Most of these studies have found that students who attend private schools score higher on achievement tests or are more likely to attend college than those who attend public schools.[19]

Because private schools are generally less expensive than public schools, these studies

suggest greater efficiency in the private sector. But these findings may be contaminated by selection bias: Students in private schools, who come from tuition-paying families, may have unobserved characteristics that increase the likelihood of their scoring higher on achievement tests, regardless of the schools they attend.[20]

The second research tradition consists of studies that evaluate the test performance of students from low-income or at-risk backgrounds who have received scholarships that give them the opportunity to move from public to private schools.[21] Although these evaluations also have reported that private schools produce higher levels of student achievement with less expenditure per pupil, their findings may also be contaminated by unobserved background characteristics of scholarship recipients. In almost all the programs studied, scholarships have been distributed on a first-come, first-served basis. They also require additional tuition payments by families, increasing the likelihood that scholarship recipients have unobserved characteristics (such as motivation) correlated with higher test scores.

A previous evaluation of the Milwaukee choice program reports no systematic achievement effects of enrollment in private schools.[22] But this evaluation compared students from low-income families with public school students from more advantaged backgrounds, leaving open the possibility that unobserved background characteristics could account for the lack of positive findings.[23] In sum, with the exception of the Milwaukee evaluation, most studies have found efficiency gains from the privatization of educational services. Yet all studies have suffered from potential selection bias, because they have relied on nonexperimental data that have included unobserved but possibly relevant background characteristics that could account for reported findings.

One way to improve on previous research is to conduct an experiment that avoids selection bias by randomly assigning students to treatment and control groups. With random assignment the members of the two groups can be assumed to be similar, on average, in all respects other than the treatment they receive. Differences in average outcomes can be reasonably attributed to the experimental condition. Only a few studies of school effectiveness have been able to draw upon data from randomized experiments, probably because it is difficult to justify random denial of access to apparently desirable educational conditions.[24] The results from the Milwaukee choice program reported here are, to the best of our knowledge, the first to estimate from a randomized experiment the comparative achievement effects of public and private schools.

Some results from the randomized experiment in Milwaukee were reported by Witte and associates in 1994,[25] but that study concentrated on a comparison of students in choice schools with a cross-section of students attending public schools. Data from the randomized experiment were underanalyzed and discussed only in passing.[26] In addition to our initial report,[27] two other studies have reported results from the randomized experiment in Milwaukee,[28] but all three studies relied on inaccurate test score data.

Subsequent to issuing our report in 1996, we discovered that the Milwaukee test score data available on the world wide web did not adjust for the fact that some students were not promoted from one grade to the next. For example, students in both test and control groups who were held back for a year at the end of third grade were scored as third graders when they otherwise would have been scored as fourth graders. When this happens, a student can receive a much higher percentile score than is appropriate. Other students are allowed to skip a grade, and if this promotion is not taken into account, it produces

an error of the opposite kind. We were able to eliminate both types of error by adjusting test scores to the correct grade level by means of conversion tables.[29]

A Hard Case

The Milwaukee choice program, initiated in 1990, provided vouchers to a limited number of students from low-income families to be used to pay tuition at their choice of secular private schools in Milwaukee. The program was a hard case for testing the hypothesis that efficiency gains can be achieved through privatization, because it allowed only a very limited amount of competition among producers and choice among consumers.[30]

The number of producers was restricted by the requirement that no more than half of a school's enrollment could receive vouchers. Because this rule discouraged the formation of new schools, no new elementary school came into being in response to the establishment of the voucher program. Consumer choice was further limited by excluding the participation of religious schools (thereby precluding use of approximately 90 percent of the private school capacity within the city of Milwaukee). Co-production was also discouraged by prohibiting families from supplementing the vouchers with tuition payments of their own. (But schools did ask families to pay school fees and make voluntary contributions.) Other restrictions also limited program size. Only 1 percent of the Milwaukee public schools could participate, and students could not receive a voucher unless they had been attending public schools or were not of school age at the time of application.

These restrictions significantly limited the amount of school choice that was made available. Most choice students attended fiscally constrained institutions with limited facilities and poorly paid teachers.[31] One school, Juanita Virgil Academy, closed a few months after the program began.[32] Although the school had existed as a private school for a number of years, it was eager to admit sixty-three choice students in order to alleviate its enrollment and financial difficulties. Even with the addition of the choice students, the school's problems persisted. To comply with the requirement that schools offer secular curricula, the school had to drop its Bible classes. Parents complained about the food service, overcrowded classrooms, a shortage of books and materials, and a lack of cleanliness and discipline. The executive director had hired a new principal away from the public schools, but she had to be relieved of her responsibilities two months into the school year. The school withdrew from the choice program the next semester, giving as its reason the desire to "reinstate religious training in the school." A few weeks later the school closed altogether.[33]

Given the design of the Milwaukee choice program, more school failures might have been expected. The three schools that together with Juanita Virgil Academy admitted 84 percent of the choice students in 1990 had modest facilities and low teacher salaries. Bruce Guadalupe Community School was in particular difficulty. Established in 1969, it sought to preserve Latino culture and teach children respect for both English and Spanish. Many teachers had once taught in Central American schools. Instruction was bilingual, often more in Spanish than English. Despite its distinctive educational mission, the school had difficulty making ends meet. Even finding an adequate school building seemed a never-ending problem; the school moved from one location to another on several occasions during its first two decades. By January 1990 things had become so desperate that the school was on "the verge of closing." But enactment of the choice program gave the school "new hope for the future," a hope that "otherwise had been snuffed out."[34] A tuition voucher of more than $2,500 per student was a boon to a

school that had had trouble collecting $650 from each participating family.

Despite the arrival of choice students in the fall of 1990, the school, still in financial distress, was forced to cut its teaching staff by a third. The school's difficulties were fully reported in the *Milwaukee Journal:* "Two staff aides were fired, the seventh and eighth grades were combined, the second grade was eliminated with children put into the first or third grade, and the bilingual Spanish program was cut. . . . Two teachers were transferred. . . . The former eighth grade teacher [was] teaching fourth grade. . . . Overall, the teaching staff was reduced from 14 to 9."[35] The school's principal described staff morale as "low."

The two other community schools with large choice enrollments, Harambee Community School and Urban Day School, had better reputations, but still suffered from serious financial difficulties.[36] Like Bruce Guadalupe, they catered almost exclusively to a low-income minority population. Established in the 1960s in former Catholic parish schools, they tried to survive as secular institutions after the archdiocese closed the parochial schools. Named for the Swahili word meaning "pulling together," Harambee presented itself as "an African American-owned school emphasizing the basics through creative instructional programs, coupled with a strong cultural foundation."[37] Urban Day was said to place "a heavy emphasis on African history and culture."[38]

Like Bruce Guadalupe, these schools could ask families to pay only a very modest tuition. Though they set their annual rates at somewhere between $650 and $800, only a few families whose children were attending the schools actually paid full tuition. Tuition scholarships were the norm, not an exceptional privilege. But parents were expected to participate in fund-raising activities. Teacher salaries were much lower than those paid by the Milwaukee public schools. As one principal observed, "The teachers who stay here for a long time are either very dedicated or can afford to stay on what we pay."[39]

The quality of the physical plant provided a visible sign of the school's modest financial resources: "Recess and physical education facilities were relatively poor in the schools. One school had easy access to a city park for recess, one relied on a blocked off street, two others asphalt playgrounds with some wood chips and playground equipment. All the schools had some indoor space for physical education, but it often served multiple purposes."[40] One of its hardest-working supporters was asked what she would most wish for the school. She said, "I'd like to see the school financially self-sufficient."[41]

To repeat, the Milwaukee choice program is a hard case to test the hypothesis that privatization can result in efficiency gains. If one finds efficiency gains under considerably less than ideal circumstances, one is likely to find gains under more opportune conditions.

School Costs

The relative costs of the public and private schools in Milwaukee remained approximately the same throughout the four years of the experiment. In the 1991–92 school year payments per pupil to schools participating in the choice program schools were $2,729. Based on interviews with school administrators, it is estimated that schools received an additional $500 per student through fees and fund-raising activities. Therefore, the total costs per pupil are estimated to have been $3,229. Per-pupil costs for the Milwaukee public schools at this time averaged $6,656, somewhat higher than the $5,748 cost of educating the average public school student in the United States as a whole.[42]

Although it appears that the cost of educating a pupil in a choice school was only 48 percent of the cost of educating a student in

the Milwaukee public schools, the actual difference was not this large. Choice school students were provided transportation by the Milwaukee public school system if they needed it. In addition, the reported per-pupil expenditures for the Milwaukee public schools included the costs of educating secondary school students (which may be more expensive than elementary education) as well as students receiving special services. But even after taking these considerations into account, the per-pupil costs of the private schools were lower.

The Milwaukee Randomized Experiment

The Milwaukee school choice program was a randomized experiment. To ensure equal access to the choice program among eligible applicants, the legislature required choice schools, if oversubscribed, to admit applicants at random. In the words of the original evaluation team, "Students not selected into the Choice Program in the random selection process represent a unique research opportunity. … If there are any unmeasured characteristics of families seeking private education, they should on average be similar between those in and not in the program."[43] The legislature asked the state's Department of Public Instruction to evaluate the Milwaukee choice experiment. Data were collected on family background characteristics and student performance on the Iowa Test of Basic Skills in reading and mathematics. These data were made available on the world wide web in February 1996.

Students did not apply to the choice program as a whole; instead, they applied each year for a seat in a specific grade in a particular school. They were selected or not selected randomly by school and by grade. Because the random assignment policy was implemented in this way, in our analysis we used a fixed effects model that took into account the grade to which a student applied and the year of application.[44] Our analysis was unable to ascertain the particular school to which a student applied,[45] but it took this factor partially into account by adjusting for the ethnicity of the applicant. More than 80 percent of the choice students attended one of three schools, and of these three schools virtually all students applying to one school were Hispanic, and almost all students applying to the two others were African American. Though the analysis took the two predominantly African-American schools as a block, it otherwise distinguished among schools by adjusting for whether the applicant was Hispanic or African American. Because the number of white students and other minority students for whom information was available was so sparse that no reliable results could be obtained, these students were removed from the analysis.

By using a fixed effects model that took into account each point at which randomization occurred, together with a control for gender, it was possible to estimate the effects of enrollment on test scores in choice schools.[46] This procedure treated each point at which randomization occurred as a dummy variable. The measures of test score performance were the students' normal curve equivalent (NCE) scores for math and reading on the Iowa Test of Basic Skills. The NCE is a transformation of the national percentile rankings that arranges the scores around the fiftieth percentile in a manner that can be described by a normal curve. A standard deviation for NCE is 21 percentile points.

Separate ordinary least squares regressions produced an estimate of the effect of one, two, three, and four years of treatment on math and reading scores. The analysis of the Milwaukee randomized experiment conducted by Rouse constrained the effects of treatment to be linear in order to estimate the effect of each year in a single regression for

math and reading, respectively.[47] Because the effects of treatment do not seem to have been linear, our approach of estimating each amount of treatment separately avoided this source of potential bias.

Coefficients for each dummy representing the points of randomization and the constant are too cumbersome to present in a book of this nature, but are available from the authors upon request, as are the coefficients for all substantive variables employed in the models. We controlled for gender in every regression, because it was available for virtually all students and produced a more precise estimate of the effect of treatment.

Our data are limited by the fact that test data were available for only 78 percent of those assigned to the treatment group and 72 percent assigned to the control group. The percentage of test scores available decreased to 40 percent of the treatment group and 48 percent of the control group by the third or fourth year following application to the program (see table 1).

Our results depend on the assumption that the missing cases did not differ appreciably from those remaining in the sample.[48] One way of estimating whether this assumption is reasonable is to examine the observed characteristics of students in the treatment and control groups. As can be seen in table 2, the background characteristics of the two groups do not differ in important respects. In the words of the original evaluation team, "In terms of demographic characteristics, nonselected. . . students came from very similar homes as choice [students did]. They were also similar in terms of prior achievement scores and parental involvement.[49]

TABLE 1
Students for Whom Data Are Available

Student category	*Choice students*	*Control students*
Percent with test scores available (table 13-3, columns 1–4)	79	72
Total number who applied, 1990–93	908	363
Percent with test scores three or four years after application (table 13-3, column 5)	40	48
Total number who applied in 1990 or 1991, making it possible to have scores three or four years after application	592	166

Results

Using the analytical procedures discussed above, we estimated the effects of choice schools on students' performance after one, two, three, and four years of attending choice schools.[50] Table 3 reports the results of our main analysis, in which we estimated the difference in test scores between students attending choice schools and those in the control group after controlling for gender using a fixed effects model that takes into account the points of randomization in the experiment.

The estimated effects of choice schools on mathematics achievement were slight for the first two years students were in the program. But after three years of enrollment students scored 5 percentile points higher than the control group; after four years they scored 10.7 points higher. These differences between the two groups three and four years after their application to choice schools are .24 and .51 standard deviation of the national distribution of math test scores, respectively. They are statistically significant at accepted confidence levels.[51] Differences on the reading test were between 2 and 3 percentile points for the first three years and increased to 5.8 percentile points in the fourth. The results for the third and fourth years are statistically significant when the two are jointly estimated.[52]

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

	All students in the study			*All students with scores three or four years after application*		
Characteristic	*Choice students*	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Controlling for Family Background

The results in the main analysis in table 3 provide the best estimate of the achievement effects of attendance in private schools, because this analysis had the fewest missing cases. But because these results do not take into account family background characteristics, they depend on the assumption that students were assigned at random to the test and control groups. Inasmuch as even the main analysis had many missing cases, it was possible that the two groups were no longer similar in relevant respects, despite their similar demographics (see table 2). To explore whether this possibility contaminated our results, we performed a fixed effects analysis that took into account gender, mother's education, and parents' marital status, income, education expectations, and time spent with the child. Table 4 reports the results.

This analysis depended on information provided in response to a written questionnaire which, unfortunately, many parents did not complete. Background information was available for only 47 percent of the selected students and 36 percent of the control group. The number of cases available for analysis was therefore considerably reduced, and the point estimates are less reliable. Nevertheless, all point estimates are positive, and six of the eight are actually larger than those reported in the main analysis.

TABLE 3

The Effect of Attending a Choice School on Test Scores, Controlling for Gender, Using Fixed Effects Model
(NCE percentile points[a])

Effect and subject	*1 year of treatment*	*2 years of treatment*	*3 years of treatment*	*4 years of treatment*	*3 or 4 years jointly estimated*
Differences in mathematics scores between choice students and control group					
Effect on math scores	1.31	1.89	5.02**	10.65**	6.81**
Standard error	1.98	2.05	3.07	4.92	2.97
N	772	584	300	112	316
Differences in reading scores					
Effect on reading scores	2.22*	2.26	2.73	5.84*	4.85**
Standard error	1.74	1.78	2.63	4.22	2.57
N	734	604	301	112	318

a. Normal curve equivalent.
* = p < .10 in one-tailed t-test
** = p < .05 in one-tailed t-test

Controlling for Prior Test Scores

The main analysis did not control for students' test scores before entering into the choice program. It is not necessary to control for pre-experimental test scores when comparing a treatment group and a control group in an experimental situation, because the two groups, if randomly assigned to each category, can be assumed to be similar. But because of the sizable number of missing cases it is possible that the two groups we compared had different pretest scores before the experiment began. This potential source of bias did not appear, however. The average pretest scores at the time of application for the two groups were essentially the same. The average math and reading pretest scores for

TABLE 4

The Effect of Attending a Choice School on Test Scores, Controlling for Gender, Education Expectations, Income, Marital Status, Mother's Education, and Time Spent with Child, Using Fixed Effects Model
(NCE percentile points[a])

Effect and subject	*1 year of treatment*	*2 years of treatment*	*3 years of treatment*	*4 years of treatment*
Differences in mathematics scores between choice students and control group				
Effect on math scores	6.01**	5.36*	8.16*	7.97
Standard error	3.39	3.39	5.82	9.85
N	378	289	149	57
Differences in reading scores				
Effect on reading scores	4.72**	1.17	8.87**	15.00*
Standard error	2.88	2.99	5.27	9.45
N	358	293	150	55

a. Normal curve equivalent.
* = p < .10 in one tailed t-test
** = p < .05 in one-tailed t-test

those selected for the choice program were the NCE equivalents of 39 and 38 percentile rankings, respectively; for those not selected they were the NCE equivalents of a 39 percentile ranking for reading and a 40 percentile ranking for math (see table 2).

Inasmuch as the students' pretest scores at the time of application were essentially the same, it is unlikely that controls for this variable would alter the result. We nonetheless tested for the possibility, and the results are reported in table 5. Because pretest scores at the time of application were available for only 29 percent of the selected students and 49 percent of the control group, the sample size for this analysis is smaller and the results are generally not statistically significant. Yet five of the eight point estimates are larger than those in the main analysis, and all but one have positive signs.

Effects on All Students Accepted into the Choice Program

The results reported so far compare students who attended private schools with students who had applied for choice but were assigned to the control group. Some students, however, were accepted into the program but chose not to participate for the full four years. Some students immediately turned down the opportunity, but others left sometime during the four-year period.

To see the effect of the choice program on all those admitted, regardless of their subsequent enrollment decisions, we conducted an analysis identical to the main analysis, except that analysis compared all students initially assigned to treatment and control groups, regardless of the schools they chose to attend. This type of analysis is known in medical research as an intention-to-treat analysis. In many medical experiments subjects may be more or less faithful in complying with the treatment. For example, some forget to take their pills three times a day as instructed. An intention-to-treat analysis answers this question: Is the treatment effective even when compliance is less than 100 percent? Those who refused enrollment in the private schools or left before the end of the experiment can be thought of as not having complied with the treatment.

This approach had the important disadvantage of including in the treatment group many students who either did not attend the

TABLE 5

The Effects of Attending a Choice School on Test Scores, Controlling for Gender and Test Scores Before Application, Using Fixed Effects Model (NCE percentile points[a])

Effect and subject	*1 year of treatment*	*2 years of treatment*	*3 years of treatment*	*4 years of treatment*
Differences in mathematics scores between choice students and control group				
Effect on math scores	2.34	3.46*	7.40**	4.98
Standard error	2.32	2.71	4.08	9.16
N	286	185	83	31
Differences in reading scores				
Effect on reading scores	1.50	3.24*	5.28*	-3.29
Standard error	2.07	2.46	3.74	7.46
N	303	189	84	31

a. Normal curve equivalent

* = $p < .10$ in one tailed t-test

** = $p < .05$ in one-tailed t-test

TABLE 6

Percentage of Students in Intention-to-Treat Analysis for Whom Data Are Available

Student category	*Selected students*	*Control students*
Percent with test scores available (table 7, columns 1–4)	89	72
Total number who applied, 1990–93	908	363
Percent with test scores three of four years after application	63	48
Total number who applied in 1990 or 1991, making it possible to have scores three or hour years after application	592	166

private schools or attended the private schools for less than the full period under study. But it had two advantages. First, departure from an ideal randomized experiment was less in this case than in the main analysis. All cases were preserved except instances in which test data were not collected. The percentage of intention-to-treat cases in the analysis was 89 percent; sixty-three percent of the intention-to-treat cases three or four years after application remained in this analysis (see table 6).[53] (There were fewer missing cases because the students who left private schools but were tested in the Milwaukee public schools were not excluded from the intention-to-treat analysis.) Second, this analysis may have better captured what might happen if choice between public and private schools were generalized; students can be expected to migrate back and forth between the two systems.

Are there efficiency gains when comparisons are made between all those randomly assigned to the intention-to-treat group and the control group? The answer to this question is given in table 7. The effects do not differ in any significant way from those reported in the main analysis. Slight positive effects are found for the first three years after application to the program, and moderately large effects are found after four years. Students who were given a choice of schools performed better than did the control group, regardless of the public or private schools they attended.

TABLE 7

The Effect of Being Selected for a Choice School (Intention to Treat) on Test Scores, Controlling for Gender, Using Fixed Effects Model (NCE percentile points [a])

Effect and subject	*1 year of treatment*	*2 years of treatment*	*3 years of treatment*	*4 years of treatment*
Differences in mathematics scores between choice students and control group				
Effect on math scores	2.68*	2.59*	3.83*	11.00**
Standard error	1.89	1.94	2.87	4.14
N	854	728	435	175
Differences in reading scores				
Effect on reading scores	2.46*	2.57*	2.10	6.26**
Standard error	1.71	1.68	2.48	3.65
N	816	738	441	175

a. Normal curve equivalent

* = $p < .10$ in one tailed t-test

** = $p < .05$ in one-tailed t-test

All results but one are statistically significant at the .1 level; fourth-year results are significant at the .05 level.

These results suggest that some of the achievement effects produced by choice may be due to a closer match between school qualities and student needs. When families are given a choice between public and private schools, they may be choosing the options best suited to their children. It is possible that public schools induced some families with students in the treatment group to return to the public schools by providing them with better public school alternatives. The Milwaukee public school system had the ability to respond in this manner because it had a number of magnet schools. It also had the incentive to react, because the system could regain funds equivalent to the size of the voucher if a student returned to the public school system.

Conclusions

The Milwaukee choice experiment suggests that privatization in education may result in efficiency gains. This finding emerges from a randomized experiment less likely to suffer from selection bias than studies dependent on nonrandomized data. The consistency of the results is noteworthy. Positive results were found for all years and for all comparisons except one. The results reported in the main analysis for both math and reading are statistically significant for students remaining in the program for three to four years when these are jointly estimated.

These results after three and four years are moderately large, ranging from .1 of a standard deviation to as much as .5 of a standard deviation. Studies of educational effects interpret effects of .1 standard deviation as slight, effects of .2 and .3 standard deviation as moderate, and effects of .5 standard deviation as large.[54] Even effects of .1 standard deviation are potentially large if they accumulate over time.[55] The average difference in test performances of whites and minorities in the United States is one standard deviation.[56] If the results from Milwaukee can be generalized and extrapolated to twelve years, a large part of between-group reading differences and all of between-group math differences could be erased.

Without data beyond the Milwaukee program's first four years, one can only speculate as to whether such generalization and extrapolation are warranted. But if they are, the effectiveness of government-financed education could be greatly enhanced. These moderately large effects on student achievement were observed even though the Milwaukee plan offered students and families only a slightly enlarged set of educational choices. These achievement effects were produced at lower per-pupil cost than that of a Milwaukee public school education.

One must be cautious concerning the universe to which these results are generalized. Efficiency gains may be greater in Milwaukee and other central cities than in suburban areas where competition among school districts is greater. They may also be greater in cities than in rural communities where opportunities for co-production in public education may be more prevalent. The magnitude of the gains reported here may not be generalizable beyond central cities.

In addition, the study was limited to students from low-income families. Other studies suggest that private schools have a larger positive effect on the achievement of disadvantaged students.[57] Perhaps the results found in Milwaukee are restricted to low-income minority populations. Finally, the results are for families who applied for vouchers. It may be that the benefits of privatization are greater for those families who desire alternatives to the public schools serving them. Their children may have been particularly at risk in public

schools, and they may be more willing to engage in coproduction than all other families.

The conclusions that can be drawn from our study are further restricted by limitations of the data made available on the world wide web. Many cases are missing from this data set. The percentage of missing cases is especially large when one introduces controls for background characteristics and preexperimental test scores. But given the consistency and magnitude of the findings as well as their compelling policy implications, they suggest the desirability of further randomized experiments capable of reaching more precise estimates of efficiency gains through privatization.

Randomized experiment are under way in New York City, Dayton, and Washington, D.C.[58] If the evaluations of these randomized experiments minimize the number of missing cases and collect preexperimental data for all subjects in both treatment and control groups, they could, in a few years' time, provide more precise estimates of potential efficiency gains from privatizing the delivery of educational services to low-income students. Similar experiments should be conducted in a variety of contexts, but especially in large central cities, where potential efficiency gains seem particularly likely.

Notes

1. Kenneth Arrow, "An Extension of the Classical Theorems of Welfare Economics," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics* (Berkeley: University of California Press, 1951); Jacob Schmookler, *Invention and Economic Growth* (Harvard University Press, 1966); and James Dearden, Barry W. Ickes, and Larry Samuelson, "To Innovate or Not to Innovate: Incentives and Innovation in Hierarchies," *American Economic Review* (December 1990), pp. 1105–06.
2. Charles M. Tiebout, "A Pure Theory of Local Expenditures," *Journal of Political Economy* 64 (October 1956), pp. 416–24; Robert L. Bish, *The Public Economy of Metropolitan Areas* (Chicago: Markham, 1971).
3. Elinor Parks, Roger B. Parks, and Gordon P. Whitaker, *Patterns of Metropolitan Policing* (Cambridge, Mass.: Ballinger Publishing, 1978).
4. Cash-transfer programs are larger but do not involve direct service provision. Although publicly funded medical services are more costly than publicly-funded schools, most medical services are provided by private vendors. In recent years the cost of defense has fallen below the cost of state-provided educational services.
5. Paul E. Peterson, *The Price of Federalism* (Brookings, 1995).
6. *Public Perspective,* Roper Center Review of Public Opinion and Polling (November-December 1993), p. 13.
7. Some of these increased school costs are due to improved services for students with disabilities and those who are otherwise disadvantaged.
8. Helen F. Ladd, ed., "Introduction," in Ladd, ed., *Holding Schools Accountable: Performance-based Reform in Education* (Brookings, 1996), p. 3; Richard Rothstein and Karen Hawley Miles, *Where's the Money Gone? Changes in the Level and Composition of Education Spending* (Washington: Economic Policy Institute, 1995), p. 7.
9. Larry V. Hedges and Rob Greenwald, "Have Times Changed? The Relation between School Resources and Student Performance," in Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (Brookings, 1996), pp. 74–92; information cited on p. 78.
10. Caroline Minter Hoxby, "When Parents Can Choose, What Do They Choose? The Effects of School Choice on Curriculum," in Susan Mayer and Paul E. Peterson, eds., *When Schools Make a Difference* (forthcoming).
11. Paul E. Peterson, "Monopoly and Competition in American Education," in William H. Clune and John F. Witte, eds., *Choice and Control in American Education,* vol. 1 (New York: Falmer, 1990), pp. 47–78.
12. Pam Belluck, "Learning Gap Tied to Time in the System: As School Stay Grows, Scores on Tests Worsen," *New York Times,* January 5, 1997, p. B1; George A. Mitchell, *The Milwaukee Parental Choice Plan* (Milwaukee: Wisconsin Policy Research Institute, 1992); and Paul E. Peterson, "Are Big City Schools Holding Their Own?" in John Rury, ed., *Seeds of Crisis* (University of Wisconsin Press, 1993).

13. Richard J. Herrnstein and Charles Murray, *The Bell Curve: Intelligence and Class Structure in American Life* (Free Press, 1994).
14. David Card and Alan B. Krueger, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100 (February), pp. 1–40; Hedges and Greenwald, "Have Times Changed?"; Christopher Jencks and Meredith Phillips, "Does Learning Pay Off in the Job Market?" in Mayer and Peterson, eds., *When Schools Make a Difference;* Jay Girotto and Paul E. Peterson, "Do Hard Courses and Good Grades Enhance Cognitive Skills?" in Mayer and Peterson, eds., *When Schools Make a Difference;* Susan Mayer and David Knutson, "Early Education Versus More Education: Does the Timing of Education Matter?" in Mayer and Peterson, eds., *When Schools Make a Difference;* Robert Meyer, "Applied versus Traditional Mathematics: New Evidence on the Production of High School Mathmatics Skills," in Mayer and Peterson, eds., *When Schools Make a Difference.*
15. John E. Chubb and Terry M. Moe, *Politics, Markets, and American Schools* (Brookings, 1990).
16. Anthony Bryk, Valerie E. Lee, and Peter B. Holland, *Catholic Schools and the Common Good* (Harvard University Press, 1993).
17. Kevin B. Smith and Kenneth J. Meier, *The Case Against School Choice* (Armonk, N.Y.: M. E. Sharpe, 1995), p. 126; but also see Hoxby, "When Parents Can Choose, What Do They Choose?"
18. Richard F. Elmore, "Choice As an Instrument of Public Policy: Evidence from Education and Health Care," in Clune and Witte, eds., *Choice and Control in American Education,* vol. 1; Carnegie Foundation for the Advancement of Teaching, *School Choice* (Princeton, N.J., 1992); and Amy Gutmann, *Democratic Education* (Princeton University Press, 1987).
19. James S. Coleman, Thomas Hoffer, and Sally Kilgore, *High School Achievement* (New York: Basic Books, 1982); James S. Coleman and Thomas Hoffer, *Public and Private Schools: The Impact of Communities* (Basic Books, 1987); Chubb and Moe, *Politics, Markets, and American Schools;* Bryk, Lee, and Holland, *Catholic Schools and the Common Good;* William N. Evans and Robert M. Schwab, "Who Benefits from Private Education? Evidence from Quantile Regressions," Department of Economics, University of Maryland, 1993; and Christopher Jencks, "How Much Do High School Students Learn?" *Sociology of Education* 58 (April 1985), pp. 128–35. But for different results, see Douglas J. Wilms, "School Effectiveness within the Public and Private Sectors: An Evaluation," *Evaluation Review* 8 (1984), pp. 113–35; Douglas J. Wilms, "Catholic School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-up Study," *Sociology of Education* 58 (1985), pp. 98–114.
20. Arthur S. Goldberger and Glen G. Cain, "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report," *Sociology of Education* 55 (April-July 1982), pp. 103–22; Peter W. Cookson, Jr., *School Choice: The Struggle for the Soul of American Education* (Yale University Press, 1994); John F. Witte, "Understanding High School Achievement: After a Decade of Research, Do We Have Any Confident Policy Recommendations?" paper prepared for the 1990 annual meeting of the American Political Science Association; and John F. Witte, "Private School versus Public School Achievement: Are There Findings That Should Affect the Educational Choice Debate?" *Economics of Education Review* 11 (1992), pp. 371–94.
21. Terry M. Moe, ed., *Private Vouchers* (Stanford University Press, 1995); Valerie Martinez, Kenneth R. Godwin, and Frank R. Kemerer, "Private School Choice in San Antonio," in Moe, ed., *Private Vouchers;* Janet R. Beales and Maureen Wahl, *Given the Choice: A Study of the PAVE Program and School Choice in Milwaukee,* Policy Study 183, Reason Foundation, Los Angeles, January 1995.
22. John F. Witte, "First Year Report: Milwaukee Parental Choice Program," Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, November 1991; John F. Witte, Andrea B. Bailey, and Christopher A. Thorn, "Second Year Report: Milwaukee Parental Choice Program," Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, December 1992; John F. Witte, Andrea B. Bailey, and Christopher A. Thorn, "Third Year Report: Milwaukee Parental Choice Program," Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, December 1993; John F. Witte and others, "Fourth Year Report: Milwaukee Parental Choice Program," Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, December 1994; John F. Witte, Troy D. Sterr, and Christopher A. Thorn, "Fifth Year Report: Milwaukee Parental Choice Program," Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, December 1995; and John F. Witte, "Achievement Effects of the Milwaukee Voucher Program," paper presented at the 1997 annual meeting of the American Economics Association.
23. Jay P. Greene, Paul E. Peterson, and Jiangtao Du, with Leesa Berger and Curtis L. Frazier, "The Ef-

fectiveness of School Choice in Milwaukee: A Secondary Analysis of Data from the Program's Evaluation," Occasional Paper, Program in Education Policy and Governance, Harvard University, 1996.

24. But see the evaluation of the Tennessee randomized experiment in Frederick Moesteller, "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children* 5 (1995), pp. 113–27. This study found that class size has a positive affect on student achievement, contrary to many econometric studies. For the latter, see Eric Hanushek, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 24 (September 1986), pp. 1141–77.
25. Witte and others, "Fourth Year Reports."
26. Paul E. Peterson, "A Critique of the Witte Evaluation of Milwaukee's School Choice Program," Occasional Paper, Center for American Political Studies, Harvard University, February 1995.
27. Greene and others, "Effectiveness of School Choice."
28. Witte, "Achievement Effects of the Milwaukee Voucher Program"; Cecilia Rouse, "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," Department of Economics and the National Bureau of Economic Research, Princeton University, 1997.
29. A. N. Hieronymous, *Teacher's Guide Multilevel Battery Level 9–14: Iowa Test of Basic Skills Forms G/H* (Chicago: Riverside Publishing, 1986); H. D. Hoover and others, *Iowa Tests of Basic Skills: Norms and Score Conversions, Form K, Complete and Core Batteries, Levels 5–14* (Chicago: Riverside, 1993).
30. The Milwaukee choice program is described as it was in its initial years, because the data on student achievement are available for only the first four years. In subsequent years the program was expanded somewhat, but the important expansion in 1995 to include religious schools has yet to be implemented due to court challenges. For a fuller discussion of the program, see Paul E. Peterson, Jay P. Greene, and Chad Noyes, "School Choice in Milwaukee," *Public Interest* (Fall 1996), pp. 38–56; and Paul E. Peterson and Chad Noyes, "Under Extreme Duress, School Choice Success," in Diane Ravitch and Joseph Viteritti, eds., *New Schools for a New Century: The Redesign of Urban Education* (Yale University Press, 1997.)
31. The number of students attending each school was made available by the State Department of Public Instruction and reported in the *Milwaukee Journal* ("Court Time on Choice Extended," October 3, 1991) and a report by the Wisconsin Legislative Audit Bureau (*An Evaluation of Milwaukee Parental Choice Program*, February 1995, table 2, p. 22, table 3, p. 23). In addition to the schools discussed in the text, a Montessori school serving a middle-class constituency admitted three students the first year and four the next. Woodland School, formerly a laboratory school for a local Catholic college, enrolled between twenty and forty choice students each year. After the first year three other private elementary schools admitted a small number of students. Test performances of a small number of students attending the high schools participating in the program were not analyzed because no appropriate control group was available. These schools were initially established to serve at-risk students referred to them by the Milwaukee public schools.
32. The students attending this school are not included in the main analysis because they were not in a choice school at the end of the first year; nevertheless, they are included in the intention-to-treat analysis in table 7.
33. Witte, "First Year Report: Milwaukee Parental Choice Program."
34. Amy Stuart Wells, "Milwaukee Parents Get More Choice on Schools," *New York Times,* March 28, 1990, p. B9.
35. Barbara Miner, "'Choice' School in Turmoil Because of Staff Cuts, Changes," *Milwaukee Journal,* November 23, 1990, p. B5.
36. Four of the original choice schools were said to be in "serious financial difficulty" and, in addition to Juanita Virgil, two more were said to be "on the verge of closing in the Spring of 1990" (Witte, Bailey, and Thorn, "Third Year Report: Milwaukee Parental Choice Program").
37. *Milwaukee Community Journal,* Special Supplement, May 4, 1994.
38. Paul Taylor, "Milwaukee's Controversial Private School Choice Plan Off to Shaky Start," *Washington Post,* May 23, 1991, p. A3.
39. Witte, "First Year Report: Milwaukee Parental Choice Program," p. 12.
40. Witte, "First Year Report: Milwaukee Parental Choice Program," p. 13.
41. *Milwaukee Community Journal,* Special Supplement.
42. U.S. Department of Education, *Digest of Education Statistics* (Washington, Office of Educational Research and Improvement, 1991), table 158.
43. Witte and others, "Fourth Year Report."
44. Siblings were exempt from the random assignment rule. We were unable to identify siblings from the information made available on the world wide web.
45. To protect the confidentiality of students, the data on the world wide web do not identify the schools they attended. To obtain this information we of-

fered to protect students' confidentiality, but we were unable to obtain access to these data.

46. Inasmuch as there were nine grades, two racial groups, and four years in which students applied, analyses could potentially include seventy-two dummy variables representing all possible points of randomization. In practice, the number of dummy variables or "blocks" included in the analyses reported in table 3 varied between eleven and sixty-five, the precise number depending on the number of grades for which students applied in particular years. See W. G. Cochran, "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society,* Series A, 128 (1965), pp. 234–65, and D. B. Rubin, "William Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies," in Podurl S. R. S. Rao, ed., *W. G. Cochran's Impact on Statistics* (New York: John Wiley, 1984).
47. Rouse, "Private School Vouchers and Student Achievement."
48. Many factors contributed to the large number of missing cases: Milwaukee public schools administered tests intermittently; students were absent on the days the tests were administered; students left the city, left the choice program, or were excluded from testing; test scores were lost; and so forth. One can speculate that the large number of missing cases may bias results in one direction or another. Low performers may be more likely to be tested (because of federal requirements) or may be less likely to be tested (designated as special students); they may be more likely to have moved (live in mobile homes) or less likely to have moved (do not have many options). If the initial assignment to test and control groups was random, one may reasonably assume that all extraneous factors operate with equal effect on both treatment and control groups. The fact that most observable characteristics of the treatment and control groups do not differ significantly is consistent with such an assumption.
49. Witte and others, "Fourth Year Report."
50. These data are from the first four years of the choice school experiment. Test score information on the control group was not available on the world wide web for subsequent years.
51. We prefer the one-tailed t-test to estimate the statistical significance of the findings, because theory and prior research both suggest that students should perform better in private schools.
52. Results for three and four years after application were jointly estimated by averaging scores for students who were tested in both years and by using the single score available for each of the remaining students. Dummy variables were included for those who had only third-year or fourth-year scores.
53. The background characteristics of students who are included in the intention-to-treat category are virtually identical to those who actually enrolled, as reported in table 2.

CHAPTER 22

Two Sides of One Story

Chieh-Chen Bowen

Witte, Sterr, and Thorn (1995) wrote a fifth-year report for the Milwaukee School Choice program. Their conclusion was that students in the Choice program did not perform differently than students in public schools. Then Greene, Peterson, and Du (1998) obtained the data from the web site put together by Witte and colleagues. They reanalyzed the same data but came to the incompatible conclusion that Choice students significantly outperformed their counterparts in public schools after staying in the program for three or fours years.

What I observe here is two totally different conclusions drawn by two groups of researchers using the same data set. Three questions come into my mind immediately. What happened? How is it possible? Which of these two studies, if either, is to be believed? In the following critique of these two studies, I will answer the first two questions and shed light upon the third. For the sake of brevity, hereafter I will refer to the Witte and colleagues study simply as Witte and the Greene and colleagues study simply as Greene.

Most of us usually think statistics are objective and will not lie to us. Statistics, however, are like any other tools created by humans to achieve our goals. Different people can use the same tool to produce very different products. Statistical procedures, correctly applied, invariably produce means when you tell them to calculate means and compare the differences between two groups when you specify those two groups. They are completely objective in the sense that they follow the laws of mathematics and logic. Beyond this sense, however, in practice the results of statistical analyses heavily depend on the knowledge, background, skills, values, and intentions of the researchers. Although one of the biggest struggles of the social sciences is to produce value-free research, no matter how hard anyone tries, social science research cannot and will not be completely value free.

There are, however, reasonable boundaries outside of which research is not scientifically credible. Let us examine the specific methodological and statistical procedures used by the two groups of researchers to produce these incompatible conclusions.

Population and Sample

The first step in conducting an empirical study is to define the relevant population and sample. In turn, the definitions of the population and sample depend on the purpose of the study.

It is interesting in this regard that although Witte and Greene used the same data set, they used different definitions of the population and sample. Moreover, neither study explicitly defined either the population or the sample.

Based on the presentation of the analyses, I inferred that the population in Witte's study was made up of students in the Choice program and all students in the Milwaukee Public Schools (MPS). The sample seems to have been composed of those among these students whose test scores and/or parents' attitudes and demographic data were available, though this was never explicitly stated. In contrast, Greene's study clearly stated that they took students who applied for the Choice program as the population and that the sample was made up of those among these students whose test scores and/or parents' attitudes and demographic data were available.

Since these are different populations and samples, it is evident that the two studies have different scopes of interest. Witte apparently thought that all public school students' performances were relevant to the evaluation of the Choice program. In contrast, Greene evidently thought that only the performances of those students who actually applied for the Choice program are relevant. Accordingly the differences in the conclusions of these two studies can be largely attributed to the differences in defining their populations and samples for evaluation of the Choice program.

Selection of the Control/ Comparison Groups[1]

In general, selection of a control group can heavily influence if not completely determine the conclusions drawn by a researcher. For example, although you probably can neither sing better than Whitney Houston nor outperform Michael Jordan on the basketball court, you very well may be able to sing better than Michael Jordan and outperform Whitney Houston on the basketball court. I hope this example helps you to understand the importance of choosing your control/comparison group, not only in terms of to whom you compare but also in terms of which dimension you want to use to make the comparison.

Unfortunately, there are no iron rules to follow when choosing a comparison group. The general guideline is to pick a group as similar as possible to the experimental group in every variable except for the treatment (or independent variable of interest) in the study. Ideally, you want everything to be completely equal between groups, except the treatment.

In the case of these two studies, the students are not at all like rats in the laboratory. Rats can be chosen from the same family, fed the same food, caused to sleep in the same room, and exposed to exactly the same lighting and temperature tightly controlled by the experimenter. So a control or comparison group and an experimental group can be essentially equal before the study begins. Students, however, vary by race, income level, gender, and family background, among many other things. Moreover, any of these variables may have effects on the test scores. If a comparison group and the experimental group were known to be different in terms of these other variables (which they brought into the study with them), then it would make the evaluation of the Choice program necessarily uncertain and unclear.

One way that evaluators often deal with equalizing the control and experimental groups is through random assignment of individuals to groups. In this case, random assignment of students to groups could help equalize the other irrelevant variables between the two groups. If this were done and no differences were found on these other variables, then any observed differences on the test scores can be clearly attributed to the treatment, in this case, the Choice program.

Witte was apparently uncertain about how to choose the best comparison group.

Without providing any rationale, they used two comparison groups, (1) all MPS students, and (2) low-income MPS students. Both groups, however, were found to be different from the experimental group on some or all of the race, income, and family background variables. Besides, the low-income MPS group is a segment of the total MPS. This makes the two comparison groups dependent upon each other and creates a lot of headaches when any statistical comparisons are done.

In contrast, Greene reported that the Wisconsin legislature required that Choice schools, if oversubscribed, must admit applicants at random. Their experimental group therefore consisted of students who were selected into the Choice program. Their control group consisted of students who applied but did not get selected. The random assignment of applicants into either experimental or control group was performed by the schools themselves. This means of selecting a control group is clever; it can help equalize the control and experimental groups.

Although neither of the studies showed statistical significance on the differences of demographic variables between their experimental and comparison or control groups, table 3 in Greene showed more similarities between these two groups than Witte's table 5a–5e. So in my judgment, based on their effort to keep everything else constant, Greene came up with a better group for comparison than did Witte.

Response Rate

A response rate does not have to be 100 percent for a study to arrive at sound conclusions, as long as a *representative* sample is carefully selected. To be *representative* means that the sample's characteristics reflect the population's characteristics proportionally. In the abstract, only random sampling can be bias-free in this regard. Random sampling from a representative sampling frame is the only sure-fire method of ensuring a representative sample.

Neither of these studies used random sampling. Instead both simply tried to access as many records as they could within their defined populations. However, in both studies the records were far from complete. Incomplete records created a problem associated with response rate.

The basic response rate problem faced by both studies is that people who took the time to answer a questionnaire and mail it back may well have had different opinions than those who took a look at the questionnaire and threw it in the trash can. This is a form of self-selection, and it is well established empirically that self-selected samples usually have some traits different from the population. Moreover, when self-selection bias enters the picture at the same time that the response rate is too low, the result is highly likely to be a distorted picture of the population.

Specifically, in terms of these two studies, the response rates for Witte were 30.9 percent for the experimental and 31 percent for the comparison group for the test scores, but they dropped to 18.8 percent and 9.3 percent, respectively, for both test scores and parent survey. The response rates for Greene were 79 percent for the experimental and 72 percent for the control group for the test scores, but they dropped to 46 percent and 35 percent, respectively, for both test scores and parent survey. Although neither of these studies had an impressive response rate, Greene had a somewhat better response rate than Witte.[2]

Whether or not low rates of response lead to erroneous portrayal of a population depends on the reasons for nonresponse (Simon and Burstein 1985). In the case of these two studies, the question is an empirical one: Are the characteristics of the nonrespondents related to the information that the researchers seek to collect? It would take a detailed historical, ethnographic, and demographic comparison between respondents and nonrespondents

to fully provide the information needed to evaluate the effect of nonresponse on either of the studies' conclusions.

Appropriateness of Statistical Analyses

To save time and space, I shall only criticize the statistical procedures with which I disagree.

Two weaknesses were obvious in Witte's statistics. First, the random sampling assumption was violated. This prevented them from fully utilizing the data. More specifically, in tables 2 to 5e they seemed to compare the differences between the data in 1994 with all data accumulated over the five years. Yet because of this violation they could only present percentages and not anything about the statistical significance of the differences between them. It is also odd for Witte to come up with the idea of comparing a part with its total as the part is already included in the total. This idea makes these observations dependent on each other, a procedure commonly known as *dependent sampling*. Special procedures for dependent sampling are available, but there is no indication that they were used in this study. Most basic statistical procedures such as chi-square, *t*-tests and regression analyses assume the observations to be independent. Yet none of these analytic procedures could have been used credibly in this study since they compared the five-year data with data collected in 1994. Had this violation been ignored and basic statistical procedures used anyhow, the results would likely be incorrect estimates of parameters and incorrect conclusions.

Second, the data in Witte were not presented in a consistent manner. The experimental group in table 8 was comprised of students who *applied* to the Choice program. In Table 9, the experimental group became students who *enrolled* in the Choice program. Yet there is no explicit explanation for the switch. And then again in table 10, the samples were different from those in table 9. As a consequence of this sort of sloppy presentation, I simply could not make any coherent sense out of the numbers in these tables. Undoubtedly, Witte collected much more data than needed. It was their responsibility, however, to sort through the data and only present meaningful and useful analyses and results. Their article failed to do this in any sort of clear and distinct manner. I would speculate that the absolute lack of consistency from one table to the next is attributable to having had the data collected by a different person every year and having had each analysis also done by a different person (perhaps different graduate students). This would explain why, for example, tables 8 and 9 contain simple descriptive statistics (such as medians, means, standard deviations, and sample sizes) but tables 10a to 10e are simply a mess, containing twenty-six separate *t*-tests.

Any student who goes through a first statistics class should have learned that one should *not* present a large number of separate *t*-tests. One reason is that one out of twenty such tests will show significant differences as a result of chance alone. The significance of the *t*-value in such tests depends on how many *t*-tests you perform at the same time. Bonferroni multiple comparison method should be introduced when reading the statistical significance of each *t*-test. Another reason is that to do this ignores the interdependencies between groups and risks misconstruing the results.

In any case, because the two comparison groups are dependent upon each other, it is impossible to make any meaningful comparisons between groups. The only remedy for this is to pick one and only one comparison group and get some meaningful between-group comparisons.

Witte presented regression analyses in tables 11a to 12b. The members of the comparison group in these tables were MPS students whose test scores and parental survey

were available. In these two tables, Witte did not explain why they picked the total MPS students as the comparison group and discarded the other group, low-income MPS students. My speculation is that Witte liked to use as many data as possible. That is why they used the total MPS students as the comparison group. Multivariate regression analyses allow statistically equalizing many other confounding variables that may have effects on the student test scores, such as prior achievement and background characteristics. Then compare the effect of Choice program versus MPS. The results of the regression analyses clearly stated that there were no consistent differences between Choice and MPS students after statistically controlling for possible confounding variables.

In contrast, Greene's reported analyses were presented in a consistent manner under the carefully chosen experimental and control groups. One major point needs to be clarified before I discuss the analyses. Greene referred to the test scores as NCE (National Curve Equivalent) percentile points in the analyses, when they should be referred to simply as NCE points. Percentile points are ordinal scales without equal unit between each point, which would not allow any meaningful calculation of means, standard deviations, and *t*-tests. In tables 3, 4, 5, and 7, Greene presented one-tailed *t*-tests and used $p < .10$ as the statistically significant level to report the results. Again, there are no iron rules for choosing one-tailed versus two-tailed tests. One usually uses two-tailed tests to see if differences exist between experimental and control groups unless there is explicitly stated theory, logic, or previous research that strongly suggests the direction of the difference, in which case one may use one-tailed tests. Mathematically, one-tailed tests require lower critical values to reach statistical significance. One-tailed tests would be preferred only with sufficient theoretical and empirical justification, which was not clearly stated in Greene's study. The simple statement of personal preference given in footnote 51—"We prefer the one-tailed test to estimate the statistical significance of the findings, because theory suggests that students should perform better in private schools"—is not enough to justify choosing one-tailed *t*-tests. In addition, they failed to state which theory led them to make such directional prediction about students' performance in the Choice program, however. Especially, Witte's study kept repeating that there were no effects or even negative effects of the Choice program. Confronted with inconsistent empirical findings, and lacking theoretical argument, they should have used two-tailed *t*-tests.

The tone of Greene's article seemed to belie a hidden agenda. One bit of evidence is that they tended to overstate the significance of their results. Traditionally, $p < .05$ is set as the statistical significance level, but Greene lowered the significance level to $p < .10$ without any explicit explanation. The combination of using one-tailed tests and lowering the significance level to $p < .10$ makes the analyses look a lot more significant than they really are. A one-tailed *t*-test with $p < .10$ only requires $t > 1.282$ to be claimed significant, whereas a two-tailed *t*-test with $p < .05$ requires $t > 1.960$ or $t > -1.960$ to be claimed significant in a sample size larger than 120. When I follow the two-tailed tests with $p < .05$ as the significant level, roughly 75 percent of the significant analyses reported by Greene become insignificant.

Another interesting observation about Greene's study also supports my argument that Greene had a hidden agenda before he published this study. There was a multivariate regression analysis in an earlier version of Greene's study that was published in 1996. It basically had the same results as did Witte's regression analyses. But Greene conveniently forgot to mention the regression in the current paper.

Perception Differences

Sometimes two sides of the story appear to be contrary to each other at first glance, but upon closer scrutiny it all depends upon how you present the facts as you perceive them. It is hard enough to keep the results of program evaluations impartial, let alone when there is a personal agenda involved. Some evidence indicates that there are some personal long-term conflicts between these two groups of researchers. For example, there were some criticisms and rebuttals going back and forth between Witte and Peterson in the earlier reports of the evaluation of the Choice program. It is obvious that one side does not like the other. And both sides think the other side does not have an adequate understanding of statistics. It's very easy to be blinded by personal preferences or perceptions and so fail to see the consistency between two sides of the story.

My recommendation is to sharpen your knowledge and skill in evaluation first, and be as objective as you can when conducting empirical research. Do *not* say anything more than what the data can support. And last but not least, do not let personal conflict cloud your judgment.

Notes

1. Throughout this book Bingham and Felbinger have carefully maintained the distinction between control and comparison groups. In these evaluations of the Choice program, Witte used comparison groups, whereas Greene used a control group.
2. Greene reported lower response rates in the earlier version of the paper, "The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data from the Program's Evaluation," which was presented at the American Political Science Association in 1996. The response rates were 76.2 percent for the experimental and 58.7 percent for the control group for the test scores, but they dropped to 36.7 percent and 21.8 percent, respectively, for both test scores and parent survey. There was no explicit explanation why the response rates reported in 1998 could have increased when they were reporting the same data set.

References

Greene, Jay P., Paul E. Peterson, Jiangtao Du, Leesa Boeger, and Curtis L. Frazier. 1996. "The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data from the Program's Evaluation." Paper presented at the American Political Science Association, San Francisco, 30 August.

Greene, Jay P., Paul E. Peterson, and Jiantao Du. 1998. "School Choice in Milwaukee: A Randomized Experiment." In *Learning from School Choice,* edited by Paul E. Peterson and Bryan C. Hassel. Washington, D.C.: Brookings Institution Press, 333–56.

Simon, Julian L., and Paul Burstein. 1985. *Basic Research Methods in Social Science.* 3d. New York: McGraw-Hill.

Witte, John F., Troy D. Sterr, and Christopher. A. Thorn. 1995. "Fifth-Year Report: Milwaukee Parental Choice Program." Madison: Robert La Follette Institute of Public Affairs, University of Wisconsin–Madison, December.

Index

About the Authors

Richard D. Bingham is Professor of Public Administration and Urban Studies at the Levin College of Urban Affairs, Cleveland State University, where he is also Senior Research Scholar of the Urban Center. He teaches courses in industrial policy and research methods. His current research interests include the economies of urban neighborhoods and modeling urban systems. He has written widely in the fields of economic development and urban studies. His latest books include *Industrial Policy American Style* (1998); *Beyond Edge Cities,* co-authored with colleagues from the Urban University Program (1997); *Dilemmas of Urban Economic Development,* edited with Robert Mier (1997); and *Global Perspectives on Economic Development,* edited with Edward Hill (1997). He is Founding Editor of the journal *Economic Development Quarterly* and is Past President of the Urban Politics Section of the American Political Science Association.

Claire L. Felbinger is Chair of the Department of Public Administration, School of Public Affairs, at American University. She teaches courses in urban service delivery and research methodology. She has written in the areas of public works management, economic development, intergovernmental relations, and evaluation methodology. Her current research is focused on the process of local goverment decision making in multidisciplinary contexts. She is the founding editor of *Public Works Management & Policy.*

Chieh-Chin Bowen received her Ph.D. in Industrial/Organizational Psychology from the Pennsylvania State University. She has strong interests in research methods and applied statistics in social sciences. She was a faculty member of the Institute of Human Resource Management in the National Sun Yat-Sen University in Taiwan for three years. She moved to the Department of Psychology at Cleveland State University in 1997. Her current research topics are workforce diversity, performance appraisal, and selection.