1-1-1995

Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test

Stephen H. Bell
Abt Associates

Larry L. Orr
Abt Associates

John D. Blomquist
Abt Associates

Glen George Cain
University of Wisconsin - Madison

Citation

Bell, Stephen H., Larry L. Orr, and John D. Blomquist, and Glen G. Cain. 1995. Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

PROGRAM APPLICANTS as a Comparison Group in EVALUATING TRAINING PROGRAMS

Stephen H. Bell
Larry L. Orr
John D. Blomquist
Glen G. Cain

Program Applicants as a Comparison Group in Evaluating Training Programs

Theory and a Test

Stephen H. Bell
Abt Associates Inc.

Larry L. Orr
Abt Associates Inc.

John D. Blomquist
Abt Associates Inc.

Glen G. Cain
University of Wisconsin—Madison

1995

W. E. Upjohn Institute for Employment Research
Kalamazoo, Michigan

Library of Congress Cataloging-in-Publication Data

Program applicants as a comparison group in evaluating training
programs : theory and a test / Stephen H. Bell . . . [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0–88099–158–5 (cloth : alk. paper). — ISBN 0–88099–157–7
(pbk. : alk. paper)
1. Employees—Training of—Evaluation—Econometric models.
I. Bell, Stephen M. II. W.E. Upjohn Institute for Employment Research.
HF5549.5.T7P73 1995
658.3'12404—dc20 95–37399
CIP

W. E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007–4686

The facts presented in this study and the observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of the W. E. Upjohn Institute for Employment Research.

Cover design by J. R. Underhill
Index prepared by Shirley Kessel.
Printed in the United States of America.

Acknowledgments

The authors wish to thank the Upjohn Institute and its staff for their support of this study. We are particularly grateful to the Institute's former executive director Robert Spiegelman and its assistant executive director H. Allan Hunt for allowing us the time required to obtain the necessary data and work through the many difficult conceptual issues involved in the study. The research staff of the Institute also provided valuable feedback on the monograph in a seminar in September 1994. We are indebted to Robert Spiegelman, Thomas Fraker, and three anonymous reviewers for insightful comments and suggestions on earlier drafts.

The study also benefited substantially from technical review at Abt provided by Stephen Kennedy, Chief Social Scientist, whose advice on statistical matters was particularly valuable. Other Abt staff who contributed to the monograph were Winston Lin (who made useful comments on portions of the volume), Annirudha Bonnerjee (who programmed the algorithm for the risk functions presented in chapter 4), and Margie Washington (who provided secretarial support throughout the project). Michael Tilkin, a graduate student at the University of Wisconsin, assisted in the estimation of impacts.

Finally, the authors are indebted to the Health Care Financing Administration of the U.S. Department of Health and Human Services, which funded the AFDC Homemaker-Home Health Aide Demonstrations on which the empirical analysis is based. We should also thank the Office of Income Statistics of the U.S. Internal Revenue Service, which provided long-term follow-up data for the homemaker sample.

This monograph is the end product of several years of collaboration among the authors, beginning with the summer of 1991 which Cain spent at Abt helping to develop the overall structure of the analysis. Chapter 1 was written by Bell and Cain and chapter 2 by Orr, Bell, and Cain. Blomquist was the principal author of chapter 3 and performed the empirical analysis presented in that chapter; the chapter builds on an empirical analysis by Cain presented to the 1993 meetings of the American Economics Association. The validation techniques presented in chapter 4 were developed by Bell and Orr. Chapter 5 was a joint effort by all four authors.

The Authors

Stephen H. Bell is a senior economist and project director at Abt Associates Inc. in Bethesda, Maryland. His research focuses on econometric methods for public policy evaluation, including measurement of program impacts using classical experiments and nonexperimental methods. His most recent publications appear in the *Journal of Human Resources* (on the cost-effectiveness of training welfare recipients) and the *Social Security Bulletin* (on employment programs for people with disabilities). Upcoming titles include a book on the National JTPA Study and papers on validating nonexperimental impact estimates and designing employment services for recently released prisoners. Dr. Bell received his Ph.D. in economics from the University of Wisconsin—Madison.

Larry L. Orr is a senior economist and vice president at Abt Associates Inc. in Bethesda, Maryland. His research interest include the analysis of public policy issues using experimental methods and the evaluation of employment and training programs; he has participated in the design of more than twenty randomized field experiments. Dr. Orr was project director of the recently completed National JTPA Study and the evaluation of the AFDC Homemaker-Home Health Aide Demonstrations and serves on the editorial board of the journal *Evaluation and Program Planning*. From 1991 to 1993, he was a member of the National Academy of Science Committee on Postsecondary Education and Training. Prior to joining Abt, he directed the Office of Technical Analysis, U.S. Department of Labor, and the Office of Income Security Policy Research, U.S. Department of Health, Education, and Welfare. Dr. Orr received his Ph.D. in economics from the Massachusetts Institute of Technology.

John D. Blomquist is a senior economist at Abt Associates Inc. in Bethesda, Maryland, His research interests include applied econometrics, survey design, labor economics, and public policy. Dr. Blomquist's recent research has addressed a variety of policy and evaluation topics, including welfare reform, housing policy, return-to-work initiatives, and transportation regulation. He has presented work at both the annual meetings of the Association for Public Policy Analysis and Management and the National Association for Welfare Research and Statistics. Dr. Blomquist received his Ph.D. in economics from the University of Pennsylvania.

Glen G. Cain is a professor of economics and is affiliated with the Institute of Research on Poverty and the Industrial Relations Research Institute at the University of Wisconsin—Madison. He joined the department of economics there in 1963 and in 1965 began evaluation research of the training programs funded by the Manpower Development and Training Act. He spent the 1966-67 academic year at the Office of Economic Opportunity in Washington, DC and studied various antipoverty programs, including the New Jersey income maintenance experiment. He was editor or associate editor of the *Journal of Human Resources* from 1968-1984. His current research is on the impact of international trade on employment and wages, on racial differences in youth employment, and on women in the labor force.

Contents

List of Exhibits

1
Methods Used to Evaluate Employment and Training Programs in the Past

Evaluation of employment and training programs has been a central focus of workforce policy decisions in the United States for nearly 25 years, yet remains controversial. Despite major advances in evaluation methods, it is not clear that the nation has the tools it needs to obtain unbiased measures of the benefits of any particular training intervention. Without such measures for past policies, wise choices cannot be made among policy options for the future.

It is generally agreed that experimental evaluations, with random assignment to program and control groups, are more likely to provide unbiased estimates of program impacts than are alternative methods. It is also widely recognized that such evaluations cannot be implemented in all situations.[1] Therefore, over the last several decades, labor economists have developed increasingly sophisticated nonexperimental econometric methods to estimate the effects of employment and training programs using (nonrandom) "comparison groups" drawn from external (i.e., nonprogram) sources to represent what would have happened to participants in the absence of the program.

Despite these efforts, there is still no generally accepted nonexperimental method for estimating the impacts of such programs on the earnings and other outcomes of participants. Different methods yield markedly different estimates, even when applied to identical samples and data.[2] The critical objective of these methods has yet to be achieved: adjustment of outcomes to remove any preexisting differences between participants and the nonexperimental comparison group that would otherwise be mistaken as program impacts.

This monograph critically reviews the many nonexperimental impact estimation approaches introduced over the years that are based on external comparison groups. It then proposes an "internal" comparison group that we believe holds considerable promise: applicants for the same programs who for various reasons do not participate. No recent studies have used the population of nonparticipating applicants as a benchmark for measuring program effects, and none has ever tested the effectiveness of that approach using experimental data.

Compared to primitive uses of the applicant-based approach during the formative years of employment program evaluation (the 1960s and early 1970s), we extend the methodology here by:

- Giving it a stronger theoretical rationale, which makes clear how certain conceptual limitations of external comparison groups are corrected through the use of internal, applicant-based comparison groups;
- Incorporating more information on preexisting differences between excluded applicants and participants than has been available in the past, including measures that capture the criteria program staff used to determine which applicants participate; and
- Testing the applicant-based measures against estimates of program impact taken from a randomized field experiment.

We begin in this chapter by reviewing the history of employment and training program evaluation, with a focus on the methodological lessons to be learned from that history.[3] We then present a theoretical rationale for the applicant-based approach in chapter 2. Chapter 3 describes the data we will use to test the approach and develops nonexperimental impact estimates from those data using applicants as comparison group members. We test the new estimates against the original experimental findings in chapter 4 to determine which, if any, provide promising alternatives to the experiment. Chapter 5 summarizes our conclusions and their implications for future employment and training evaluations.

The Importance of Employment and Training Programs

The U.S. government has invested in worker training and employment programs at least since the late 1950s.[4] By fiscal year 1991, 14 federal departments and agencies ran 125 such programs at a cost of $16.3 billion per year, consuming just over 1 percent of all federal expenditures.[5] Additional state programs are numerous, though not nearly so large (many are funded in part by federal dollars), while local governments, private foundations, and employer groups also contribute to the nation's workforce training effort.[6]

In total, these programs serve many millions of American workers each year in an attempt to increase worker productivity and incomes. A great deal may be at stake in such investments. Increasingly, the skill level and employment success of the nation's workforce are viewed as the key to America's standard of living and competitive position in the world economy.[7] Thus, the importance of evaluating the nation's many workforce programs to distinguish effective from ineffective investments can hardly be overemphasized.

Early Evaluations of MDTA

Serious evaluation of government employment and training programs began with the Manpower Development and Training Act (MDTA) programs of the 1960s. The U.S. Congress enacted the MDTA in 1962 to expand federal retraining services for workers who lost jobs due to technological change[8] and, for the first time, to attempt to improve the long-run earnings capacity of low-skill workers in general. Operationally, MDTA focused not just on classroom skill training as had earlier programs, but on on-the-job training and basic education as well.

Beginning with Borus (1964), several researchers attempted to measure the impact of MDTA on participants' employment and earnings.[9] In hindsight, reviewers of these early studies found them to be uneven and generally unsatisfying in terms of quality and statistical validity.[10]

Some of these studies measured impacts as the change in participant outcomes over time from the preprogram to the postprogram period.[11] Under this approach, any program that evidenced a substantial upward trend in employment and earnings tended to be viewed as a success, at least if the earnings gain exceeded that for all workers over the same period.[12] Unfortunately, this approach ignored the possibility that people enter employment programs at low points in their labor market histories (e.g., following job loss) and therefore stand to improve their fortunes more than the average even without special government assistance.[13] If this is true, pre/post measures of program impacts have a built-in bias toward favorable conclusions.

Later findings of sharply downward trends in earnings just prior to program entry—the so-called "preprogram dip"—noted by Ashenfelter (1978) and Ashenfelter and Card (1985), among others, seemed to confirm the importance of this problem. So did still later experimental evaluations of job training programs, where a random subset of those who would otherwise have entered training were precluded from doing so. Follow-up data for these experimental control groups showed sharply rising earnings paths in the period after program application even in the absence of any intervention.[14]

The possibility of "dip and recovery" led evaluators to develop an alternative benchmark with which to judge program effects. If the initial position of program participants provided an unreliable standard, then perhaps a benchmark could be derived from the experience of similar workers who did not receive training assistance. Other early studies of MDTA adopted that tack, usually adjusting for any remaining baseline differences between the participant and comparison groups using statistical matching or multivariate regression techniques.[15] For a time, comparison group strategies of this sort were accepted as an appropriate basis for judging past policies and, implicitly, for making future policy.

Confronting the Selection Bias Problem

Later evaluations of MDTA added new sophistication to the comparison group strategy.[16] Here, evaluators focused squarely on the problem of "self-selection"—that individuals who self-select into employment and training programs are systematically different from other apparently similar workers who do not seek assistance. In view of this possibility, it becomes necessary to control not only for differences in general demographic characteristics (e.g., age and education) between program participants and comparison group members at baseline, but also for the particular factors that motivate program entry at a point in time. Here, complex econometric techniques enter the employment and training program evaluation literature for the first time.

In his overview of the econometric evaluation of training programs, Moffitt (1987) cites Ashenfelter (1978) and Bloch (1979) as the first to confront the selection bias problem head on.[17] Attention focused on possible corrections for selection bias through the use of preprogram earnings to predict a valid postprogram earnings benchmark. Goldberger (1972) and Cain (1975) noted the potential for this approach to remove selection bias under the strong assumption that systematic selection into the program was based only on observable variables, such as preprogram earnings. Ashenfelter (1978) was the first to apply the approach to real data in his analysis of MDTA. A number of refinements and commentaries on the approach followed, including Kiefer (1979), Cooley, McGuire, and Prescott (1979), Director (1979), and Bloom (1984a).

The CETA Evaluations.

These models provided the foundation for the next generation of training program evaluation, which focused not on MDTA but on programs funded under its successor, the Comprehensive Employment and Training Act of 1973 (CETA). Extending the MDTA approach, CETA offered public service employment and (for particularly disadvantaged workers) unpaid work experience in addition to classroom and on-the-job training. Barnow (1987) summarizes the many analyses of CETA impacts commissioned by the U.S. Department of Labor in the 1970s and early 1980s.[18]

Without exception, the CETA studies focused on the comparison of earnings for CETA participants and similar individuals in the population at large.[19] They also used preprogram earnings differences to equalize the two populations at baseline in all cases. As Barnow (1987) notes, these studies "vary considerably in their findings and conclusions on the impact of CETA" (p.175) and "the results are sensitive to the specific methods adopted" (p. 157). In particular, Barnow concludes that an important source of variation in the estimates was the way different evaluators used preprogram earnings to predict postprogram earnings:

> Earnings in the year immediately prior to participation in a training program tend to decline from the trend in the years preceding it. The treatment of the 'preprogram dip' in the analysis can play a

> substantial role in the estimates of program impact. If the dip is a transitory phenomenon, then it could influence selection into the program without having a long-term impact on earnings. . . . On the other hand, if the dip indicates a permanent decline in human capital (or the value placed by society on the human capital), then earnings in the period immediately prior to program participation is likely to be a key variable in explaining later earnings (pp. 184-185).

This observation raises a serious problem for the design of evaluations using external comparison groups. If the preprogram dip is purely transitory, one need only match the participant and comparison groups on earnings prior to the dip (and follow both groups long enough to measure postprogram earnings beyond the dip) to obtain a comparison group that is well-matched to participants on permanent income.[20] But if the loss of earnings that triggered program entry signifies a permanent break in the earnings trend for participants, earnings prior to that break contain very little information about postprogram earnings, and therefore cannot be used to identify an appropriate external comparison group (or, what is the same thing, to adjust for differences in postprogram earnings that are not due to the program).

This uncertainty casts serious doubt on any method that relies heavily on preprogram earnings to predict postprogram earnings. If the preprogram dip is both unprecedented (for the individual) and permanent, this strategy cannot work by definition. If instead it represents a mix of transitory and permanent changes for any group of program participants, one can never be sure of the mix, much less how to predict future earnings for the subset of participants experiencing permanent shifts.[21] Finally, even the best scenario—a situation where all pre-program earnings changes are transitory—does not solve the problem, since the analyst has no means of recognizing that situation when it occurs.[22]

On the basis of his review of the CETA studies, Barnow concluded that:

> [Randomized field experiments] appear to be the only method available at this time to overcome the limitations of nonexperimental evaluations (p. 190).

Experiments create "internal" comparison groups of control group members who, because they are a random subset of would-be participants, will on average follow the same permanent and transitory earnings paths that participants would have absent the program. Hence, subject only to sampling error, the control group provides an appropriate benchmark, or counterfactual, for measuring program effects. As noted below, Barnow's conclusion that controlled experiments are the preferred method for evaluating training programs eventually came to be shared by most of the evaluation community.

Two-Stage Methods.

Another external comparison group strategy for addressing the selection bias problem was proposed concurrent with the CETA studies: the use of two-stage selection models to jointly explain participation in employment and training programs and its effects on earnings. The most widely cited two-stage technique for addressing selection bias in the labor market is that introduced by Heckman (1974, 1976, 1979).

Under this approach, specific statistical assumptions about the relationship between the decision to participate in a training program and the participant's future earnings provide a way to equalize the starting point for the program and comparison groups when measuring program impacts. These assumptions require that the factors that influence both program entry and later earnings, such as educational level and motivation, are either controlled for in the analysis through measured variables or jointly influence these two outcomes according to the well-behaved statistical patterns of the bivariate normal distribution.[23] If these assumptions hold true, the resulting estimates of program effect are unbiased. In fact, the model has been found to be sensitive to the assumption of bivariate normality in studies by several econometricians.[24]

The two-stage model for selection bias adjustments has not been widely used for employment and training evaluations, although it has in many other econometric applications.[25] Only one of the CETA studies (Westat 1984) attempted the methodology, but the estimates of program impact were reported to be too sensitive to the variables included in (or excluded from) the model to be useful.Manski (1989) summarizes the current state of the econometrician's unease with the method when he refers to the two-stage selection model's "fragility," in which "seemingly small misspecifications may generate large biases in estimates" of the program's effects (p. 356).

To summarize, then, the problem of selection bias—while perhaps much better understood—appeared just as intractable following the CETA studies as before. Direct empirical support for this conclusion appeared almost immediately thereafter.

Testing Nonexperimental Estimates Against Experimental Findings

As the CETA findings emerged, several researchers began to examine the problem of selection bias in the various comparison group strategies employed by the CETA researchers using data from controlled field experiments. The use of experimental methods for social policy evaluation began in the late 1960s and early 1970s in other policy contexts, specifically with regard to the effects of a national negative income tax.[26] Under the experimental approach, the group of individuals that would normally be subjected to a policy or program is split at random prior to the intervention and only a portion "treated" with the policy or program. The remaining group—which differs from the participants only by random sampling error—then serves as a control group" for measuring the effects of the intervention, in much the same way that controlled experiments are used to test new drugs in a laboratory or clinical setting. In large samples, chance differences in preexisting characteristics between the treatment and control groups tend to disappear (and, in any case, can be taken into account in standard statistical tests), effectively removing the self-selection problem that is at the heart of any nonexperimental impact analysis.

The first training program to use random assignment to select participants was the National Supported Work Demonstration, which provided intensive training and work assistance to severely disadvantaged workers such as long-term welfare recipients, disadvantaged youth, and ex-offenders.[27] In the mid-1980s, LaLonde (1986) and Fraker and Maynard (1987) reanalyzed the original Supported Work data with nonexperimental methods, as though the experimental control group was not available, and compared the resulting estimates with the experimental findings. They used the same technique that had been applied to the CETA data, drawing external comparison groups from national data bases by selecting a sample of individuals who were similar to the participants on the basis of certain observed characteristics. They produced estimates of earnings impacts that varied as much from one another as the original CETA estimates.

More important, LaLonde and Fraker-Maynard for the first time demonstrated that few of the nonexperimental estimates came close to the experimental estimate, which was presumed to be free of selection bias. Moreover, estimates derived from more sophisticated and more theoretically compelling techniques performed only a little better than more primitive approaches and still left a wide margin for error.[28] Most observers saw this as a graphic illustration of the potential for selection bias to invalidate even the most sophisticated nonexperimental techniques.[29] An immediate consequence was a widespread and rapidly growing preference among policy makers, both in Congress and among executive agencies, for experimental over nonexperimental training program evaluations.[30]

Responses to the Unfavorable Test Results

Realizing that controlled field experiments could not, or would not, be used in all applications, some evaluators responded to the LaLonde and Fraker-Maynard results not so much as an indictment of external comparison group techniques but as a challenge to improve them. We review those responses below.[31]

Model Specification Tests.

The most direct response came from Heckman and Hotz (1989), who argued that many of the estimation techniques considered by LaLonde and Fraker-Maynard could—and should—have been rejected prior to the comparison to the experimental benchmark on the basis of their conceptual implausibility and/or their demonstrable inconsistency with the nonexperimental data.[32] Making these exclusions, Heckman and Hotz contended that the remaining plausible estimates are much more similar to one another, and—in their policy implications—to the experimental estimate than the original group. Others, however, have not found these tests to be helpful; see for example, Friedlander and Robins (1992).

A conceptual problem at issue in this method is the absence of explicit criteria for choosing among econometric methods and their various estimates when there is no experimental estimate against which to compare them. Heckman and Hotz's response to this problem was to develop a series of model specification tests, based on methods first introduced in Heckman and Robb (1985, 1986). They argued that evaluators should accept or reject each nonexperimental estimation technique based on how well its assumptions accord with the available data. Given enough preprogram data, many nonexperimental techniques can be tested in the absence of a controlled experiment (which, of course, is the only situation in which such tests are needed). These include approaches that assume earnings are steady over time (testable with two or more preprogram observations) or that earnings vary at random around some steady-state trend line (testable with three or more preprogram observations).

In the best case, model specification tests would reduce the range of nonexperimental estimates to a tight band around the experimental benchmark. If the "tightness" of this band—or at least some measure of consistency among the remaining estimates as to policy implications (e.g., whether a program has a positive or negative effect)—can be established from nonexperimental data, one should have greater faith that the group of estimates as a whole comes close to the (unobserved) experimental benchmark. One's faith in the approach should grow further still with each instance in which it replicates the results of a true experiment, of which Heckman-Hotz was the first attempt.

Better Comparison Groups and Baseline Data.

A second, related response to the limits of existing nonexperimental estimators was pioneered by the National JTPA Study sponsored by the U.S. Department of Labor in the late 1980s. This $23 million study of the Job Training Partnership Act (JTPA) for the first time combined both experimental and nonexperimental elements in its design. Approximately $5 million was used to study the selection of program participants and to assess the validity of nonexperimental techniques. To provide a basis for nonexperimental comparison groups, the project identified and interviewed 2,300 individuals in the study areas who were eligible for JTPA services but did not participate. The eligible population was viewed as an external comparison group in which the preexisting factors separating participants from nonparticipants could be identified and included in the model to eliminate selection bias from the estimated program impact.[33]

While results are not yet available from this undertaking, its design has many desirable features. This external comparison group was selected on the basis of its similarity to the group assigned to JTPA in terms of location and current economic circumstances that determine JTPA eligibility. Interviews with these individuals focused on detailed employment and earnings histories over the five years prior to eligibility determination and 18 months after. Data were also collected on respondents' understanding of and inclination to pursue eligibility for a variety of employment assistance programs, including JTPA. The purpose of this data collection strategy was to discover the reasons that some eligible individuals applied to and entered JTPA at a point in time, while others applied and did not enter and still others (the external comparison group) did not even apply to the program. Visits to the study sites by the principal researchers were designed to heighten this understanding by looking at the program intake process itself.

In many respects, this research project represents the limit of what can be accomplished through reliance on comparison groups generated external to the program under study. It maximizes the comparison group match to participants, the information available to control for any remaining differences, and the econometric expertise needed to make those adjustments. Thus, once completed, the study should provide a useful test of the potential validity of external comparisons.

Nonparametric Bounds on Effects

In the interim, an entirely new approach has been introduced by Charles Manski. First applied to the measurement of the effects of family structure on high school graduation (see Manski et al. 1992), this strategy uses nonparametric methods to place bounds on the selection bias in estimating program effects. In contrast with the current econometric methods of modeling the selection process, which require rather restrictive assumptions about functional form and other parametric assumptions, Manski's "nonparametric" method is virtually assumption-free.

The technique is best illustrated when the outcome is binary, such as graduating from high school or obtaining a job. In this case, the impact of the program must be within a fixed range that is determined by the outcomes of participants and nonparticipants and the relative shares of the population in each group. To use the method for continuous outcomes, such as earnings, more restrictive assumptions are required. Whether the bounds derived by this method will be tight enough to give useful guidance to policy decisions is an open question, as Manski acknowledges.[34]

The real payoff to the approach may come only as carefully selected assumptions are added to the model to narrow the initial bounds to some meaningful level.[35] In any case, the method has the virtue of imposing a "from the ground up" assessment of the implicit assumptions imbedded in all previous (and future) nonexperimental estimators, making clear the tradeoff between the strength of the assumption and the progress it provides in narrowing the bounds of uncertainty.

Comparison Site Designs

A fourth strategy, more popular with policy makers than with researchers in the late 1980s and early 1990s, is to design evaluations around random assignment of local areas such as counties or other units of local government to program or comparison status.[36] In these "comparison site" designs, comparison groups are taken from the population of potential participants (e.g., AFDC recipients) in alternative geographic areas, either by purposively matching comparison sites to predetermined program sites or by picking matched pairs of counties and then deciding at random which one will host the program.

Some types of effects can *only* be analyzed with comparison site designs. If, for example, the interest is in estimating the impact of a "saturation" treatment or effects at the community level, the program must include all individuals within the community; it cannot be implemented for a just sample of individuals who are randomly assigned to treatment. Comparison site designs can also capture effects that occur prior to the point at which random assignment could feasibly be implemented, such as changes in the rate at which individuals apply to a program.

In principle, when pairs of sites are randomly assigned to treatment or control status, this approach removes the selection bias problem just as effectively as random assignment of individuals, without the added complication of deciding individual fates one at a time.[37] As with random assignment of individuals, treatment sites do not differ systematically from comparison sites on the nonprogram factors that affect outcomes. But they may still differ substantially on those factors by chance alone, given the small number of sites involved in most such studies.[38] However, if most of the variation in the outcome of interest (e.g., earnings) is at the individual level, so that average outcome levels tend to be similar across localities, a relatively small number of randomly assigned sites could provide highly reliable impact estimates.

Comparison site designs have the disadvantage that they cannot be used to evaluate existing programs without discontinuing local operations in the comparison sites. Moreover, problems can arise even when the approach is applied to demonstrations of new programs in selected counties. If the program is voluntary, the preferred comparison of participants in program sites with "participant-like" individuals in nonprogram sites becomes impossible, since one has no way of identifying who would have participated in the nonprogram sites had the program been offered. The most obvious alternative—comparisons of participants with the entire eligible population in the nonprogram sites—reintroduces the self-selection problem common to earlier comparison group approaches. The best that can be done in this situation is to compare those who meet the program's eligibility rules between the two sets of sites, adjusting for the fact that most eligibles do not participate.[39] Unless the participation rate among eligibles in the program sites is quite high, however, the resulting impact estimates will be relatively imprecise.

Overall, comparison site designs remain an option of necessity more than of choice when evaluating mandatory employment and training demonstrations. And they certainly are not a solution to the more general problem of self-selection when evaluating existing voluntary programs such as JTPA.

Instrumental Variable Approaches

A long-standing approach to dealing with the endogeneity of selection into certain states, such as participation in training programs, is to apply various econometric techniques used in simultaneous-equation estimation. These methods have only recently been applied to the evaluation of training programs. In this context, the first equation models program participation, and the second equation models participant outcomes. The equation modeling participation must include one or more determinants (variables) that do not, on their own, influence the outcomes. In the nonexperimental evaluation of the Job Corps, for example, distance from the nearest Job Corps center was found to be a good predictor of participation, but not of earnings.[40] If such factors can be found, they can provide reliable information on the effects of participation per se, free from the influence of selection.

In practice, econometricians have frequently found it difficult to identify a factor that might influence participation that does not otherwise influence earnings. Caution in choosing such "instruments" is well justified, since making an erroneous exclusion restriction from the earnings equation can easily lead to substantial bias in the impact estimate.[41]

Angrist and Imbens (1991) and Imbens and Angrist (1992) recast the search for an exclusion restriction in a two-stage model as a need for an "instrumental variable" that can be used to estimate program impacts in a single stage. If a factor can be identified that affects participation but not earnings (except through participation), it can be used as an "instrument" in place of the usual indicator for participation in an earnings impact equation. Angrist and Imbens discuss possible instruments in several applications, though not that of evaluating the earnings effects of employment and training programs.

In general, the use of instrumental variable methods of nonexperimental analysis has to be carefully justified in a particular context. The conditions necessary for accepting assignment to a treatment group as a valid instrument for participation are widely accepted; those involving other instruments are not. Sometimes, nonrandom variation in access to programs occurs naturally due to geographic or other factors, but these same factors may affect future earnings in ways not otherwise controlled for in the model. Thus, while valid instruments for program participation (other than random assignment) may exist, they must be discovered and justified in each specific evaluation application. Random assignment, on the other hand, always provides a strong starting point for deriving valid instruments.

Lessons from the Literature

On the basis of this review, we draw four major lessons from the thirty-year history of employment and training program evaluation:

1. Assumptions about the selection process that distinguishes program participants from nonparticipants (and from their own prior experience) are inevitable in any meaningful analysis of program impact.
2. The best and most credible impact estimates are those whose assumptions are clearest, most limited, and most plausible *a priori*, and most testable *ex post*.
3. It will be difficult to use data on the characteristics of participants and nonparticipants to replace knowledge of the selection process as the best starting point for measuring program impacts.
4. In voluntary programs, it is particularly critical to take account of the time path of participants' earnings around the point of program entry. Participants tend to enter a program at a low point in their earnings history—the "preprogram dip"—and, absent intervention, may or may not emerge with their earnings restored to previous levels.

None of these points is a new insight. Manski makes point 1—the inevitability of assumptions—most sharply by starting without assumptions and showing what must be added to obtain meaningful results. The same point is driven home by the long history of evaluators introducing new techniques that avoid the assumptions of earlier approaches and ending up simply shifting the debate to the validity of their own set of assumptions.

The importance of limiting and testing assumptions wherever possible—point 2—is also fundamental to much of the work reviewed here. Angrist and Imbens (1991, pp. 1-2) make this point most succinctly: "Disagreements over evaluation methodology notwithstanding, research . . . allowing for fewer assumptions in observational analyses is likely to remain important." The development of model specification tests (by Heckman and others) has improved but not assured the success of methods relying on external comparison groups and tests of assumptions.

Point 3 has also appeared in various forms in the literature for at least twenty years, beginning with Goldberger's (1972) observation that knowing the selection rule and having data on its determinants is sufficient for unbiased estimation. The same point is fundamental to mainline evaluation handbooks in the education field (e.g, Campbell and Stanley 1966; Cook and Campbell 1979), which urge evaluators to impose well-understood and carefully monitored selection rules when designing impact evaluations.

Finally, while point 4 has been well known for many years, its implications have perhaps been less than fully appreciated. In particular, early evaluations based on external comparison groups essentially ignored this point in attempting to use individuals who are (on average) in steady state in their earnings histories as benchmarks for individuals with transitorily low earnings, while typically controlling for only fixed factors such as race, sex, and education. The more sophisticated attempts to adjust for preprogram earnings differences are also fraught with difficulties. In particular, the loss of earnings that typically triggers program entry among participants may signify a permanent break in earnings trends, so that preprogram earnings contain essentially no information about subsequent "without program" earnings levels. In this case, even comparison groups that are well matched on *permanent* preprogram earnings (e.g., by matching on earnings before the pre-program dip) will yield biased estimates of program impact.

These conclusions suggest that external comparison groups may not provide the best benchmark for measuring training program impacts. As an alternative, evaluators might consider internal comparison groups of nonparticipating program applicants, whose division from participants is based on simple and well-understood selection rules and whose comparability to participants—especially with respect to the time path of earnings—can be established with a minimum of assumptions and data. While such a strategy will not necessarily avoid all of the problems that have surfaced in the literature over the years, we believe it is worth trying. We begin one trial of the approach in the next chapter.

NOTES

1. The use of experiments is sometimes limited by the operational and ethical problems that arise when randomly excluding individuals from program services. See, for example, Burtless and Orr (1986) or Manski and Garfinkel (1991) for a discussion of this issue.

2. See, for example, LaLonde (1986), Fraker and Maynard (1987), and Barnow (1987).

3. Moffitt (1991) provides a similar review of the literature through 1989, drawing substantially different conclusions from those presented here.

4. O'Neill (1973) provides a succinct overview of early programs, then called "manpower" programs, many of which were supported under the Manpower Development and Training Act of 1962.

5. These figures include spending on postsecondary education as well as job training and placement programs for adults and non-college-bound youth. See U.S. General Accounting Office (1992) for details.

6. Miller and Buckley (1993) estimate that U.S. employers invest 1 to 2 percent of their payroll expenditures in worker training, a figure in the tens of billions of dollars.

7. See Reich (1983), Johnston et al. (1987), and U.S. Congress (1990) for three of the many recent "call to arms" statements on this theme.

8. The Area Redevelopment Act of the late 1950s provided skill training and placement assistance to displaced workers prior to MDTA.

9. We define "impact" as the change in outcomes due to the program—i.e., that portion of the outcomes that would not have occurred absent the program. Operationally, this can be thought of as the *difference* between the outcome given the program (usually observed) and the outcome that would have occurred for the same person had he or she not participated in the program (which cannot be observed directly). Other evaluations of MDTA focused exclusively on program administration and the observed postprogram outcomes of participants, rather than on impacts.

10. For example, O'Neill (1973) concludes that the early studies "vary tremendously in terms of quality of data and statistical methodology" (p. 10). Other reviews, not all as critical as O'Neill, include Somers (1968), Hardin (1969), Borus and Buntz (1972), Goldstein (1972), and Perry et al. (1975).

11. See, for example, Goldfarb (1969), U.S. Department of Labor (1970), or Smith (1970).

12. Smith (1970) stood out among the early evaluators by comparing trainee wage gains to those of workers in the economy in general before interpreting upward trends as program effects.

13. This phenomenon, which is known in statistics as "regression to the mean", had been noted by a number of researchers; see, for example, Cain and Hollister (1969). Ashenfelter (1978) and Kiefer (1979) provide excellent discussions of the problem and methods for dealing with it. Note that the point does not necessarily apply to employment and training programs in which participation is mandatory, such as those that have been the focus of much of the recent literature on evaluation of programs for AFDC recipients (see, for example, Gueron and Pauly 1991). When participation is imposed from the outside, as in mandatory work-welfare programs such as the

AFDC JOBS program or the food stamp employment and training program, one would not necessarily expect participants to begin at unusually low points in their labor market histories.

14. See, for example, Bell and Orr (1994) and Bloom et al. (1993).

15. See, for example, Borus (1964), Main (1968), Stromsdorfer (1968), Hardin and Borus (1971), Prescott and Cooley (1972), and Farber (1972).

16. See Ashenfelter (1978), Kiefer (1979), and Bloom (1984a).

17. Others had previously addressed the effect of self-selection on non-training-related labor market outcomes using sophisticated econometric techniques. See, for example, Ashenfelter and Johnson (1972), Greenberg and Kosters (1973), and Heckman (1974).

18. These analyses included Westat (1981), Bloom and McLaughlin (1982), Bassi (1983, 1984), Westat (1984), Bassi et al. (1984), Dickenson, Johnson, and West (1984, 1986), and Geraci (1984). Additional analyses of CETA not included in the Barnow review appear in Bryant and Rupp (1987), Rupp et al. (1987), and Card and Sullivan (1988).

19. Data for this comparison were taken from a nationally representative sample of CETA enrollees interviewed by the U.S. Bureau of the Census, members of the U.S. population at large interviewed as part of the Bureau's March Current Population Survey, and several years of matched social security earnings records for both samples. Collectively, this data base was known as the Continuous Longitudinal Manpower Survey.

20. In practice, of course, it may be difficult to identify the "pre-dip" period and to obtain data on earnings during that interval, either because of data constraints or because sample members do not have extensive employment histories (e.g., youths and women entering or reentering the labor force). It is also true that as the preprogram and postprogram earnings observations are separated further in time, preprogram earnings becomes a less powerful predictor of postprogram earnings in general.

21. Ashenfelter and Card (1985) also recognized this problem and attempted to address it, but in the end concluded that only an experimental design could be relied upon to yield unbiased estimates in the face of this uncertainty.

22. Observation of the subsequent earnings of trainees cannot resolve this problem, since later earnings reflect both the natural "rebound" (or lack of rebound) from the preprogram dip and the effects of the training program intervention.

23. Maddala (1983, pp. 260-71) provides a useful discussion of these assumptions and other aspects of the two-stage model for correcting for selection bias.

24. See Goldberger (1983). Horowitz and Neumann (1987) and Newey, Powell, and Walker (1990) explore the implications of relaxing the bivariate normal distributional assumption in other applications. To our knowledge, this extension has not been undertaken in the context of training program impact analysis.

25. See Benus and Byrnes (1993) for a recent exception.

26. See Greenberg and Shroder (1991) for an overview of these and a large number of other social experiments.

27. See Hollister et al. (1984).

28. Couch (1992) repeated a portion of this analysis with longer-term follow-up data and obtained much the same result. See also LaLonde and Maynard (1987) for a summary and discussion of the earlier analyses.

29. See, for example, Stromsdofer et al. (1985), who recommended an experimental evaluation of the next generation of federal employment and training programs—those authorized by the Job Training Partnership Act of 1982—largely on this basis. Others to make the case for experiments over nonexperimental methods included Ashenfelter and Card (1985), Burtless and Orr (1986), and Barnow (1987). For dissenting opinions, see Heckman, Hotz, and Dabos (1987), Heckman (1991), Manski and Garfinkel (1991), and Heckman and Smith (1993).

30. Gueron and Pauly (1991) summarize more than a dozen evaluations of employment and training programs for welfare recipients initiated as controlled experiments in the 1980s. Greenberg and Shroder (1991) provide an even more complete catalog ranging over many years, policy interventions, and target populations (e.g., displaced workers, youth ex-offenders). The preference for experimental research continues unabated into the 1990s, as evidenced by recent decisions at the U.S. Department of Health and Human Services, the Social Security Administration, and the U.S. Department of Labor to fund major experimental evaluations of training programs for welfare recipients, persons receiving disability benefits, and disadvantaged and dislocated workers. See Wiseman (1993) and Bell et al. (1993) for details of the first two initiatives; the Department of Labor studies are just underway and will focus on the national Job Corps program and job search demonstrations in three states.

31. Two further new directions in the recent employment and training evaluation literature do not bear directly on the relative merits of different impact estimation techniques. These concern the synthesis of findings from multiple program evaluations using "meta analysis" techniques (see Greenberg and Wiseman 1992) and the examination of different aspects of multidimensional treatments (see Greenberg, Meyer, and Wiseman 1992).

32. Model specification tests were also advocated by Ashenfelter and Card (1985).

33. Chapters VI and VII of Bloom et al. (1988) provide the original motivation and design for this approach. A more recent version appears in Hotz (1991).

34. Angrist and Imbens (1991) explore a possible bounding strategy for continuous outcome measures, though not one free of assumptions.

35. Manski et al. (1992) illustrate this process.

36. Several of the work-welfare initiatives of the last six years have employed this approach. (See Fishman and Weinberg 1991 for a summary.) Among the most visible is the evaluation of the Washington State Family Independence Program (Long and Wissoker 1992).

37. See Harris (1985), Ginsburg (1985), Orr (1985), and Garfinkel, Manski, and Michalopoulos (1991) for a more extensive discussion of the strengths and weaknesses of comparison site designs in relation to other options.

38. Friedlander and Robins (1992) explore the potential for error through random selection of program and comparison sites using data from the WIN demonstrations of the 1970s. Working with data from multicounty work-welfare experiments, they combine treatment group observations from one set of randomly selected "program" counties with control group data from another set of randomly selected "comparison" counties. The results show that impact estimates are quite sensitive to the particular counties selected, even after controlling for certain preexisting differences between counties and individuals.

39. Bloom (1984b) provides a formula for this adjustment. Angrist and Imbens (1991) specifically advocate this approach to the design of experiments.

40. See Mallar et al. (1982).

41. Leamer (1978, 1982) demonstrated this result with regard to identifying restrictions on two-stage models generally.

2
The Case for Applicant-Based Comparison Groups

Our review of the history of employment and training program evaluations in chapter 1 identified several weaknesses in the most common approach to measuring program effects—using external comparison groups. In this chapter we explore the use of alternate "internal" comparison groups composed of nonparticipating applicants as a means of addressing the key weakness of selection bias.

We begin by identifying the qualities we would require of a nonexperimental impact estimation technique in order to make it a reasonable alternative to randomized field experiments. We then discuss the different subgroups of nonparticipants that might serve as internal comparison groups and how each of these groups differs from participants. A subsequent section reviews prior uses of applicant-based comparison groups and what became of that strategy in recent years. A final section explores ways to overcome several of the limitations of previous applications as we seek to reintroduce the method to the literature.

Desired Qualities in a Nonexperimental Estimation Technique

Three qualities are essential for any nonexperimental method to replace experiments as an accepted means of measuring the effects of employment and training programs. Such a nonexperimental method must be:

- Operationally feasible and affordable across a variety of contexts;
- Free of any important amount of selection bias;
- Widely accepted among policy makers as free of selection bias.

As we saw in chapter 1, the literature on nonexperimental estimation has appropriately emphasized the second of these three criteria by focusing on the risk of selection bias. But this is not, of course, the only factor to be considered.

Much of the debate over experiments concerns their operational feasibility and cost. This category has several components: data collection and analysis costs, logistical difficulties when implementing random assignment in the field, and political and ethical concerns about random exclusions from program services. We will not attempt to summarize the debate surrounding these issues.[1] Instead, we will note its principal lesson: that operational constraints lie at the root of many of the perceived difficulties with experimental analyses.[2] This being the case, it is in the realm of feasibility and cost that nonexperimental methods must improve on experiments if they are to be considered; they will not be able to do so with regard to selection bias.

Often overlooked in the debate is the need for credibility among policy makers who will use research results. It is not enough that an estimation technique be considered free of selection bias in the minds of researchers; it must also be *viewed* as bias-free by policy makers if its results are to be trusted and used. The presumption of accuracy on the part of policy makers, based on a method's "face validity," constitutes one of the key strengths of experimental research. For example, it was this "power to persuade" that propelled the findings of the work-welfare experiments of the 1980s to a pivotal role in the passage of national welfare reform in the Family Support Act 1988. As several commentators have noted, as important as it was for the work-welfare findings to be unbiased, it was even more essential that they be *viewed* as unbiased and, therefore, beyond dispute.[3] An example in the opposite direction is the set of Comprehensive Employment and Training Act (CETA) studies reviewed in chapter 1, none of which produced estimates viewed as credible by policy or research experts and which, as a result, had little consistent influence on policy.

The credibility requirement may in fact pose the greatest challenge to the success of any nonexperimental impact estimation technique. To achieve widespread recognition as a valid approach, a nonexperimental estimator must do what none of the CETA estimators did:

- Offer a compelling reason to believe, *a priori*, that by its very construction it has the potential to overcome the selection bias problem; and
- Demonstrate its validity through favorable comparisons to estimates that are free from selection bias.

In the words of one of the leading exponents of nonexperimental research on education and training programs: "We are looking for a behaviorally motivated and empirically validated nonexperimental impact estimation technique."[4] We will return to this requirement, and the others noted above, as we assess the overall potential of the applicant-based estimation approach in the rest of the monograph.

Potential Nonparticipant Comparison Groups

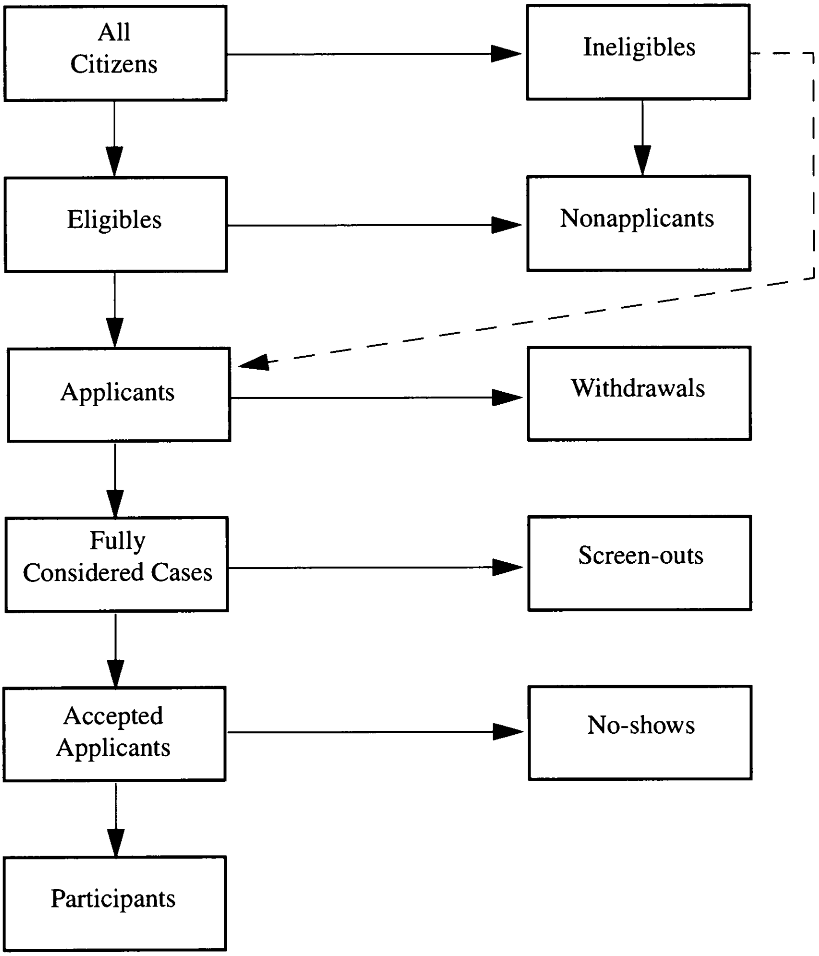
To understand why the use of nonparticipating applicants as comparison group members may meet these criteria, we begin by reviewing the typical employment and training program intake process that generates the nonparticipating applicant population. Exhibit 2.1 summarizes that process and the steps by which some individuals become nonparticipating applicants and others participants or nonapplicants.[5] Of all these groups, only participants receive the classroom training, job search assistance, and other services that define the program, the effects of which the evaluator seeks to measure.

Throughout the program intake process, the pool of potential participants divides over and over again until only the participants remain. At various points of division, groups of nonparticipating applicants are created, which can serve as "internal" comparison groups for measuring program effects. We discuss the strengths and weaknesses of these various groups below.

Nonapplicants

As shown in the exhibit, the first division is between citizens eligible to participate in a given program and those who are not. Typically in government-sponsored employment and training programs, the categorical requirements of eligibility focus on household income levels (e.g., only low-income individuals qualify) and/or current employment status (e.g., only those out of work qualify). Qualifications may also extend to reasons for nonemployment (e.g., plant closing), income transfer program status (e.g., AFDC or food stamp receipt), and/or educational or competency levels.

Exhibit 2.1 Flow Diagram of Entry into Employment and Training Programs



As shown at the second level of the exhibit, both the eligible and ineligible populations divide into applicants and nonapplicants. The vast majority of both groups never apply, with applications from the pool of ineligibles being particularly unusual. Individuals may fail to apply to a program for a variety of reasons other than the realization that they are not eligible. A principal reason for nonapplication among eligibles is that the individual feels no need for the services that the program provides. In employment and training programs, individuals typically apply only when they are unemployed. Since eligibility for training programs is usually based on broader criteria, such as family income, a large proportion of the eligible population may not be unemployed at any given time.[6]

Even when matched on formal eligibility criteria and current labor market status, nonapplicants may differ sharply from participants in their interest in self-improvement and motivation to work—both factors that could substantially affect subsequent earnings. Yet despite these likely differences between nonapplicants and participants, eligible nonapplicants are routinely used as external comparison groups against which program effects are measured. At times, even the information needed to identify eligible individuals—as opposed to nonparticipants in general—is unavailable when forming comparison groups. As a result, evaluators have often had to rely on external comparison groups of nonapplicants—shown near the top of exhibit 2.1—to compare with participants at the very bottom of the exhibit. That these two groups match up well on the factors that determine future earnings (other than program participation) seems unlikely, even if one controls for observed differences in age, education, and previous income.

There are two situations, however, in which eligibles might make an acceptable comparison group. In the first instance, all eligibles may be equal in their motivation and future earnings potential at the point of eligibility, but differ in terms of the timing of that eligibility. Those who become eligible at a time when the program is perceived to have training "slots" available to new applicants apply, while the majority—those who become eligible when availability is believed to be highly limited—do not. Here, it is only the timing of eligibility that distinguishes the two groups, which in steady state may not generate systematic differences in underlying earnings potential.[7] The simpler version of this "rationing" of applications occurs when geographic location limits access to program services for individuals who are otherwise identical to program participants. The use of nonapplicant eligibles as comparison groups could be justified in both instances.

Withdrawals and No-Shows

Some of the same motivational differences will often distinguish participants from other populations defined during intake. Those who withdraw from the program's intake process while their application is being processed—the "withdrawals" shown at the third level of the exhibit—also differ from eventual participants on these factors. So too do accepted applicants who fail to show up for services once admitted into the program—the "no-shows" shown at the fifth level of the exhibit. Even so, the withdrawals at least match the circumstances and motivations of participants up to the point of applying for the program. No-shows parallel participants even further into the process, separating only after acceptance into the program.

Of the two groups, no-shows should be better matched to participants on motivational factors, in that they waited longer—i.e., until after they were approved for program entry—before withdrawing. No-shows also differ from withdrawals in that, like participants, they have passed the screens used by program staff to select program participants.

Screen-Outs

Between these two stages of self-selection comes the crucial step of program selection and exclusion. Most government-funded employment and training programs lack the resources to serve all of the eligible individuals who would like to participate. Thus, some form of program selection is essential to ration program services. The exact form of that selection may vary, but the result is always the same: a group of "screen-outs," individuals at least as motivated to obtain program services as the no-shows—and possibly as motivated as the participants—but who are not allowed to participate. These individuals differ from participants on program-related factors, but not necessarily on the types of personal factors that lead to self-selection among withdrawals and no-shows.

In summary, then, the groups of nonparticipants that might be considered for use as comparison groups, and the principal ways in which they may differ from the participants, include:

- Nonapplicants drawn from the general population, the group used to form external comparison groups. While these individuals may differ from participants in the personal characteristics that influence selection, it is possible to control for many of these in the analysis. Several important characteristics are particularly difficult to control for, however, and may be systematically different for program participants and nonapplicants.
- Withdrawals, who voluntarily withdraw from the intake process after application but before acceptance or rejection by program staff. Because they applied to the program, withdrawals are presumably more like participants than are nonapplicants in the personal characteristics and labor market status that determine self-selection. But they may still differ from participants on some of the factors that determine interest in the program, and on the program selection factors applied by intake staff.
- Screen-outs, who clearly differ from participants in terms of the characteristics that lead program staff to reject their applications. While these individuals resemble participants in their initial decision to seek services, they include some individuals who—like withdrawals—would not have participated were they selected.
- No-shows, who differ from participants primarily in terms of factors (e.g., motivation, finding a job on one's own) that lead some accepted applicants to enter the program once accepted, while others do not.

We refer to the last three of these groups collectively as "nonparticipating applicants."

The History of Applicant-Based Comparison Groups

Interest in nonparticipating applicants as comparison group members is not new, particularly in the field of education research. Beginning with Thistlethwaite and Campbell (1960), a vast literature has emerged on how best to use comparison groups of students excluded from educational programs through program selection—the group we call screen-outs in our discussion. The focus of this literature is on the "regression discontinuity" approach to analyzing screen-out data, introduced by Thistlethwaite and Campbell (1960) and then independently reinvented over thirty years by several other researchers.[8]

Under the most common version of this approach, potential program participants are rated on criteria of suitability for admission to the program and only those with the highest ratings are admitted. The rest—particularly those whose ratings fall just below the cutoff for program admission—serve as comparison group members in measuring the impact of the program on those just above the cutoff. If the ratings are the only systematic determinant of program admission, it becomes feasible to use the ratings as a variable that controls for selection and, in theory, produce an unbiased estimate of the program effect. (Several necessary conditions to obtain this desired result are discussed below.)

This particular use of "screen-outs" as comparison group members is distinguished by the notion—also common to randomized social experiments—that admission decisions can be designed and controlled by the evaluator to ensure reliable impact estimation. To our knowledge, controlled selection processes of this sort (other than random assignment) have never been used to evaluate employment and training programs, although they are common in educational research.

Less premeditated uses of nonparticipating applicants as comparison group members were fairly common among the early evaluations of MDTA reviewed in chapter 1, however. Applicants who dropped out of the intake process before receiving services—both withdrawals and no-shows—were used as comparison group members by Borus (1964), Cain (1968), Stromsdorfer (1968), Borus, Brennan, and Rosen (1970), Hardin and Borus (1971), and Prescott and Cooley (1972), several of which were noted as playing a role in the broader history of nonexperimental methods. In addition, Gibbard and Somers (1968), Solie (1968), and Robin (1969) used applicants who were rejected by program staff—screen-outs—as their comparison groups.[9]

These analysts were generally aware of the potential selection bias problems posed by using nonparticipating applicants as comparison group members, but defended their choice on the grounds that "sometimes they are the only groups available for comparisons" (Borus and Buntz, 1972, p. 238). In surveying their work, Hardin (1969, p. 107) writes:

> [D]efending the kinds of control groups used in Hardin and Borus. . . . I derive comfort from the erratic judgment by the local employment service screening the applicants, failure of notices of enrollment to reach the applicants, temporary illness preventing enrollment in the course, and haphazard ineligibility for training allowances.

Hardin thus appeals to randomness in the selection procedure as the basis for trusting comparison groups of nonparticipating applicants. A general weakness of these earlier studies is that the analysts often did not attempt to control for selection bias by modeling the nonrandom factors that lead to participation among applicants.

The early applicant-based studies and their methodologies have largely escaped notice in recent years. They were not viewed as terribly rigorous or reliable at the time of their release, and the approach they represented fell out of favor in the 1970s. Three factors may have led to their demise:

- Some of the early analyses were, because of data limitations or simple oversight, fairly unsophisticated in motivating and applying the applicant-based approach.[10]
- It is not clear that its early practitioners appreciated or advanced the strong theoretical arguments that can be made for the approach. To our knowledge, we are the first to propose the applicant-based approach as potentially preferable to other nonexperimental methods over a wide range of employment and training contexts.
- Finally, and partly as a result of the first two points, a National Academy of Science report (1974) took a fairly visible stance favoring the use of external comparison groups when evaluating manpower programs, especially those serving prime age male workers such as MDTA. This conclusion reflected both skepticism regarding the value of internal comparison groups in relation to their costs (new follow-up data would have to be collected for applicant comparison groups, but already existed for external comparison groups) and the belief that the selection problem was

not as severe for programs serving prime-age males as for those serving more disadvantaged populations, such as Job Corps.

At the same time, two concurrent developments focused attention on nonapplicant comparison groups drawn from national data bases:

- The increasing availability of large, nationally representative survey data bases containing detailed measures of the earnings of low-income individuals. These included the Current Population Survey, the Continuous Longitudinal Manpower Survey, and the Survey of Income and Program Participation. These data were both more accessible and cheaper than special surveys of nonparticipating applicants.
- The development of elegant and apparently powerful econometric techniques for correcting or avoiding selection bias in a variety of labor market applications. As discussed in chapter 1, these techniques began their ascendancy with the seminal work of Heckman (1974, 1976) and, for employment and training program evaluation, Ashenfelter (1978) and Ashenfelter and Card (1985).

Together, these two developments offered promise for valid program evaluation based largely on existing (and ever improving) general purpose data sets and econometric techniques. As noted above, the underdeveloped method of using applicant-based comparison groups was left by the wayside.

The emphasis in nonexperimental employment and training program evaluation has since turned completely away from the use of applicant comparison groups. Beginning with Ashenfelter (1978) and Kiefer (1979), it shifted to modeling the overall selection process depicted in exhibit 2.1—distinguishing eligibles as a group from participants—and the implications of that distinction for future earnings.

Before closing our discussion of applicant-based comparison groups, it is worth noting that they have been used sporadically over the last fifteen years in other fields besides educational research.[11] Of particular relevance is a volume by Collignon et al. (1989) providing an extremely thorough discussion of alternative applicant-based comparison groups for measuring the effects of state vocational rehabilitation programs. Both the theoretical and empirical appeal of alternative comparison groups are explored, although the authors do not have the benefit of an experimental control group with which to test the various options (as we do here). Largely on theoretical grounds, then, they conclude that—in the vocational rehabilitation context—some applicant-based comparison groups should provide a useful lower bound on true program impacts when used as a benchmark (e.g., applicants screened out as not sufficiently disabled to merit rehabilitation services), while others should provide a complementary upper bound (e.g., those screened out as too disabled to rehabilitated). The tightness of these bounds is not considered, however, nor is the question of whether similar groups of screen-outs can be found for this purpose in evaluating standard employment and training programs.

Reviving and Strengthening the Applicant-Based Approach

The principal purpose of this monograph is to reintroduce nonparticipating applicants as a comparison group for evaluating voluntary employment and training programs.[12] In this section, we present our conceptual argument for why comparison groups composed of nonparticipating applicants may be able to overcome the problem òf selection bias. In the process, we establish that the applicant-based approach has the capacity to meet the other two criteria for successful nonexperimental methods introduced earlier in the chapter: operational feasibility and an *a priori* plausibility for obtaining unbiased estimates. We provide modest but promising evidence on the final criterion, empirical reliability, in chapters 3 and 4.

Possible Advantages of the Approach

We begin by revisiting the program intake diagram in exhibit 2.1. We argued earlier that later attriters from the intake process—withdrawals, screen-outs, and no-shows—may correspond much more closely to participants, on both observed and unobserved characteristics, than the more commonly used comparison group of eligible non-applicants. One expects that, simply by applying to the program, nonparticipating applicants have revealed themselves to be more similar to the participants than are nonapplicants in such critical dimensions as labor market status, the transitory component of earnings, and motivation to seek training.

The reverse could also be true—i.e., nonapplicants could make a better comparison group than either withdrawals or no-shows.[13] Suppose that nonapplicants are a mix of those who expect to find good jobs on their own and those who do not. Suppose also that individuals who apply for assistance further sort themselves between those who enter the program and those who drop out of the intake process. Suppose that individuals who succeed in finding employment on their own become withdrawals and no-shows, while less successful applicants become participants. Nonapplicants—some employed, some not—fall between these two groups, making them a better comparison group for participants than withdrawals or drop-outs.

While not implausible, the above situation seems to us less likely than one in which participants match more closely to nonparticipating applicants than to any other group on the transitory elements of earnings and motivation that determine immediate postprogram earnings. This is not a foregone conclusion, however, making it essential that our theoretical arguments in favor of nonparticipating applicant comparison groups be subjected to tests of their empirical validity in later chapters.

Of particular importance to the *a priori* argument on behalf of applicant-based comparison groups is the fact that nonparticipating applicants are likely to be well matched to the participants on the time path of earnings just prior to application. Like participants, most nonparticipating applicants will have experienced an earnings loss that prompted them to apply to the program. In contrast, eligible nonapplicants with similar earnings may have chronically low incomes (and a disinterest in employment and training services) and thus will be in steady state rather than in the midst of a transition when selected for inclusion in a comparison group.

If the preprogram dip in earnings for participants signifies a permanent break, the same type of break should be evident to some extent in the subsequent earnings of nonparticipating applicants. There is no reason to expect either of these patterns to be especially strong for eligible nonapplicants, if our hypothesis of sustained low income for those individuals is correct. In this case, applicant-based comparison groups will better mimic the earnings of participants than would nonapplicants. If this is the case, reliance on applicant-based comparison groups will avoid the problems of trying to adjust comparison group earnings for individuals in steady state to match the without-program earnings of participants in steady state.[14]

Even if nonparticipating applicants match well to participants on the transitory aspects of earnings, they may still differ from participants in two other, potentially important, respects:[15]

- Withdrawals and no-shows selected themselves out of the intake process before acceptance and/or entry into training. This may mean that they are less motivated, capable, or willing to work than participants. Or it may mean that, unlike participants, they found an attractive job opportunity before they could begin training.
- Screen-outs were selected out of the intake process by program staff. If the screening criteria used by program staff are related to future earnings, the subsequent earnings of screen-outs may not be a good proxy for what participants would have earned in the absence of training.

Controlling for Self-Selection

To deal with the first of these problems—self-selection among program applicants—the initial step is to control for observable baseline characteristics such as age, sex, and ethnicity. While common in studies using external comparison groups, controls of this sort were not always present in the early applicant-based evaluations of Manpower Development and Training Act. Even with these variables included in the model, however, an unknown degree of selection bias may still occur due to selection on unobservable characteristics such as motivation and inherent ability. Traditionally, evaluators using external comparison groups have attempted to correct for these differences by adjusting for differences in preprogram earnings (see, for example, Ashenfelter 1978), or by explicitly modeling the selection process as part of a two-stage estimation process (see Barnow, Cain, and Goldberger 1980).[16]

As the discussion in chapter 1 makes clear, these methods are not always successful in eliminating systematic differences between the participants and the comparison group. In the absence of a randomly assigned control group as a benchmark, however, one can never know how successful they have been in any particular application. Our strategy for minimizing the selection bias problem is to choose a comparison group that is as similar to the participants as possible in terms of its circumstances and motivation at the time of application to the program. We believe that nonparticipating applicants provide such a comparison group. Nevertheless, because of the differences between nonparticipating applicants and participants noted above, there is no guarantee that they will perform as well as a randomly assigned control group—or even that they will constitute a better comparison group than nonapplicants, although their similarities to participants argue strongly in their favor.

Controlling for Program Selection

The second problem—program staff selection of suitable participants—is, at least in principle, more tractable from an analytic standpoint.[17] Screen-outs differ from participants on factors that are necessarily external to the individuals involved. Whether acting on conscious or unconscious considerations, intake workers—if they are systematic at all—necessarily work from external signals when deciding whom to admit and whom to exclude from their programs, signals that are, at least in principle, also discernible by evaluators.[18]

Some of these signals should be captured by standard background variables like age, level of education, and prior work experience. Presumably, these predictors of later labor market outcomes and ability to benefit from training substantially influence which individuals intake staff see as most suitable for or most in need of program services.[19] Thus, these variables alone will help to remove preexisting differences in earnings potential between screen-outs and participants that might otherwise be confounded with program impacts.

To deal with any remaining differences between screen-outs and participants, we propose an approach that was introduced to the educational literature concurrent with the beginning of employment and training evaluation. In their seminal article on regression discontinuity analysis cited above, Thistlethwaite and Campbell (1960) set out a model that is explicitly designed to deal with the fact that screen-outs differ systematically from participants on those characteristics that lead program staff to reject certain candidates. They point out that if these characteristics can be captured and added to the model, an unbiased estimate of the impact of the program can be obtained, at least for those near the "cutoff point" between participants and those screened out of the program. (They conceptualize the selection factors as a single index variable; we adopt this formulation for convenience of exposition, although it is not essential to the model.) We now examine this model in detail.

The Regression Discontinuity Model for Screen-Out-Based Impact Analysis

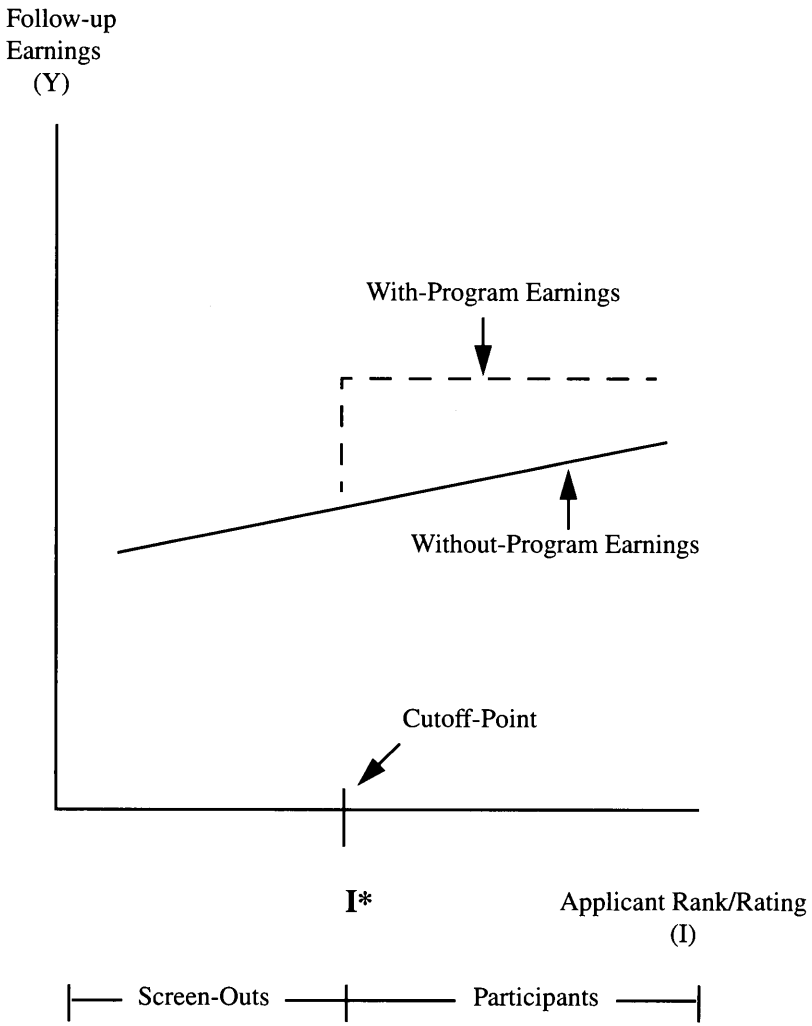
Exhibit 2.2 shows how the regression discontinuity approach to impact estimation works—and how it got its name. At its core, the technique represents a promising way of estimating what the earnings of program participants would have been absent the program, given earnings data on nonparticipants—in this case, screen-outs.

Projecting the Without-Program Earnings of Participants

Exhibit 2.2 depicts graphically the relationship between intake staff ratings of the suitability of applicants for admission to the program and applicant earnings. The postprogram earnings of participants and screen-outs appear on the vertical axis of the graph; the index used to rate applicants and make screening decisions is shown on the horizontal axis.[20] In the simplest case, program staff establish a cutoff index value, I*, for entry into the program. The cutoff value divides the applicant population intǫ screen-outs (those with scores to the left of I*) and participants (those with scores to the right of I*).

The solid line in the exhibit shows how much earnings will rise with the ratings absent entry into the program. The earnings of the screen-outs in the postprogram period determine the slope of this line to the left of I*. This segment can be estimated with observed data. Above I*, however, there are no "without-program" cases to observe, since all applicants with index values greater than I* have been accepted into

Exhibit 2.2 Diagram of Regression Discontinuity Analysis



the program. The diagram assumes that the linear relationship shown below I* continues above I*.

The postprogram earnings of participants may be higher than this benchmark level, because of the impact of the program. This enhanced level of earnings is denoted by the broken line to the right of I*, which appears above the solid line. The space between these two lines—the broken line representing participant outcomes with program services and the solid line representing participant outcomes without services—is the impact of the program.

Now consider how one might estimate the relationships shown in the exhibit. The solid line to the left of I* and the broken line to the right of I* can be estimated from observed data on screen-outs and participants, respectively. A linear specification gives the pattern shown here, although nonlinear specifications could be considered. In the linear case, postprogram earnings of screen-outs and participants are regressed on a constant, the index I, a dummy variable for participation, and the dummy for participation interacted with I.[21] The first two terms from this equation determine the height and slope of the screen-out segment of the earnings line (the solid line to the left of I*), while the last two terms shift those parameters to create the participant portion (the dashed segment to the right of I*). The vertical "jump" between these two line segments at I*—the discontinuity in the regression line—provides an estimate of the effect of participation on the marginal participant (the individual with index value I*).

To obtain measures of the impact of the program on other participants—those with index values greater than I*—one must extrapolate the relationship between the index and the "without-program" earnings of screen-outs to index values *above* I*. We observe this relationship below I* (i.e., for screen-outs) but must simulate it above I*. This is done in the exhibit by extending the solid without-program earnings line to the right of I*, on the assumption that the linear relationship between earnings and index values observed for screen-outs would have applied to participants had they too been kept out of the program. Other nonlinear extrapolation methods are possible, though all necessarily involve extension of without-program patterns among screen-outs to index values for participants.

Given the centrality of the index variable in this exercise, it is important to recognize that the index need not be a perfect correlate for future earnings. Rather, it is a measure of the applicant's suitability for admission to the program (as judged by intake staff), *which may or may not correlate with applicant's future earnings*. If it does not, and the index is the sole basis for program admission, we have no selection bias problem. More likely, some correlation (either positive or negative) is present, and the index becomes a tool for measuring program impacts.

While it may seem like a tall order to define a single index that captures all the considerations that went into the decisions of program staff to accept or reject specific applicants, it is in principle quite feasible. As noted by Campbell and Stanley (1966), a simple ranking of the N applicants from 1 (most acceptable) to N (least acceptable) by program intake staff would capture all of the relevant considerations and make additional objective measures such as education and prior work experience superfluous. Similarly (though perhaps less obviously), if program staff were to rate each candidate's suitability for training on a scale of 1 to 100, such ratings should also capture all of the information used in selection decisions, since staff would presumably never assign a higher rating to an applicant they rejected than to one they accepted.[22]

Potential Limitations of the Approach

There are some potential limitations to the regression discontinuity approach, however.[23] First, as noted above, it requires assumptions about the relationship between the selection index and the potential without-program earnings of participants (individuals above I* on the selection index). Depending on the accuracy of these assumptions (e.g., whether the true relationship really is linear as assumed in exhibit 2.2), the technique may give valid impact estimates only at the boundary between acceptance and rejection to the program (i.e., at I*). More generally, one's confidence in the impact estimates it produces will depend on how close to the cutoff the participant observations are. Estimates become increasingly dependent on the assumptions used to project without-program earnings trend of screen-outs to individuals above the cutoff line as one moves farther above that level. There is no guarantee that the relationship observed below I*—whether projected linearly as in the exhibit or nonlinearly as suggested by possible nonlinear patterns in the data to the left of I* (not shown)—would continue above the cutoff point.

Assuming that there are no discrete jumps in without-program earnings as index values change,[24] all smooth projections beyond I* will give fairly similar, uniformly reliable impact estimates for those just above the cutoff. Different approaches may give quite different answers further to the right, as the various means of projection diverge.

The regression discontinuity approach is critically dependent, then, on two assumptions: (1) that the selection index variable fully controls for the systematic determinants of participation in the program, and (2) that the relationship between the selection index and without-program earnings can be reliably estimated from the earnings of screen-outs.

The approach is also dependent on obtaining complete, consistent data on applicant rankings.[25] As noted above, ratings of applicants on a common numeric scale will serve just as well as rankings for this type of analysis, and may be the only feasible option in programs whose intake period extends over many weeks or months (precluding the joint ranking of all applicants for the purpose of making admission decisions). But dependence on either type of index carries certain costs. First, one must arrange to collect rating data on all applicants, or for as many as possible up to a reasonable sample size. (See below for a discussion of the operational feasibility of the method). Also, unless instructed to do so, intake workers cannot all be expected to use the same cutoff point on a ratings scale to determine who enters the program.[26] Or, in what amounts to the same thing, different workers may adopt different admission standards and then score candidates so that, in each instance, those who just meet a rater's personal standard receive the preordained cutoff score of I*. Nor is the cutoff point adopted necessarily the same at all points in time, depending on temporal ups and downs in the supply of (and demand for) training services.

Ironically, arbitrary variations in admission standards will strengthen, rather than weaken, the regression discontinuity analysis. Exceptions to the cutoff rule provide observations on screen-outs both to the left and right of the I* cut-off point in exhibit 2.2. If sufficiently numerous, these observations can be used to test how well the discontinuity analysis projects without-program outcomes into the participant portion of the scale or, better still, may serve as the basis for actually estimating that portion of the line from observed data. In the extreme, if the cutoff point varied randomly from one applicant to the next, admission to the program would be random and screen-outs would be equivalent to a randomly assigned control group.[27] In practice, these data points are likely to provide a less complete set of without-program observations above I* than would an experimental control group and will not necessarily represent that portion of the population on a statistically valid basis.[28] Nevertheless, they unquestionably improve the information available to the evaluator regarding the key with-program/ without-program contrast between otherwise comparable individuals.

The mere occurrence of screen-out observations above I* is not sufficient to guarantee better information with which to measure impacts, however. The overlap of participants and screen-outs along the ratings scale must be due to true differences in the standards applied for program admission, having accurately recorded the ratings. It cannot be the result of variations in how ratings are recorded for individuals perceived as equally qualified for admission yet rated differently. In that case, the variation in ratings would tell us nothing about the selection process itself, only about errors in the ratings data.

A more complex problem arises when two intake workers use not just a different cut-off standard on a common rating scale, but two entirely different scales based on *different sets of rating factors.* For example, one worker might try to admit those applicants who look like they can do best in the labor market *with* the program's help, while another worker admits those who would do worst *without* help and who therefore need help more. As a result, the first worker would rate highly and admit applicants with relatively high without-program earnings (the case shown in exhibit 2.2), while the other worker does just the opposite.

One response to this situation would be to analyze the intake process separately for each intake worker. However, this would substantially reduce the statistical precision of the analysis, even when combined across workers. Alternatively, one could perform a combined analysis to estimate the (weighted) average of various selection rules. This would produce meaningful impact estimates for the overall program only if the *average* trend in earnings across a variety of indexes—estimated within the screen-out sample—provides an unbiased projection of *average* without-program earnings of participants at higher levels of the index scale.[29]

In light of all these possibilities, the most difficult problem facing evaluators using the regression discontinuity approach may be *knowing how to interpret the ratings data* at their disposal. Here, there is simply no substitute for ensuring that the intake process is either successfully controlled and monitored by the evaluator in all times and places, or else adequately documented at a level of care and disaggregation not normally considered by program evaluators.

Desired Qualities: How Does the Screen-Out-Based Approach Fare?

At the outset of the chapter, we posited three desired qualities in any nonexperimental approach to measuring the impact of employment and training programs: operational feasibility and affordability, freedom from selection bias, and general acceptance as unbiased based on past performance and *a priori* plausibility. Here, we assess the potential of the screen-out-based method of analysis with regard to these factors.

The second and third qualities are related, inasmuch as true freedom from selection bias and the role of past performance in generating the method's acceptance can only be assessed once the approach has been used and tested. These tasks we undertake for the first time in chapters 3 and 4. Here, we confine our attention to the *a priori* plausibility and operational feasibility/affordability of the method, plus a closing comment on the replicability of the testing process initiated in subsequent chapters.

A Priori Plausibility

To offer a viable alternative to experimental research, an applicant-based impact estimation strategy must not only avoid serious selection bias, but do so by methods that are conceptually plausible across a variety of employment and training contexts. In this chapter, we have attempted to demonstrate both the theoretical appeal and the limitations of the various applicant-based strategies. In our judgment, comparison groups composed of screened-out applicants, combined with data on intake workers' assessments of suitability for training, have stronger theoretical appeal than the external comparison groups traditionally used in nonexperimental evaluations.

Our real goal here is not to argue for one nonexperimental method above another when both are flawed, however, but to assess the overall strength of the case for accepting screen-out-based impact estimates as free of substantial selection bias in most applications. The one impact estimation methodology where this case has been made successfully to date is that of controlled random experiments. It is instructive, therefore, to make a conceptual comparison between ratings-adjusted screen-outs and the comparison group used in experiments: randomly excluded but otherwise acceptable program applicants. The contrast to screen-outs is immediately apparent: screen-outs are nonrandomly excluded and unacceptable program applicants. This points to two major points of contrast, one essentially a "red herring" but still important in an assessment of the method's *a priori* plausibility, and the other much more fundamental:

- Unlike random assignment, the selection and exclusion of screen-outs are not directly controlled by the evaluator; and
- Program staff rationing of training slots among interested and eligible applicants contains both random and systematic elements, rather than strictly random elements as in random assignment.

We first consider the reasons why evaluators might want to control the intake process and, in light of those reasons, why the absence of control should not pose a concern regarding the plausibility and face validity of the method. It seems clear from our discussion of the regression discontinuity approach that the advocates of evaluator-controlled selection (see, for example, Campbell and Stanley 1966) have two purposes in mind: uniformity and "knowability" of the intake decision rule. Both of these qualities obviously apply to a random assignment intake process and to any other evaluator-controlled selection approach. But, as Thistlethwaite and Campbell (1960) first observed, they can also be attained by letting program operators follow their own predilections when selecting participants as long as both the selection criteria used and the information to which those criteria were applied are fully specified to the evaluator.

Thus, in principle, any screen-out-based evaluation can meet these two requirements, if planned in advance of program intake. All that is required is for the evaluator to work with program operators to elicit their preferred selection approach and develop an overall rating scale that reflects that approach for individual applicants. A regression discontinuity analysis follows immediately, regardless of whose selection rule is followed. This approach has the side benefit of allowing programs to be run on their own terms, which should make screen-out-based evaluations much more salable to program operators than the highly demanding selection requirements of randomized field experiments. Documenting, rather than dictating, the basis for program selection may be the *only* option available when evaluating an ongoing program; there, the exact mechanisms for selection are part of the program to be evaluated "as is" and cannot be dictated by the researcher as in a demonstration project.

Even where it is not possible to assure well-documented screen-out procedures and data inputs in advance, a reasonable approximation to that result may be possible after the fact as long as intake workers' self-defined ratings of applicants are collected from the outset. This is in fact the situation we face in using data from the AFDC Homemaker-Home Health Aide Demonstrations to test the screen-out approach in later chapters. The effectiveness of this *ex post* approach, explored more fully in chapters 3 and 4, may not yet fully illustrate the potential of more deliberate attempts to employ screen-outs as comparison group members when undertaken *ex ante*.[30] We will return to the question of how this might be done in our concluding chapter.

The distinction between random and systematic selection is more fundamental. In comparing random assignment to applicant-based methodologies in this regard, it is again helpful to refer to the diagram in exhibit 2.2. As noted earlier, the one difficulty (apart from possible limitations of the ratings variable) that this approach entails is the need to project rather than observe without-program outcomes of higher-ranked applicants. Random assignment solves this problem by excluding a random share of applicants *at all ratings levels* from the program, effectively tracing out all points along the solid without-program earnings line to the right of I* using observable data on controls. This most fundamental—and potentially sizable—advantage of experimental analyses over regression discontinuity modeling cannot be overcome by any program-defined selection rule, since program operators will always have preferred applicants under any approach or rule for rationing training slots. Barring unintended departures from that rule (see above), by definition all favored applicants will appear to the right of I* and be observed only as program participants.

In the end, then, the *a priori* plausibility of the screen-out-based approach rests on one's faith in how well it projects the without-program earnings line beyond the point on the ratings scale where one ceases to observe without-program individuals. If one thinks the projection has been accomplished within an acceptable margin of error, one should believe that the systematic differences between screen-outs and participants which distinguish this approach from a randomized experiment have been removed. If not, one will have to judge the method's *a priori* appeal not against the experimental norm but in relation to the plausibility of other nonexperimental methods that try to equalize participants and comparison group members by other means.

The accuracy of the projection of without-program earnings into the range of the index where only participants are observed depends in large part on the evaluator's ability to choose the correct functional form for the regression equation. If the relationship between the index and without-program earnings is close to linear, a standard linear regression specification may provide a reasonable projection into this range. If, however, this relationship is highly nonlinear in that range, a linear projection could yield a very inaccurate measure of what participants would have earned in the absence of the program. If nonlinear relationships are evident below the cutoff level for intake (i.e., for screen-outs), this can be projected instead, but again succeeds or fails according to the reliability of this projection above the cutoff.

One more comment on the theoretical appeal of the screen-out approach should be noted before turning to a discussion of the method's operational feasibility. Even if we assume that regression discontinuity analysis will adjust screen-outs to match non-screen-outs in all relevant respects, a portion of the selection bias problem remains. Some of the individuals admitted into the program through staff screening may elect not to participate and instead become no-shows. This final self-selection process pares down the participant group on a nonrandom basis, leaving a subset of participants who do not match screen-outs as well as the original, complete set of accepted applicants did.

If screen-out-based comparison groups have trouble matching up to the participants who remain following the exodus of no-shows, the same will also be true of experimental control group members. Like screen-outs, control group members are removed from the intake flow at the point of admission to the program and are fully comparable only to the full set of accepted applicants (no-shows plus participants). We discuss in the next chapter how this problem is addressed in the context of experimental research; the same approach could be applied to a screen-out comparison group if it was thought to be otherwise comparable to no-shows and participants as a group. Thus, this aspect of the selection problem does not represent a shortfall of the screen-out-based approach in comparison to the experimental standard.

Feasibility and Affordability

A final consideration concerns the feasibility and affordability of applicant-based comparison groups in general, and screen-out strategies in particular. Here, the experience of the National JTPA Study[31] and other social experiments is instructive. Routinely, and with little hesitation on the part of program operators, evaluators add a form to the program intake process on which applicants record their baseline characteristics, and intake staff subsequently document their own decisions and actions with regard to the applicant. Nearly 30,000 of these forms were collected in 96 local offices as part of the National JTPA Study—at a modest cost and with minimal missing data rates.

In this and most other program contexts, nearly all of the information required on the evaluation intake form would be collected anyway for operational purposes. Thus, the evaluation imposes little added burden on the program; in fact, in some instances, it simplifies and improves data collection to the point that program operators adopt the evaluator's form as a permanent part of their intake process once the evaluation ends.[32] And as long as it is routinely administered at the point of application, such an intake form should identify a sufficient pool of nonparticipating applicants and participants for research purposes in most contexts.[33] As we shall see in the next chapter, the break-down of the former group into withdrawals, screen-outs, and no-shows is easily accomplished by the inclusion of just a few items on the form to be completed by program staff.

The other essential requirement for applicant-based impact analysis is accurate follow-up data on both participants and nonparticipating applicants. Historically, employment and training program evaluators have relied on follow-up surveys to measure participant earnings following program entry, and—for experiments—to obtain corresponding measures for control group members. The latter requirement has been viewed as one of the drawbacks of experimental designs, since special control group surveys are unnecessary in nonexperimental analyses based on external comparison groups drawn from existing national data bases such as the Current Population Survey, the National Longitudinal Survey, or the Survey of Income and Program Participation. It would also seem to apply to applicant-based analyses, where special-purpose data collection is again unavoidable for the group that is to be compared to participants.

Increasingly, however, evaluators are turning to much less expensive administrative sources for earnings information, for all types of comparison groups and for both the preprogram and postprogram periods. The National JTPA Study, for example, compiled earnings data on treatment and control group members from three sources: follow-up surveys, state unemployment insurance wage records, and (in grouped form to protect confidentiality) Internal Revenue Service (IRS) employer reports.[34] The Bloom et al. (1993) study shows that both of the latter administrative data sources provide earnings impact estimates similar to those produced by the traditional survey method across broad populations. Our work with IRS data for the homemaker demonstration sample (see chapter 3) confirms this result in an entirely different sample and in relation to a different type of follow-up survey.

Given their low cost, broad coverage, and proven reliability, these administrative records systems may be a particularly cost-effective source of follow-up data for applicant-based impact evaluations for many years to come. And, by removing the requirement of costly survey data collection for participants as well as comparison group members, they make applicant-based evaluations even more affordable than past evaluations that took only comparison group data from existing (and therefore low cost) national data bases. Given the essentially zero cost of obtaining additional observations from administrative records systems, using such data to track comparison group members once they are being used to track participants should reduce total evaluation costs to a level below that of just the special-purpose (participant) survey requirements of past evaluations using external comparison groups.[35]

Replicability

Replicability is also a key requirement for the applicant-based approach, both in its testing and in subsequent applications. Like any nonexperimental approach that aspires to broad acceptance, the applicant-based strategy will need to be tested in more than one setting where experimental data are also available as a validation benchmark. The tests of applicant-based comparison groups presented here constitute only a single observation on their performance; they will not provide a definitive judgment as to whether such methods are an acceptable substitute for experiments. Only by replicating this type of analysis many times in many different contexts can such a judgment be made.

Fortunately, the fact that data on applicants used here first emerged as an accidental byproduct of an experiment suggests that this type of analysis could readily be conducted in other settings. The large number of experiments currently in the field or in the planning stage provide an opportunity for a broad range of validation tests across a variety of contexts over the next five or ten years.

We present the first such test in the next two chapters.

NOTES

[1] See, for example, Burtless and Orr (1986) and Manski and Garfinkel (1991).

[2] The operational difficulties facing experiments may do the most harm in confining them to nonrepresentative samples of program participants. This restriction calls the findings of many experiments into question even after selection bias has been removed through random assignment. (Not all experiments suffer from this difficulty; the Food Stamp Employment and Training Program evaluation described in Puma et al. 1990 applied random assignment methods to a nationally representative sample). Other common criticisms of experiments stem from the feasibility constraint: unlike evaluations based on natural variations within national data bases, experiments are said to lack flexibility when addressing a range of policy questions, and to distort outcomes in relation to a permanent national program when limited to a few localities and/or years. While the first of these criticisms may not be valid—randomization *creates* a new population for study (the

control group) without destroying any existing populations, thus increasing research flexibility—the second surely is and results directly from operational restrictions on the scope and duration of experiments.

3. See the articles collected in Wiseman (1991).

4. James Heckman, January 5, 1993, to the annual meeting of the American Economic Association in a panel entitled "Evaluating Employment and Training Programs Using Nonexperimental Methods."

5. Maddala (1983, p. 266) also frames the problem of self-selection in program evaluation in terms of such a diagram.

6. For example, Sandell and Rupp (1988) found that while 68 percent of participants in Job Training Partnership Act (JTPA) programs are unemployed at application, only 12 percent of the eligibles fell into this category. The remaining eligibles are either employed or not in the labor force.

7. We are indebted to Stephen Kennedy for pointing out this possibility.

8. Campbell notes several of these rediscoveries in the forward to Trochim (1984). Examples from the econometric literature include Goldberger (1972, 1980) and Cain (1975). The current monograph also stems from the independent development of the technique by Orr and Bell specifically for the purpose of evaluating employment and training programs. The approach is discussed in detail in Campbell and Stanley (1966) and Cook and Campbell (1979) and receives book-length treatment in Trochim (1984).

9. A more detailed review of these early studies is available from the authors on request, including an assessment of how well the different applicant-based approaches were implemented. For present purposes, the research merit of these studies is of less interest than the place they occupied in the overall history of nonexperimental evaluations of employment and training programs.

10. For example, the variables used to control for preexisting differences between participants and nonparticipating applicants were fairly limited or nonexistent in several studies.

11. For example, screen-outs are used by Bloom and Singer (1979) as a comparison group for evaluating the effects of alternative prisons and by Bound (1989) for evaluating the effects of disability insurance (also discussed in Bound 1991 and Parsons 1991). Dean and Dolan (1991) use no-shows to evaluate vocational rehabilitation services, while Berger and Black (1992) rely on a special group of applicants not normally available to evaluators of employment and training programs—those on the waiting list for services—to measure the effects of child care services.

12. Mandatory employment and training programs do not formally allow for nonparticipation among those who fall within the program's mandate, either at the discretion of the individual involved or based on decisions by program staff, although nonparticipation cannot be avoided in a free society. We do not discuss how to deal with this type of nonparticipation. Note, however, that programs of this sort, such as the Job Opportunities and Basic Skills (JOBS) program for able-bodied AFDC recipients, may still apply the participation requirement selectively when program openings must be rationed due to limited capacity. If so, much of what we discuss here in terms of voluntary employment programs applies to those programs as well.

13. We are indebted to Stephen Kennedy for bringing this possibility to our attention.

14. It is possible, of course, to match an external comparison group on changes in earnings in the period immediately prior to program entry, mimicking the preprogram dip. Several of the CETA studies used comparison groups drawn from the Current Population Survey matched on this basis (see Westat 1981, 1984). Because there is no assurance that the dip represents the same mix of permanent and transitory changes in earnings for both groups, however, such matching provides little assurance that the two groups are well matched on the *transitory* component of earnings, and therefore on subsequent total earnings.

15. Collignon et al. (1989) explore these same two factors when considering the comparability of various groups of nonparticipating applicants and participants in state vocational rehabilitation programs.

16. See chapter 1 for a detailed discussion of these methods.

17. Angrist and Imbens (1991) make the same point, but for different reasons, in their paper on the use of instrumental variables to address the selection bias problem. They say at the outset that, in contrast to previous econometric analyses of the selection bias problem offered by Heckman (1990) and Chamberlain (1986):

> An important distinction, and an underlying theme of [our] paper, is the difference between identifying information derived from models of program participants' behavior and from information about program eligibility rules. We argue that the latter is more likely to provide a convincing empirical identification strategy (p. 4).

Unlike the argument presented here, the Angrist-Imbens approach hinges on the existence of a group of individuals who have zero probability of participating—a group they believe is more likely to be created by program exclusions than by individual tastes for employment and training services (pp. 13-16).

18. The only alternative—that program exclusions are totally arbitrary (i.e., not based on any external signals), as suggested by Hardin (1969)—presents an even greater opportunity for unbiased impact estimation. Arbitrary exclusions would prevent participation from correlating with future earnings at all, and would therefore preclude any degree of selection bias. Under this scenario, screen-outs would be the ideal comparison group, differing from participants strictly at random, as do experimentally designed control groups.

19. Several studies have identified some of the factors used to screen applicants to employment and training programs, many of which tend to recur across a variety of program settings. The most consistently influential variables are prior education and recent work experience (positively related to selection) and age and current employment (negatively related to selection). See, for example, Anderson, Burkhauser, and Raymond (1992) and Sandell and Rupp (1988), on selection in Job Training Partnership Act (JTPA) programs, and Rupp, Bryant, and Mantovani (1983) on selection into the earlier CETA programs. An earlier study of special relevance is Bell and Orr (1988), which examines screening factors in the demonstration projects analyzed in chapters 3 and 4.

20. We assume here that those with higher index values earn more in the follow-up period at all points on the scale. This assumption is not essential, however; what is essential is that follow-up earnings vary in some fashion with the value of the index.

21. In practice, the regression could also include control variables for the measured characteristics of the sample, including prior earnings. We abstract from these factors here to focus on the critical role of the index in the regression analysis.

22. While true for individual intake workers (assuming that their cognitive processes and admission standards are consistent across applicants), this point does not necessarily hold for different workers or over different program intake periods. We discuss the implications of these exceptions below.

23. Potential problems with the method have been explored at length in the education literature, most recently in Stanley (1991), Reichardt, Trochim, and Cappelleri (1992), and related articles. Here, we make an independent assessment of the potential limitations of the method when evaluating employment and training programs.

24. This assumption, too, could be violated, although the odds of a discrete jump occurring right where it would most interfere with impact estimation—at the cutoff point for program admission, I*—seem fairly remote.

25. Also, in what may be an easy requirement to meet, some variation in ratings among screen-outs is essential in order to project follow-up earnings as a function of those ratings into the participant portion of the range.

26. For example, Worker A might admit all candidates rated above 50, while Worker B admitted all candidates rated above 30. Thus, one worker would admit candidates rated 31-50, while the other would reject them.

27. Admission to the program would also be random, and screen-outs equivalent to a true control group, if the ratings themselves were completely capricious.

28. If some applicants above I* are admitted while others are not, it is likely that admissions in this range of the index will correlate with both index values and outcomes; if so, these observations may provide biased estimates of the relationship above the cutoff. The only exception would be if violations of the cutoff rule occur on a haphazard, unintentional basis, providing essentially a random (though very small) subsample of applicants above I* who do not participate. Exceptions due to variations in admission standards among workers and over time are more problematic, but still potentially quite valuable. For example, if the composition of applicants is fairly constant over time and applicants are assigned to intake workers without regard to their (the applicants') characteristics, screen-outs above I* should be a random sample of all the applicants in that range. (I.e., they would represent the particular applicants who happened to encounter workers and/or time periods for which the cut-off level was above I*.) While these screen-outs would not cover the entire range of ratings (since all workers and times would probably exercise some level of selectivity), they could still be used to estimate without-program earnings over the range in which they did occur.

29. One could easily imagine a case where an unbiased projection of this sort would not result, depending on the curvature of the without-program earnings lines for the various intake workers. Under such circumstances, the analysis should be kept separate for the different workers and the independent projections (for the respective participant samples) averaged at the end of the process. This is computationally equivalent to running separate impact estimates for each intake worker and then averaging the estimates.

30. In particular, the analyses in chapter 3 might have been stronger had the demonstrations adopted a more detailed rating scale or collected data on which intake worker rated each applicant.

31. See Bloom et al. (1993).

32. This was the case in several of the sites in the AFDC Homemaker-Home Health Aide Demonstrations and the National JTPA Study.

33. Obviously, applicant-based comparison groups cannot provide valid impact estimates when very few applicants drop out or are screened out of the intake process. This seems rarely to be the case in an age when government resources are severely constrained and most employment and training programs oversubscribed. An insufficient number of nonparticipating applicants is not the reason large nonparticipating applicant samples are unavailable for most of the National JTPA Study sites; there, data limitations resulted from administering the form too late in the intake process, a constraint that other evaluations can easily avoid if they are (as that one was not) designed with applicant-based comparison groups in mind.

34. The social security earnings records used in several evaluations are derived from the same source as IRS data—employers' reports of individual workers' earnings. Unlike IRS data, however, they have the limitation that only earnings below the Social Security tax limit are recorded in this data source. This limitation is much less serious now than it was when the tax limit was much lower.

35. The same is true of the costs of social experiments.

3
Estimating Program Effects in the AFDC Homemaker-Home Health Aide Demonstrations

This chapter describes an application of applicant-based nonexperimental impact estimation techniques to data from the AFDC Homemaker-Home Health Aide Demonstrations. It begins with a brief description of the homemaker demonstrations and their data, followed by an explanation of the methodology we employ and a detailed account of our findings.[1]

Our analysis of the homemaker data focuses on the roles of alternative comparison groups and intake workers' ratings of applicant potential in forming nonexperimental estimates of the program's impact. We therefore do not explore all of the econometric modeling techniques available in the literature (see chapter 1) that might be applied with these two basic inputs. Instead, we use basic multiple regression methods to highlight the role of these two key inputs.

The Homemaker Demonstrations and Their Data

The data employed in our analysis come from Abt Associates' evaluation of the AFDC Homemaker-Home Health Aide Demonstrations. These demonstrations, conducted as random assignment experiments during the 1980s, included seven state-run programs, each providing selected recipients of Aid to Families with Dependent Children (AFDC) with four to six weeks of training and up to a year of subsidized employment as homemakers and home health aides.[2] Over a thirty-month follow-up period, this intervention produced substantial, sustained earnings gains and important welfare reductions in most states.[3]

Like many government-sponsored employment and training programs, the demonstrations served volunteer applicants, whose applications were screened on a selective basis by program intake staff prior to admittance. Screening was based on formal eligibility criteria and subjective assessments of the applicant's potential to benefit from or succeed in the program. Formal eligibility criteria restricted the demonstrations to individuals who had received AFDC for at least three months and who had not worked as homemakers or home health aides in the previous six months. Some states imposed further requirements, such as the availability of private transportation or a minimum educational or literacy level. All seven demonstrations tended to "cream" applicants, admitting a larger share of those with better education and/or more employment experience. The same factors have been found to correlate with higher subsequent earnings among control group members, though not with larger program effects.[4]

The data from the evaluation are particularly well-suited for a critical examination of alternative applicant-based comparison groups for several reasons. First, they contain information on randomly assigned treatment and control group members, which is essential for the creation of an experimental benchmark estimate of program impacts.[5] Second, they include a measure of program selection—ratings by intake workers of the applicants' potential as homemakers and home health aides on a scale of 1 (excellent) to 4 (poor). Also, unlike many other evaluation data sets, long-term follow-up measures of both participant and nonparticipating applicant earnings are available from a single, consistent source (see below).

The analysis sample breaks down into five subgroups defined by the demonstration intake process:

- 909 **withdrawals**, who applied to the program but chose to leave the intake process before being screened by intake staff;[6]
- 931 **screen-outs**, who applied but were rejected by intake staff
- 282 **no-shows**, who applied to and were accepted into the program, but decided not to participate;
- 1,573 **participants**;[7]
- 1,826 **controls**, who were accepted by program staff but were then randomly selected for exclusion from the program.[8]

All applicants filled out Background Information Sheets, which collected detailed information on a wide range of baseline characteristics, including demographics, employment experience, caregiving experience, and public program participation. In addition, staff rated each applicant on a four-point scale to indicate his or her perceived potential as a homemaker or home health aide. Reasons for nonparticipation were also recorded to distinguish screen-outs from withdrawals.

Demonstration intake ran from early 1982 through May 1985. The current analysis focuses on the cohort of individuals who went through intake between July 1984 and May 1985, the only cohort for which complete data are available on all demonstration applicants.[9]

The demonstration data have been supplemented with long-term follow-up data on annual earnings for all five subsamples. These data come from the U.S Internal Revenue Service (IRS), which collects employers' reports of gross wages and salaries paid to individual workers throughout the United States. The earnings reported correspond reasonably well to earnings measures based on survey self-reports from the original evaluation.[10] To preserve anonymity, the IRS provided these data in aggregate form (as group means and standard deviations) for groups of 10 or more individuals. Earnings data are available on all applicants in the July 1984-May 1985 cohort for the years 1984 through 1990. This provides four years of postprogram data (1987-1990) for participants, allowing us to consider the performance of alternative nonexperimental estimators over a longer period of time after the end of the program than in most evaluations of training programs.[11]

Exhibit 3.1 presents the sample sizes and descriptive characteristics for the five analysis samples. As shown in the exhibit, there are 82 group observations for withdrawals, 83 screen-out groups, 24 no-show groups, 144 participant groups, and 166 control groups. There are never more than 19 nor less than 10 individuals in any single group. Group sizes were sometimes set above 10 to increase within-group homogeneity on baseline variables, important in assuring maximum explanatory power from key group-level baseline variables, such as the intake rating.[12] The average number of individuals in each group is slightly above 10, ranging from 10.9 for participants to 11.8 for no-shows. Thus, additional homogeneity is achieved at only a small cost in terms of the number of overall group observations. Details of the grouping procedure and the resulting degree of homogeneity are presented in appendix A. As explained there, grouping does not affect the

Exhibit 3.1 Sample Sizes and Selected Characteristics of Applicant Groups

	Withdrawals[a]	Screen-outs	No-shows	Participants	Controls
Sample size					
Number of persons	909	931	282	1,573	1,826
Number of groups	82	83	24	144	166
Average group size	11.1	11.2	11.8	10.9	11.0
Group characteristics[b]					
Average age	30	31	29	30	29
% White	31	39	18	29	29
% Black	55	53	63	59	57
% Education < 12 years	53	50	53	46	49
% Married, spouse present	13	12	6	9	9
Average number of dependent children	2.0	2.0	1.9	1.9	1.9
% Ever worked for pay	82	85	86	87	86
Average previous hourly wage[c]	$3.00	$3.17	$3.37	$3.33	$3.33
Average 1984 annual earnings[d]	$823	$592	$710	$840	$749
% Arkansas	4	11	16	16	16
% Kentucky	17	31	3	14	11
% New Jersey	31	15	15	18	18
% New York	15	14	28	13	14
% Ohio	26	30	4	15	14

% South Carolina	4	4	10	8	9
% Texas	3	4	24	16	18
Average intake rating[e]	3.0	3.1	2.4	2.3	2.3
% Intake rating = excellent or good	53	44	84	86	86
% Intake rating = fair or poor	47	56	16	14	14
% Intake rating missing	65	36	18	12	12

a. Of the 82 withdrawal groups, 19 consist of individuals for whom reason for nonparticipation is missing. These individuals are assumed to have withdrawn from the intake process voluntarily without providing a reason to the demonstration staff.

b. Observations with missing or invalid values are excluded from the calculations except where otherwise noted. See appendix A for missing data rates and additional details on the baseline characteristics of the sample.

c. Includes only individuals whose highest wage on previous job was less than or equal to $25 per hour. Those who had never worked for pay are coded zero and included in the average. Those who worked for an unknown wage rate are assumed to have earned the average reported wage rate for their subpopulation (withdrawals, etc.), intake rating, and race.

d. Includes $0 values for those with no earnings reported in the IRS data.

e. Intake rating is scaled as: 1=excellent; 2=good; 3=fair; 4=poor.

precision of mean earnings estimates for subpopulations allocated to mutually exclusive groups. Grouping does, however, affect the precision of mean earnings estimates for subpopulations mixed together in the same groups, as well as the precision of multivariate relationships involving one or more such "mixed" subpopulation. (See note 13.)

As can be seen from the exhibit, the typical applicant was a thirty-year-old black single parent with two children and less than twelve years of education. Though not shown, virtually all applicants were women, as was true of almost all adult recipients of AFDC in the mid-1980s. Most applicants had previously worked for pay at or just below the (then) minimum wage of $3.35 per hour. Annual earnings prior to application were very low, around $700 in 1984 dollars. Nonparticipating applicants and participants differed little on these factors, except that participants tended to have somewhat more education and—except when compared to withdrawals—somewhat higher annual earnings. The distribution of the four populations across states differed considerably, however, with withdrawals disproportionately concentrated in New Jersey and Ohio, screen-outs in Kentucky and Ohio, and no-shows in New York and Texas, compared with the participant group. A final column of the exhibit shows the control group to be well matched to the combination of (a few) no-shows and (many more) participants that form the experimental treatment group.

Not surprisingly, given prior evidence of "creaming," almost all of those admitted to the demonstrations—no-shows, participants, and controls—were rated as having excellent or good potential as homemakers or home health aides, while less than half of screen-outs were so rated. As a result, average rankings varied considerably between the two groups: 2.3 for those admitted versus 3.1 for those screened out on a scale of 1 (excellent) to 4 (poor). The overlap in ratings between screened in and screened out applicants is particularly striking: 44 percent of screen-outs were rated good or better, while around 14 percent of "screen-ins" were rated fair or worse. Clearly, intake decisions did not move in lock-step with the reported ratings, a pattern with many possible interpretations that we will explore later when interpreting our findings.

The exhibit also shows one drawback of the data: a rather high proportion of withdrawals and screen-outs (65 and 36 percent, respectively) have no potential ratings recorded. Since the original design of the evaluation was experimental, obtaining subjective ratings for screen-outs was not emphasized. The dearth of ratings for withdrawals reflects the fact that these individuals tended to leave the intake process prior to a full assessment by demonstration staff.

Impact Estimation Methodology

For each of the three nonexperimental comparison groups discussed in chapter 2—withdrawals, screen-outs, and no-shows—we calculate three separate impact estimates. For a given comparison group, each successive estimate accounts for progressively more information about the selection process, as explained below. Once obtained, we compare the range of estimates produced to see how much of the difference in estimated impacts is attributable to the choice of the comparison group and to the addition of information on the intake workers' ratings of potential. Experimental estimates are provided as reference points for this purpose, and each follow-up year is considered separately. However, the formal comparisons of nonexperimental and experimental estimates needed to gauge the overall success of these nonexperimental approaches in eliminating selection bias are deferred until chapter 4.

Differences in Means

The simplest measure of program impact is the unadjusted difference in mean annual earnings between participants and a particular comparison group. This is calculated as:

$$\delta_1 = \bar{y}_p - \bar{y}_c$$

where:

$\bar{y}_p$ = *average annual earnings outcome for participants,*

$\bar{y}_c$ = *average annual earnings outcome for comparison group.*

No other information is incorporated, so that if participants differ from comparison group members on any selection factors, the estimates will be biased.

Controlling for Standard Selection Factors

We use two techniques to control for readily measured differences in the participant and comparison samples due to self-selection factors. First, we expand the comparison of means into a multivariate regression analysis that not only distinguishes participants and comparison group members using a dummy variable, *p*, for participation, but also incorporates information on the baseline characteristics of both groups using variables of the sort shown in exhibit 3.1:

$$y = \underline{X}\underline{B} + \delta_2 p + \varepsilon$$

where:

y = *annual earnings,*
$\underline{X}$ = *a vector of baseline characteristics,*
p = 1 *for participants,* 0 *for comparison group members.*

Baseline variables include age, race, education, marital status, number of dependent children, whether the individual has ever worked for pay, maximum wage previously earned, earnings in 1984 (the baseline year), and state of residence.[13]

Since the analysis must be conducted at the group level, background variables are expressed either as mean values for the group or as percentages of the group possessing a given characteristic.[14] Weighted least-squares regression is used to adjust for variations in group size. (See appendix B for details.)

Regression-adjusted estimates obtained in this way are comparable to the regression-adjusted estimates used in studies involving external comparison groups. The estimates produced by this model differ from previous commonly used estimates solely in their reliance on applicant-based comparison groups. If applicants select or are selected into the program on the basis of observed characteristics, δ_2 should approximate the actual program impact.[15] It will differ from the true impact only by chance or because of the existence of additional selection factors not included in the model.

Additional Controls for Program Selection

To adjust for variables omitted from the model that were known to the intake workers but are not among the variables that were explicitly measured, we re-estimate the equation adding intake workers' ratings of applicants as potential homemakers and home health aides. This provides the second focus of our analysis, again within the multiple regression model. The following equation is estimated for each comparison group:

$$y = \underline{X}\underline{B} + \underline{S}\underline{\theta} + \delta_3 p + \varepsilon$$

where:

p = 1 *for participants,* 0 *for comparison group members,*
$\underline{S}$ = *a vector of subjective rating variables.*

This model controls for all measured selection factors, including demographic characteristics and intake workers' ratings of applicant potential. The intake rating variables indicate the percentage of each group rated at each level on a scale of 1 (excellent) to 4 (poor), plus the percentage with missing ratings data.[16] This represents the grouped-data version of the "regression discontinuity" model discussed in chapter 2. The particular parameterization chosen—essentially dummy variables at the individual level for each value on the rating scale—avoids the problem of having to assume a functional form for how ratings affect follow-up earnings.[17] Inclusion of a separate missing data category assures that data on those without ratings will not confound estimates for those with ratings. The program impact estimate, δ_3, that emerges from this model attempts to isolate the effect of the program from the effects of program selection on both observable and subjective factors, and from the effects of self-selection on observable factors.

Adjusting for Self-Selection in the Experimental Sample

In order to interpret the impact estimates obtained from the above procedures, we need a benchmark against which the various nonexperimental estimates may be compared. To this end, we exploit the original experimental design of the demonstrations wherein participants were randomly assigned to treatment status or control status following acceptance into the program. To obtain an experimental estimate in each year, we reestimate our final equation using data on experimental treatment group members (no-shows plus participants) and control group members, where p now represents assignment to the treatment group. This regression gives us an unbiased estimate of the average program effect on the full treatment group.

We then convert that estimate into a measure of the average effect on *participants* by dividing it (and its standard error) by the share of the treatment group who actually participated. Algebraically, the effect on participants, P, is obtained from the equation

$$T = r \quad P + (1 - r) \quad N,$$

where T is the treatment effect for those assigned to the treatment group, N the treatment effect for no-shows, and r the proportion of the treatment group that participated in the program. Assuming $N = 0$ (no effect on no-shows) and solving for P, we get a formula for the program's effect on the average participant:

$$P = T/r$$

This procedure—common in the evaluation literature—is based on two assumptions: the demonstration program had no effect on no-shows; and the distribution of no-show "types" in the experimental sample matches the distribution of nonparticipants among those admitted to an actual program of the same design.[18]

The first of these two assumptions is well known, discussed by Bloom (1984b) and others and examined empirically by Heckman, Smith, and Taber (1994). The second assumption is less well recognized; we are indebted to Arthur Goldberger for bringing it to our attention.[19]

The assumption of equal patterns of nonparticipation in the "sample" (the demonstration or experiment) and the "population" (the real program) allows us to ignore potential effects on those who participate in one instance and not in the other. It does not make any difference whether demonstration no-shows would have participated or benefitted from a real program, as long as the selection process that produces them is the same in both instances. If the selection process differs between the two settings, the "sample-based" demonstration analysis will yield a biased estimate of the "population" parameter of interest, *P*, unless the effect of the program is the same for individuals whose participation differs between the two settings and those whose participation is constant.[20]

Details of the Regressions

All regression equations are estimated as a cross section, separately for each of six years, 1984 through 1990. The 1984 regressions omit the baseline (1984) earnings measure from the specification and are presented for comparison purposes only, since they do not represent measures of demonstration impact. Appendix B contains the detailed regression results for the fully specified model (including the ratings variables).

Estimation Results

The experimental impact estimates for each year are given in exhibit 3.2. Impacts per treatment group member in the first column are taken directly from the coefficient on treatment status in our final equation ($_3$). As can be seen, the largest estimated effects are evident during the in-program years 1985 and 1986. These periods included up to a year of subsidized employment for each participant; thus large earnings effects are to be expected. Impacts decline steadily beginning with the first postprogram year (1987). By 1990, four years after the program, treatment group members earned an average of $402 more than the average control member. Most of the impact estimates are statistically significant at or near the .10 level, and most at the .01 level. They fall in

Exhibit 3.2 Impact Estimates from the Experiment, with Participant Mean Earnings (standard errors in parentheses)

Year	Average effect on: Treatment group members		Average effect on: Participants		Participant mean earnings	
1984: Preprogram year	$ 99	(74)	$ 116	(86)	$ 840	(706)
1985: In-program year	2,237	(119)	2,610	(139)	4,206	(1,288)
1986: Mostly postprogram year	1,023	(135)	1,194	(158)	3,776	(1,526)
1987: Postprogram year	551	(157)	643	(183)	3,936	(1,824)
1988: Postprogram year	500	(191)	583	(223)	4,626	(2,056)
1989: Postprogram year	338	(209)	394	(244)	5,174	(2,172)
1990: Postprogram year	402	(226)	469	(264)	5,623	(2,363)

the middle of previously available long-term impact estimates for experimentally evaluated AFDC employment programs.[21]

The third column of exhibit 3.2 applies a no-show adjustment to the results in column 1 to arrive at estimates of the impacts per participant. Adjusting for no-shows results in larger impact estimates, since the no-shows did not receive training and therefore contributed nothing to the average impact per treatment group member. Column 3 is the relevant experimental benchmark for comparison with nonexperimental estimators, which universally (in this study and others) measure impacts on program participants only, not participants plus no-shows. The estimated impact of the program on the earnings of the participants during the four purely postprogram years is a modest increase (8 to 16 percent), as shown in the last two columns of exhibit 3.2.

Withdrawal-Based Estimates

The estimates obtained using withdrawals as a comparison group are presented in exhibit 3.3.[22] The first column shows the unadjusted difference in mean earnings between participants and withdrawals. As expected, the difference is largest during the period of training and subsidized employment (1985), with the estimated impact falling off sharply in the first postprogram year and declining slowly thereafter. This mirrors the pattern seen earlier in the experimental estimates (repeated in the last column of the exhibit for reference). The nonexperimental impact estimates are, however, uniformly larger than the experimental estimates for all years after 1984. This implies that those who withdrew early in the intake process experienced lower earnings than the participants would have in the absence of the program.

Exhibit 3.3 Impact Estimates Based on Withdrawal Comparison Group

Year	Unadjusted difference in means	Based on regression with conventional independent variables[a]	Based on regression with subjective rating[b]	Experimental estimate of impact on participants[c]
1984: Mostly preprogram year	$ 41	$ 125	$ 121	$ 116
1985: In-program year	2,774	3,413	3,422	2,610
1986: Mostly postprogram year	1,687	2,076	2,003	1,194
1987: Post-program year	949	824	635	643
1988: Post-program year	907	951	800	583
1989: Post-program year	887	1,564	1,316	394
1990: Post-program year	852	2,337	2,074	469

a. Independent variables: age, race, education, marital status, number of children, ever worked for pay, highest wage attained, 1984 total earnings (except in 1984 regressions), and state of residence.

b. Includes all independent variables in previous note plus intake workers' rating of potential.

c. Experimental treatment-control difference, controlling for independent variables in note a, and subjective rating of potential and adjusted to apply to participants only.

We come to this conclusion based on the following reasoning. The control group's earnings, adjusted for no-shows, provide a selection-bias-free estimate of what participants would have earned in the absence of the program. As a result, the entire difference between the (no-show-adjusted) experimental impact estimate in column 4 and the participant-withdrawal difference of means in column 1 results from the difference between control group earnings and withdrawal earnings. A similar interpretation applies to the comparison of column 4 to columns 2 and 3, where additional demographic and selection factors have been used to adjust the earnings of withdrawals to more closely resemble those of controls.

The second column of exhibit 3.3 gives withdrawal-based impact estimates after controlling for conventional demographic and selection factors, including age, race, education, marital status, number of children, work experience, state of residence, and (except in 1984) previous earnings. This column displays the curious result that accounting for observable characteristics produces generally higher impact estimates, rather than lower estimates as one might have expected given that the original estimates were too high in most years. This suggests that the measured variables inflicted a downward bias on the impact estimates before they were neutralized, but were overwhelmed by an even stronger upward bias attributable to unmeasured factors. Removing only the first set of influences leads to an even larger upward bias in the column 2 estimates.

The third column of exhibit 3.3 shows the results of estimating the full equation described above; adding the subjective rating dummy variables. We do not expect intake workers' ratings to make large differences in this instance, since they played no role in nonparticipation decisions and, in any case, are missing for two-thirds of all withdrawals. Even so, including the subjective ratings does move the impact estimates toward the experimental estimates in all years except 1985, though not by much. Estimates with the rating variables are $150 to $260 closer to the experimental benchmark estimates in each of the postprogram years beginning in 1987. The small extent of these improvements may be due to the fact that potential ratings are available for only about a third of the withdrawal sample.[23]

For both the regression models, the deviation of the nonexperimental estimates from the experimental benchmarks widens late in the follow-up period. This deviation begins fairly large, narrows considerably in the early post-program years, and then widens again. By 1990, the gap is over $1,600 for the fully-adjusted estimator. Hence, in most years, the withdrawals in this sample do not appear to provide an adequate comparison group even after adjustments for baseline differences. (Chapter 4 considers this question more formally.) This finding is not unexpected, since withdrawals exit the employment and training program intake process for a wide variety of reasons that we could not expect to fully measure. Hence, systematic differences between withdrawals and participants may remain even after controlling for available selection factors.

Screen-Out-Based Estimates

Exhibit 3.4 presents the estimates obtained using screen-outs as the comparison group. These estimates are not obtained by directly comparing participants and screen-outs in the standard fashion, however, as was previously done with participants and withdrawals. Here, we compare participants plus no-shows to screen-outs in order to obtain a measure of impact on the average "screen-in" (participants plus no-shows) and then apply the no-show correction to that estimate (and its standard error). As with the experimental estimates, this latter step converts impacts on the average "screen-in" into impacts on the average *participant* on the assumptions that the demonstrations had no effect on no-shows and that self-selection of no-shows in the demonstrations would not change in a real program of the same design.

The reason for this departure from the standard methodology was noted in chapter 2. Unlike the withdrawal comparison group just analyzed, we expect to achieve the best match of screen-outs to a with-program population by comparing them to participants and no-shows combined. As explained in chapter 2, the objective of the screen-out-based approach is to control for the program selection that separates screen-outs from "screen-ins," using intake workers' ratings at the point of selection. After that point, adjustments for self-selection into the participant group (versus the no-show group) can be handled—just as in an experiment—through the no-show adjustment proposed by Bloom (1984b).[24] The same framework does not apply to the other applicant-based comparison groups considered here, however. Since there is no presumption that withdrawals and/or no-shows are comparable with participants *plus no-shows*, we stay with the simpler comparison of withdrawals or no-shows to participants.

Exhibit 3.4 Impact Estimates Based on Screen-Out Comparison Group[a]

Year	Unadjusted difference in means	Based on regression with conventional independent variables[b]	Based on regression with subjective rating[c]	Experimental estimate of impact on participants[d]
1984: Mostly preprogram year	$ 274	$ 192	$ 183	$ 116
1985: In-program year	3,107	3,448	3,401	2,610
1986: Mostly postprogram year	1,927	1,862	1,802	1,194
1987: Post-program year	1,388	1,111	1,030	643
1988: Post-program year	1,326	1,036	1,014	583
1989: Post-program year	1,138	850	696	394
1990: Post-program year	1,026	931	768	469

a. All estimates are based on a comparison of participants plus no-shows to screen-outs and then adjusted to apply to participants only.

b. Independent variables: age, race, education, marital status, number of children, ever worked for pay, highest wage attained, 1984 total earnings (except in 1984 regressions), and state of residence.

c. Includes all independent variables in previous note plus intake workers' rating of potential.

d. Experimental treatment-control difference, controlling for independent variables in note a, and subjective rating of potential and adjusted to apply to participants only.

With this refinement to the matching process, we expect the screen-out-based approach to produce better results than our earlier analysis of withdrawals. As can be seen in the first column of the exhibit, however, this is not the case for the basic estimates prior to controlling for baseline characteristics and intake ratings. The unadjusted difference-in-means estimates (which include the no-show correction) exceed the corresponding estimates in exhibit 3.3 and, therefore, are further from the experimental benchmarks than the withdrawal-based estimates. That they substantially surpass the experimental benchmarks confirms inferences made by Bell and Orr (1988) that demonstration intake staff "creamed" applicants with the best earnings prospects, producing screen-outs whose subsequent earnings were well below what participants would have earned absent demonstration services.

Except near the end of the follow-up period, discrepancies between unadjusted screen-out estimates of program impacts and experimental estimates are the largest of any comparison group considered and generally exceed $500 per year. This of itself is not a problem, since the real strength of screen-outs as a comparison group is not their initial similarity to participants (which we do not expect) but their likeness on self-selection factors once the determinants of program selection have been taken into account. Consistent with this theory, column 2 shows that—unlike the experience with withdrawals—including conventional control variables moves the estimates toward the lower experimental benchmarks in most years. It appears, then, that program selection on the measured variables produces a screen-out population with lower earnings than participants, all other things equal. Hence, controlling for those variables through the regression reduces the earnings difference between screen-outs and participants, resulting in impact estimates closer to the experimental estimates in most years. Nonetheless, the screen-out-based estimates in column 2 rarely come within $400 dollars of the experimental estimates in any postprogram year.

Theory further suggests that the impact estimates generated from the screen-out comparison group should show the largest improvement with the addition of the intake rating variables. Indeed, the addition of these variables in column 3 of the exhibit closes the gap with the experimental estimates by nearly $100 in each postprogram year except 1988. In all years beginning with 1985, the rating variables improve the estimates to some degree, with later years showing particularly strong improvements. In percentage terms, the gap is reduced by 17 percent in 1987, 5 percent in 1988, 34 percent in 1989, and 35 percent in 1990. Hence, despite the facts that the subjective rating variables in our data are limited to fairly coarse four-way distinctions, that these distinctions do not track entirely with admission decisions, and that somewhat over a third of the screen-outs have no rating recorded, the ratings have some explanatory power in the predicted direction.

That they could not do more to remove the initial differences between the screen-out-based and experimental estimates may be due in part to the inexact relationship between the ratings variable and actual selection decisions. As noted in discussing exhibit 3.1, a substantial number of highly rated applicants were excluded from the demonstrations and a number of poorly rated applicants admitted. This necessarily reflects inconsistent admission standards (across intake workers and/or over time), measurement error in the ratings variable, or the use of factors beyond the ratings in making admission decisions. As noted in chapter 2, the first of these factors increases the ability of the regression discontinuity approach to successfully adjust for incoming differences between the participant and screen-out samples (by providing more overlap between the two groups). In contrast, the other two factors reduce the usefulness of the rating variable as a means of refining screen-out-based estimates. Unfortunately, we do not know what accounts for the observed variation in admissions by rating level in these data.[25]

Another noteworthy aspect of the findings is the marked improvement in the screen-out estimates relative to the experimental benchmarks over time. In 1986, the first largely postprogram year, the difference between the fully adjusted nonexperimental and experimental estimates is nearly $500, while by 1990 the impact differential is less than $300. We consider the possible reasons for this trend in chapter 4.

No-Show-Based Estimates

Nonexperimental impacts based on the no-show comparison group are displayed in exhibit 3.5. Unadjusted difference-in-mean impacts (estimated using the original methodology applied to withdrawals) hover around $1,000 for each of the postprogram years, well above the experimental estimates shown in the final column. Thus, like the other two nonparticipating applicant groups, no-shows earned less than participants would have absent the demonstration throughout the follow-up period. These results are similar to those found earlier for withdrawals, the other self-selected comparison group considered here.

Exhibit 3.5 Impact Estimates Based on No-Show Comparison Group

Year	Unadjusted difference in means	Based on regression with conventional independent variables[a]	Based on regression with subjective rating[b]	Experimental estimate of impact on participants[c]
1984: Mostly preprogram year	$ 143	$ 61	$ 26	$ 116
1985: In-program year	2,751	2,077	2,100	2,610
1986: Mostly postprogram year	1,489	1,088	1,145	1,194
1987: Post-program year	1,051	898	916	643
1988: Post-program year	1,039	669	706	583
1989: Post-program year	1,192	836	814	394
1990: Post-program year	1,154	882	867	469

a. Independent variables: age, race, education, marital status, number of children, ever worked for pay, highest wage attained, 1984 total earnings (except in 1984 regressions), and state of residence.

b. Includes all independent variables in previous note plus intake workers' rating of potential.

c. Experimental treatment-control difference, controlling for independent variables in note a, and subjective rating of potential and adjusted to apply to participants only.

Unlike the withdrawals, however, adjusting the no-show-based impact estimates for objective baseline characteristics produces distinct improvements in their performance. These adjustments actually overcompensate for preexisting differences in 1984 through 1986, leading to conventional regression-adjusted estimates in column 2 that fall below the experimental benchmarks. In later years, the nonexperimental estimates again exceed the experimental benchmarks. The explanations offered earlier in connection with similar but less extreme changes in the withdrawal-based estimates may again account for these patterns.

Adding the intake rating variables to the analysis has very little effect on the no-show-based estimates. Apparently, intake workers' ratings do not discriminate well among those who make it through the demonstrations' intake screens. This should perhaps not be surprising, since most of those admitted to the program (86 percent) were rated either "excellent" or "good" on the crude four-point scale used here. The estimates derived from the full regression model are quite close to the experimental estimates in 1986, the first substantially postprogram year, but lie $100 to $400 above the experimental benchmarks in the remaining four follow-up years.

Summary

Exhibits 3.3 through 3.5 illustrate the application of applicant-based nonexperimental impact estimation techniques to data on a voluntary employment and training program for AFDC recipients. They show that, in this particular application, controlling for standard demographic selection variables and subjective intake ratings generally moves applicant-based impact estimates in the direction of the experimental estimates for both the screen-out and no-show comparison groups. The withdrawal-based estimates are not consistently improved by these variables, however, although of the three comparison groups considered withdrawals do produce simple difference-in-means impact estimates that are closest to the experimental benchmark.

In the next chapter, we consider more formally how well each of these estimators has performed relative to the selection-bias-free experimental estimates.

NOTES

1. Further discussion of the empirical work is presented in Bell et al. (1993) and Cain et al. (1993). The results presented in this chapter include certain corrections and refinements to the earlier results and should be considered final.

2. The seven states that ran demonstrations under the sponsorship of the U.S. Department of Health and Human Services, Health Care Financing Administration, were Arkansas, Kentucky, New Jersey, New York, Ohio, South Carolina, and Texas. For a full description of the demonstrations, see Bell et al. (1987).

3. Bell and Orr (1994) summarize the results of the original experimental evaluation.

4. See Bell and Orr (1988).

5. Random assignment was centrally controlled by Abt and rarely violated in the field. See Bell, Enns, and Flanagan (1987) for details.

6. For purposes of analysis, this group also contains a small number of applicants who left the intake process prior to admission with no reason reported. We assume that had intake staff known the reasons these individuals did not enter the demonstrations, they would have reported them. Hence, we take the absence of reported reasons to imply that the staff did not exclude the individuals involved and, therefore, that the applicants withdrew voluntarily (i.e., were withdrawals).

7. No-shows plus participants comprise the experimental treatment group of 1,855 individuals, a number roughly equivalent to the size of the control group as a result of the 50-50 random assignment ratio employed.

8. Controls remained eligible for other employment and training services in the community.

9. The decision to collect data on nonparticipants in order to study the demonstration intake process was made after the demonstrations were well under way. As a result, complete data on nonparticipants are available only for the third annual cohort of applicants.

10. Average earnings of the experimental sample (no-shows, participants, and controls) are almost identical between the two sources in 1985, the year of greatest overlap. In the sample of 3,432 individuals for whom both types of data are available, earnings averaged $3,434 from the survey and $3,361 from the IRS data. The correspondence is not so tight at the level of individual observations, however, as is often the case for earnings measures from different sources. (See, for example, Appendix E of Bloom et al. 1993.) As noted in the text, IRS data were provided as means and standard deviations for groups of 10 or more individuals, to protect confidentiality. The correlation coefficient between the IRS and survey measures at the group level is .63, with 54 percent of the groups differing by less than $1,000 between the two sources. Forty-six percent of the groups differed by more than $1,000, and 17 percent by more than $2,000. Partly as a result of these deviations, the experimental impact estimate for 1985 for the overlapping sample is $1,527 from the survey and $1,209 from the IRS data. Thus, while not as closely aligned as one might hope, the two sources do give basically the same picture as regards demonstration impact. (This modest divergence in impact estimates is also due partly to the extrapolation technique used to project survey-based earnings to the end of 1985 for individuals interviewed before the end of the year, a technique that biases downward control group—but not treatment group— mean earnings.)

11. Participants spent an average of twelve months in the program. With an average point of entry of December 1984, 1984 can be considered the preprogram year and 1985 the in-program year. Most, but not all, of 1986 represents postprogram experience.

12. This point may be explained with an example. If a baseline characteristic, such as an "excellent" intake rating, was evenly distributed across groups so that all groups had the same percentage, the variable would have no explanatory power in a regression of group-level observations. Conversely, maximum variation and explanatory power at the group level is achieved by making groups totally homogeneous with respect to a particular variable.

13. The specific variables used to represent each of these factors appear in the tables of regression results in appendix B.

14. Additional details on the homogeneity of the groups appear in appendix A. Since all groups are homogeneous with respect to intake status, the participation dummy in the equation is a standard 0/1 variable.

15. Bell and Reesman (1987, table B.2) find that a collection of baseline characteristics, including those considered here plus prior caregiving experience, non-AFDC welfare status, and availability of private transportation, are highly significant as a group in predicting participation in each of the seven state demonstrations. So, too, are many individual characteristics. Unfortu-

nately, the proportion of variation in participation accounted for by these factors cannot be computed because the 0/1 participation equation is estimated using maximum likelihood methods rather than ordinary least squares (which would have provided an R-squared coefficient measuring "explained" variation).

16. One of these five variables must be excluded from the model (which also includes an intercept) since they sum to 1 for each grouped observation. Values of 2 and 3 on the ratings scale correspond to "good" and "fair," respectively.

17. Note that this specification is possible only when there are both participants and comparison group members at each rating level, a feature that differs from the original regression discontinuity approach in ways that become important as we interpret our findings below.

18. See Bloom (1984b) for an early discussion of the technique. Angrist and Imbens (1991) and Imbens and Angrist (1992) reintroduce the same method as an instrumental variable approach, using assignment to the treatment group as an instrument for participation. The procedure has been applied to several large-scale evaluations of voluntary employment and training programs; see, for example, Bell and Orr (1994) and Bloom et al. (1993).

19. Unpublished comments presented at a session on "Does Job Training for the Disadvantaged Work? Findings from the National JTPA Study," Annual Meetings of the American Economic Association, Boston, Massachusetts, January 3, 1994.

20. Goldberger further suggested that the selection process be examined directly, noting that participation patterns may depend, among other things, on the pool of persons eligible to apply for the program, the sites included in the evaluation, and the incentives to participate presented by each local program office.

21. Couch (1992) estimates that the National Supported Work Demonstration increased the earnings of AFDC recipients a statistically significant $490-655 per year (in 1978 dollars) in the fourth to eighth years following demonstration exit. Friedlander and Burtless (1992) report much smaller long-term earnings gains from a Virginia work-welfare initiative—$200-300 per year in the late 1980s (in unadjusted dollars) for the fourth and fifth year of follow-up. Gueron and Pauly (1991, table A.1) report that Supported Work cost substantially more per participant to implement than did the AFDC Homemaker-Home Health Aide Demonstrations, and the Virginia demonstration substantially less.

22. For simplicity, we omit standard errors and significance tests from this and the following two exhibits. We discuss approximation errors and how they should affect the interpretation of the numbers in chapter 4. For now, our objective is simply to illustrate the behavior of the nonexperimental impact "point estimates" shown in the exhibits as we vary the comparison group and regression control variables used.

23. Many withdrawals left the intake process prior to the point where intake workers evaluated potential.

24. In practice, the use of the alternate methodology makes little difference to the size of the estimates, except in 1985 where estimates from the standard approach (not shown) are $300-400 smaller than those shown here.

25. Nor can we examine that variation by intake worker because the intake worker who made the selection decision was not identified on the forms.

4
Testing Alternative Estimates for Selection Bias

Our purpose in reintroducing applicant-based impact measures to the evaluation of employment and training programs is to find a nonexperimental measure that provides adequate protection from the selection bias problems that have plagued other nonexperimental techniques. The crucial question for the empirical portion of our analysis is therefore: Which, if any, of the applicant-based approaches are sufficiently protected from selection bias to provide adequate substitutes for a randomized experiment?

This chapter presents our attempt to answer that question for the various applicant-based impact estimates presented in chapter 3, using evidence from the associated experiment. If we can find a technique that eliminates selection bias, or reduces it to a tolerable level, we can reasonably propose that approach as a possible alternative to experiments in future evaluations carried out in similar contexts. If we cannot, the applicant-based strategy will have failed its first test. Either way, further tests of the methodology using other experimental data sets should be undertaken; given the limitations of available test procedures and the fact that any result we obtain is conditioned on a specific demonstration project and data set, a single test should not be viewed as conclusive in either direction.

We begin by reviewing how others have attempted to validate nonexperimental impact estimates by comparing them to experimental results. We then propose a more comprehensive and rigorous approach for gauging the risk of selection bias in any nonexperimental measure, an approach that allows careful comparisons among measures and provides a formal basis for judging each measure against the unbiased experimental norm. The concluding sections of the chapter apply this approach to our applicant-based estimates to determine which might be adequate substitutes for an experiment.

Previous Validation Methods

Just as nonexperimental estimation techniques have evolved over the years (see chapter 1), so too have methods for assessing the reliability of those techniques using experimental data. Two sets of previous studies have attempted to validate nonexperimental estimators against experimental estimates:

- Analyses of data from the National Supported Work Demonstration, which employed relatively informal validation methods; and
- Analyses of AFDC work incentive (WIN) employment program demonstrations, which went a good deal further in formalizing and systematizing validation methods.

Supported Work Analyses.

All of the Supported Work studies follow one of two approaches to assessing alternative nonexperimental impact measures against the experimental norm: (1) they compare the magnitude of the experimental and nonexperimental estimates (LaLonde, 1986; Fraker and Maynard 1987), applying judgmental standards to decide which nonexperimental estimates came "close enough" to the experimental benchmark and which did not; or (2) they compare the statistical inferences produced by the two types of estimates (Heckman and Hotz 1989; Couch 1992), contrasting statistically significant estimates with insignificant estimates (or with significant estimates of the opposite sign).

In the first approach, nonexperimental impact estimates that differ substantially from one another (e.g., by more than several hundred dollars of earnings per year) are typically considered to be inconsistent, and those that differ from the experimental benchmark by similar magnitudes are considered to be biased. Implicitly, estimates that fall within these bounds are judged consistent or unbiased. The magnitudes used for this purpose tend to be phrased in terms of "large relative to the experimental estimate" or simply "large," although the derivation—or even the explicit statement—of a cutoff between large and small magnitudes is not given. Thus, while clearly focused on a criterion one might apply to determine the acceptable degree of bias—the relative magnitudes of nonexperimental and experimental estimates—the standards used by LaLonde and by Fraker and Maynard are incomplete and subject to dispute.

Comparisons based on statistical inferences are less subjective but also incomplete and, as we argue below, even less discriminating than the comparison of magnitudes. Under this approach, a nonexperimental estimate is considered to be an acceptable substitute for the experimental result if it leads to the same policy conclusion among three possibilities:

- The program increased participant earnings;
- The program decreased participant earnings; or
- The evidence is inconclusive as to whether the program increased, decreased, or failed to change participant earnings.

Distinctions among these three situations are based on the results of tests of the statistical significance of the impact estimates. If a test rejects the null hypothesis of no impact because the estimate is positive and too large to reflect chance variations in the data (random sampling and measurement errors) for a program that truly had no impact, one concludes that the impact was positive. The corresponding situation for a negative impact estimate implies that the impact was negative. When the null hypothesis cannot be rejected by either a positive or negative estimate because the estimate lies within the range of chance variation, the evidence is judged inconclusive.

Ultimately, the success of an employment and training program hinges not on whether it produces a significant earnings gain, but on whether the earnings gain is large and sustained enough to offset the program's cost. For practical reasons, this question has not been addressed comprehensively by formal hypothesis tests.[1] Yet it remains the "bottom line" consideration for employment and training evaluations.[2]

If one cares most about the net of program benefits over program costs (a measure called "net social benefits" in the literature), it is helpful but not sufficient to correctly categorize the main benefit measure—impact on participants' earnings—in relation to zero. This is particularly true when the usual hypothesis test shows a program to have a significantly positive effect. A numerical example illustrates why. Suppose that for some employment and training program, the experimental estimate and the nonexperimental alternative both show statistically significant earnings gains—say, $500 per participant for the experiment and, because of upward selection bias, $700 per participant for the nonexperimental alternative. In this example, the nonexperimental estimate passes the test for "same statistical inference as the experiment." It might not give the same benefit-cost conclusion for the program, however. If program costs equal $600 per participant, for example, the nonexperimental approach would show the program to be a good social investment (with a net social benefit of $100 per participant) while the unbiased experimental estimate provides the opposite conclusion (a net social loss of $100 per participant).

The basic problem here is that using similarities or differences in policy conclusions as the litmus test for nonexperimental estimators may give different answers depending on the policy question posed. An estimate that is close enough to the experimental benchmark to agree with it on one policy question may not do so on another. Moreover, in considering the reliability of a nonexperimental method in future applications, one cannot stipulate what policy question will be of greatest interest. It might be whether the program increases earnings, as in the case considered by Heckman and Hotz and by Couch. It might be whether the program produces net social benefits, as a number of other analysts have proposed. Or it might be another question altogether, such as whether the program produces net budgetary savings or raises earnings more than an alternative policy (e.g., wage supplements).

Clearly, what is needed here is an assessment tool that considers the correspondence of policy conclusions between the experiment and possible nonexperimental estimators over a range of policy questions.

The AFDC WIN Demonstration Analyses

One recent paper begins to move the literature on model validation in this direction. In their exploration of alternative comparison-site methods for measuring the effects of mandatory work-welfare programs, Friedlander and Robins (1992) for the first time exposit specific criteria for comparing experimental and nonexperimental estimates. Three criteria are stated, including one that is new:

- The magnitude of the numerical difference between the two estimates;

- Agreement between the two estimators on the statistical inference regarding the null hypothesis of no effect; and
- The statistical significance of the difference between the estimates.

We have already reviewed the strengths and weaknesses of the first two of these validation criteria. Friedlander and Robins apply these criteria much more comprehensively and uniformly than earlier authors. They also for the first time establish explicit standards for judging the magnitude of numerical differences under the first approach, although the standards they use are necessarily judgmental and subject to debate (as they themselves acknowledge). The authors also explore the limitations of the second criterion—the test of agreement of statistical inferences—but focus on a different point than those raised above. As they point out, the test is artificially more accepting of nonexperimental estimates derived from small samples than of nonexperimental estimates derived from large samples.[3]

Perhaps the most important of Friedlander and Robins' many contributions is their decision to test for statistically significant differences between experimental and nonexperimental estimates. This is a logical extension of the first validation criterion, the simple comparison of magnitudes. Except for a passing mention by LaLonde, none of the earlier papers that focus on magnitude comparisons even mentions the possibility of testing the experimental/nonexperimental differences on which it focuses.[4] Of course, measured differences do not give the exact bias in the nonexperimental estimator, since the observed data only approximate the degree of selection bias due to chance sampling, outcome, and measurement error. Friedlander and Robins make these uncertainties explicit and central to their analysis, insisting that observed differences in magnitudes between the two types of estimators exceed the bounds of chance variation before concluding that a particular nonexperimental approach suffers from selection bias.

For reasons that we explore more fully in the next section, this important advance does not yet produce a fully satisfactory means of identifying reliable versus unreliable nonexperimental estimators. One major problem is recognized by the authors themselves but not remedied: like any test for statistically significant differences between two measures, the technique is more likely to accept the nonexperimental estimates as equivalent to the experimental estimates when samples are small than when they are large. As Friedlander and Robins emphasize, the ability of the bias test to detect any given degree of bias depends on the size of the data samples employed. All other things equal, larger samples lead to greater statistical certainty and, therefore, more discerning tests for selection bias.[5]

A more fundamental issue—which all of Friedlander and Robins' validation criteria share with those used by earlier analysts—concerns *the assumption that no bias is present until the data prove otherwise,* rather than the reverse. We consider this assumption to be unwarranted in an exercise arising out of concern for the near-universal presence of substantial selection bias among nonexperimental estimators. To address this concern, we propose a reformulation of the validation question to strike a better balance between proving what does not work and demonstrating what does.

Reframing the Validation Question

Ultimately, the search for a nonexperimental means of reliably measuring the effects of employment and training programs has to address two questions with regard to any particular candidate measure:

1. Is there compelling evidence that the candidate measure suffers from an unacceptable amount of selection bias? If so, the measure should be rejected out of hand as an inadequate substitute for an experiment.

2. If not, is there compelling evidence that selection bias, assuming it exists at all, is necessarily confined to an acceptable level? That is, having failed to prove the *existence* of substantial selection bias, can we prove the *absence* of bias of any appreciable magnitude?

The answers to these questions will tell us which, if any, of the applicant-based impact estimates presented in chapter 3 can be viewed as adequate substitutes for the experiment.

In prior validation efforts, attention has focused on question 1—proving the existence of bias. The most direct such proof—testing for a statistically significant difference between the experimental and nonexperimental estimates, as in Friedlander and Robins—begins by postulating that no selection bias exists and then looks for proof that it does. Where proof of bias is found—and the magnitude of the measured bias is of substantive importance—this sort of test is quite helpful, answering question 1 in the affirmative and making question 2 moot. When simple to carry out, such tests can serve well as the first—and possibly only—step in the analysis.

However, tests of question 1 are not simple to carry out when dealing with voluntary employment and training programs. In that circumstance, the calculation of the appropriate test statistic becomes substantially more complex than in the case of mandatory programs like those studied by Friedlander and Robins. The reasons are explained in appendix C, which explores the implications of the voluntary/mandatory distinction for the current study and in relation to past validation analyses.

Given the complexities involved in addressing question 1 for voluntary programs, we concentrate instead on question 2—proving *lack* of bias. In many instances, this question will become paramount even when question 1 is considered: an impact estimate that is not *proven* to be biased in a test of question 1 may still *be* biased to an unacceptable degree, if the data are too weak to provide conclusive evidence.

To protect against this possibility, we would turn the test objective around, passing over question 1 to address question 2 as the first step. This reflects our belief that the reliability of any nonexperimental method must be proven before that method can be recommended for use. The real objective is not to prove some estimation methods biased and let the others "pass," but to prove that the bias associated with one or more nonexperimental methods is nonexistent or, at worst, falls within an acceptable margin. Only then can we consider a method an acceptable alternative to an experiment in terms of selection bias.

The next several sections of the chapter are devoted to developing a statistical framework for answering question 2—proving the absence of substantial selection bias. This development could take many forms, several of which we explore. All of the alternate frameworks examined involve complex statistical concepts that some readers may wish to skip, since ultimately we choose a framework that can be understood intuitively without benefit of the statistical arguments.

The Classical Approach to Proving Absence of Bias

Our goal, then, is to use the standard tools of classical statistical analysis to formulate a test of question 2. We begin by considering the simplest possible approach: basing a proof of "no selection bias" on a simple comparison of magnitudes between the nonexperimental and experimental estimates. As noted earlier in connection with question 1, this strategy overlooks the fact that the difference between the nonexperimental and experimental estimates is only an approximation of the underlying bias in the nonexperimental estimate, which is defined as the average difference in the two measures in repeated applications of the technique.[6] To take appropriate account of the approximation (or estimation) error, we must use statistical tests, not simple comparisons of magnitude.

There are two routes one could take to conducting a statistical test of question 2:

- Maintain the null hypothesis of no selection bias, but alter the test procedure so that the null is rejected rather than accepted when the data are inconclusive; or
- Switch the null hypothesis to one that says substantial selection bias is present, so that inconclusive results that allow the null to stand will again lead to a finding of bias.

We develop each of these strategies in detail below. When both prove less than fully satisfying, we turn to an alternative strategy that relies on Bayesian tools of inference in place of classical hypothesis-testing methods.

Changing Emphasis in the Original Test Procedure

The most obvious way to address question 2 is to repeat the test procedure others have used for question 1, but shift priorities between the two hypotheses. As noted earlier, the original test procedure insists on strong contrary evidence before rejecting the null hypothesis of "no selection bias" in favor of the alternative hypothesis of "substantial selection bias." To address question 2, we should turn this emphasis around, requiring strong evidence that the null hypothesis is right before accepting it, and overturning the null in all other cases. We can accomplish this result by changing the significance level of the test.

Normally, one chooses a rule for accepting or rejecting the null hypothesis that gives a correct null hypothesis little chance of rejection. Statisticians refer to the probability of rejecting a correct null as the "significance level" of the test. Low significance levels, on the order of .05 or .10, are appropriate when one needs compelling evidence that the null is *wrong* before rejecting it. Rejecting the null at the .05 or .10 significance level provides a strong assurance that the null is wrong, since the probability of rejection when the null is *right* is only .10 or below.

The desire to be quite sure before rejecting the null hypothesis, while common in statistical practice, is not always appropriate. One could instead choose to be quite sure before allowing the null to stand, as we wish to be with a null that states that a nonexperimental estimator has no appreciable selection bias. We need to make the null harder to accept when false, not harder to reject when true. In so doing, however, we also make it harder to accept the null when true. Thus, we can only ensure probable rejection of false nulls by raising the significance level of the test above .10.

How far should we raise the significance level of the test to tip the balance in favor of the alternative hypothesis by an appropriate amount? Presumably, until the probability of rejecting the alternative hypothesis when it is false drops to .10 or below, according it the status usually reserved for the null. This will protect against accepting the null too quickly, and do so to the same degree that one normally protects against rejecting the null too quickly.

This result can be expressed in terms of statistical power. The "power" of a statistical test is the probability of rejecting the null hypothesis when the null hypothesis is false. We are looking here for a test with high power—a test very likely to reject the null hypothesis of "no selection bias" when selection bias is present. The test just described—one which rarely *accepts* a false null hypothesis—necessarily *rejects* a false null in almost all cases. Thus, our goal can be restated as raising the significance level of the test far enough that the power of the test—the odds of rejecting the null hypothesis when false—reaches .90.

A complication arises at this point, however—one very familiar to statisticians: the power of a test depends on the true level of the parameter in question—here, the amount of selection bias. Thus, the odds of rejecting a false null hypothesis depend on which specific alternative to the null actually holds. With true selection bias of, say, $100, a test procedure will be less likely to reject the null hypothesis of no selection bias than when the true selection bias is $1,000, since the latter situation is much more likely to generate measured differences in estimates of a magnitude that will lead to rejection.

One can address this problem in two ways. First, one can establish a degree of selection bias that one is willing to tolerate and then choose a test procedure with high power to detect bias of at least that magnitude. Alternatively, one could conduct a series of tests, each yielding 90 percent power at a specific level of true bias, and then somehow convert the range of test results into an overall conclusion regarding the method. Both of these strategies—setting a bias tolerance limit and synthesizing a range of test results—present problems that we discuss below in connection with the strategy of reversing the null hypothesis.

Reversing the Null Hypothesis

A second classical approach to question 2 is to reverse the null and alternative hypotheses used in the original test of question 1, leaving the significance level and power of the test unchanged. Here, instead of assuming that selection bias equals $0 (the null hypothesis under question 1), assume that selection bias is not $0 (the null hypothesis under question 2).

Unfortunately, this new null hypothesis is too broad to serve as the basis for constructing an informative test. In general, statistical tests are formulated by assuming that the null hypothesis is true. Based on this assumption, one can tell what sorts of estimates to expect from the data. If those expectations are not met (e.g., when the experimental and nonexperimental estimates differ markedly in a test of question 1), one rejects the null hypothesis. In this instance, we have a broad null hypothesis that says that selection bias is some positive or negative number, a situation that leaves open too many options to have clear implications for the data. (For example, the null could be true because bias is only slightly different from $0, making small or zero differences between the nonexperimental and experimental estimates fairly likely. Or it could be true because bias is quite large or very negative, making extreme differences reasonably likely). Testing a null hypothesis this broad assures acceptance, and amounts to simply assuming that the null is true.[7]

To remove this problem, we must change the null hypothesis to one that can be meaningfully tested—a narrower null that states that selection bias is confined to a smaller range than all points other than $0. A null that puts bias at or above some threshold level, say plus or minus $200, seems the most sensible. If the data reject this null hypothesis, we have found a nonexperimental method whose selection bias almost certainly does not exceed $200 in absolute value—and can declare it an acceptable alternative to an experiment if we are willing to tolerate bias up to that limit. The essence of this approach is to define an acceptable level of selection bias and then test whether actual bias exceeds that level.[8]

A Confidence-Interval-Based Test Procedure

We now have a null hypothesis for which it is possible to construct a meaningful test: that selection bias exceeds $X. There may be more than one way to test this hypothesis; we describe one such approach below, based on the concept of a "confidence interval."[9]

In classical statistics, a confidence interval represents the range in which an unknown parameter is thought to lie. One determines that range by moving far enough beyond the point estimate of the parameter on either side to encompass the true value of the parameter with great certainty. The standard error of the estimate determines how far out one must go to reach that objective.

Consider the confidence interval for true selection bias implied by the difference between a nonexperimental and experimental impact estimate. Like all confidence intervals, such an interval is defined so that we can be quite confident that the true selection bias falls somewhere between its endpoints. Following the usual standard, let us suppose that the interval is broad enough to have a 90 percent chance of containing the true level of selection bias.

To take a specific numeric example, suppose that the 90 percent confidence interval for selection bias for one of the nonexperimental impact estimates in chapter 3 centers on $50 and ranges from $-75 to $175.[10] In this instance, we would be fairly certain that the true bias is confined to this interval and, therefore, that it does not exceed $200 in either direction. If $200 is our tolerance limit, we would be willing to reject the null hypothesis of substantial bias and accept the approach as a reasonable alternative to the experiment. In other instances—say, any time the 90 percent confidence interval for true bias extended more than $200 above or below $0, we would not reject the null hypothesis, concluding instead that the nonexperimental method entails an unacceptable risk of selection bias.

This sort of calculation, while statistically complete, raises a major practical issue introduced earlier in thinking about other statistical approaches: How does one determine how much selection bias to tolerate? The $200 figure in the preceding example was chosen arbitrarily and is but one of many possibilities. Presumably, the choice of a "tolerance level" of this sort will hinge on how much selection bias one is willing to risk in relying on a nonexperimental impact estimator in place of an experiment. This question has different answers under different circumstances.

Choosing a Tolerance Limit

In general, the degree of accuracy required in measuring program impacts will depend on the context. Three factors stand out in defining the context:

- The type of decisions to be made by policy makers when using the impact estimate;
- The proximity of true impact to the level of impact on which those policy decisions will hinge; and
- The degree of risk the policy maker is willing to accept that the wrong decision will be made because of selection bias.

An extension of our earlier example illustrates the importance of these three factors when deciding on an "acceptable" level of selection bias.

Suppose that policy makers want to expand or contract a specific training program depending on whether the earnings gains it produces exceed or fall short of the program's costs. In particular, suppose the program costs $500 per trainee, so that a true impact above $500 would justify its expansion while impacts below that level would not. Suppose also that the program's *true* impact on earnings is $400, so that a positive bias of $100 or more would make the program appear to have met the $500 standard when, in fact, it did not.

Here, policy makers should be unwilling to rely on the nonexperimental impact estimator from our earlier example, which was proven to be free of selection bias in excess of $200 but not necessarily free of smaller biases. Such an estimate could be biased upward by as much as $100 or $200, in which case it would produce an impact estimate above $500 and make the program appear to provide gains in excess of costs when it does not. If instead policy makers were prepared to support the program if it produced *any* noticeable earnings gain—say one of $100 or more—a method that comes within $200 dollars of the true impact of $400 would be fully reliable for policy purposes.

The same method would also be acceptable when earnings gains above $400 are required but the true impact of the program is far different from $400, say $0 or $1,000. Here, an error of $200 or less either way could not lead to the wrong policy decision.

Finally, under the original scenario—true gains of $400, required gains of $500, and a margin of error of plus or minus $200—a policy maker might still be willing to risk an erroneous policy decision if all that is at stake is a modest program expansion, not an overall decision on continuing or discontinuing the program.

Accommodating a Range of Tolerance Levels

It is clear from these examples that changing any of the factors that influence policy makers' tolerance for errors could alter the degree of selection bias one should test against. Since the exact level of each factor—the type of policy decision to be made, the size of the true impact in relation to the degree of selection bias, the policy maker's willingness to run risks—cannot be anticipated in advance, a whole range of tests are needed to appropriately gauge the performance of any nonexperimental method. In addition, a means of summarizing the conclusions of various tests (an infinite number, in principle) will be needed, if we are to arrive at a conclusion regarding the overall acceptability of a nonexperimental approach.

We will not attempt such an ambitious generalization of the classical statistical approach in this monograph. Rather, we will start over in the next section, not with any specific statistical procedure in mind, but with the sense that, to be useful, *tests for an acceptable level of selection bias must reveal the likelihood of incorrect policy decisions over a range of policy situations.*

A Bayesian Approach to Model Validation

We begin the development of an alternative approach for validating nonexperimental estimators by rephrasing our objective in terms of policy reliability: If a nonexperimental method can be shown to provide reliable policy conclusions over a range of policy questions, the initial presumption that the nonexperimental measure is not an adequate substitute for the experiment can be overturned. If not, the presumption remains.

At its barest fundamentals, this is the framework we will use to validate or refute nonexperimental impact estimates throughout the rest of the monograph.

It is worth noting that elements of this same framework have appeared previously in other studies. Heckman and Hotz (1989), Couch (1991), and Friedlander and Robins (1992) all examine the reliability of the policy conclusions provided by nonexperimental methods in relation to an experiment for a single policy question: Do impacts exceed zero? Under the more general approach proposed here, an acceptable nonexperimental approach must provide reliable policy conclusions across a *range* of potentially relevant policy questions, not simply with regard to the question of whether impacts exceed zero.

To examine each nonexperimental estimation approach in relation to a range of policy questions, we must first formulate a model of how policy decisions are made.

A Policy Decision Rule

We begin by assuming that policy decisions about employment and training programs are made on the basis of policy makers' beliefs about the size of true program impacts. Specifically, we assume that these decisions hinge on whether the true impact is above or below some critical value. Letting I represent a program's true impact (on, say, participant earnings), we might imagine that policy actions such as future funding decisions and changes in program design depend on whether policy makers believe:

$I > C$,

where C is the minimum level of achievement required to justify the action under consideration.[11]

For example, in some cases $C = 0$, as when policy makers believe that a program that increases participant earnings by any amount ($I > 0$) is worth supporting. Alternatively, the level of impact needed for policy support may equal the per-participant cost of running the program; this would be the case if the crucial consideration is whether the program produces earnings gains sufficient to justify its cost to taxpayers. Or C may represent the level of earnings gain that could be accomplished for the same individuals by some other policy change, such as a wage supplement; here, policy makers will continue to support the training program only if they think it can raise participant earnings by more than the next best policy tool available to them.

In the ideal world, one would pick an impact estimation technique according to how well it distinguishes between earnings gains above the cutoff level, C, and earnings gains below C. Alternatively, policy makers might put top priority on accurate distinctions in a particular direction (e.g., favor techniques that never show impacts above C when true impact is below C, but which can make the reverse mistake) or on avoiding "large" incorrect distinctions (placing impact on the wrong side of C when true effects are far from C). Whatever the priority, it is always the relationship of true impact to a cutoff value, C, that will drive the policy decision.

Unfortunately, one cannot normally anticipate this "make or break" impact level when choosing evaluation methods. Absent knowledge of C—which, as we have seen, may vary for different policy makers and/or policy decisions—evaluators have no alternative but to choose an estimation approach based on its performance across some *range* of relevant *C* values. We can do this by considering the likelihood that the average result of an impact estimation technique will give the "right" answer from a policy maker's perspective for various *C* values, where the right answer is "Yes, support the policy" when the true impact of the program exceeds *C* and "No, don't support the policy" when the true impact is less than *C*.

Using Current Data to Guide Future Policy Decisions

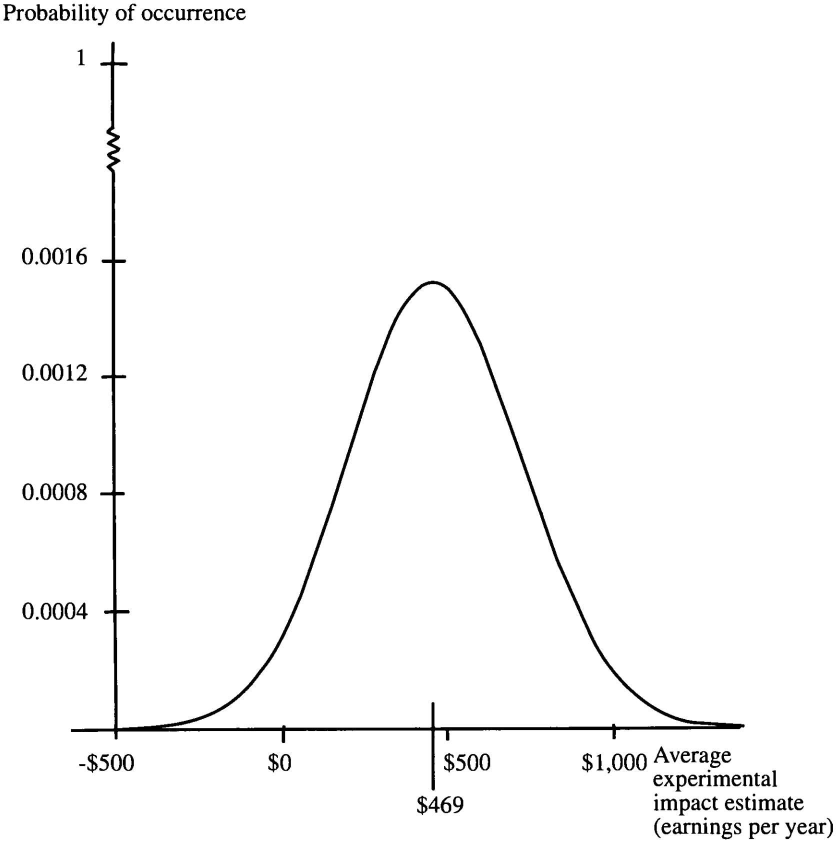
Our overarching objective, then, is to find a nonexperimental method that—in light of the homemaker demonstration data—looks like "as good a bet" as an experiment when making a variety of policy decisions. Absent empirical data, we were unwilling to "bet" on any nonexperimental method in this sense, since we believe that all nonexperimental estimators have the potential for serious selection bias until proven otherwise. But what should we believe about alternative options once we have examined the data, if we want that data to guide our choice of estimation methods in the future?

Statisticians have developed a formal framework for answering this question—i.e., for combining prior beliefs with current information to obtain a new set of beliefs on which to base decisions. Known as "decision theory" or "Bayesian statistics," this approach posits a distribution of the expected average result of a particular estimation method which recognizes that the true average result is unknown. A subjective probability is then attached to each of the values in the distribution, creating a probability distribution of beliefs across all possible average results.

Different subjective probabilities apply—beliefs change—once new information is taken into account. The homemaker demonstrations provide that information in this case, information regarding the likely outcome of using applicant-based impact estimation methods to guide policy. A fundamental theorem of Bayesian statistics states that, when one begins with an agnostic view of what the average result of an estimation approach might look like (a situation referred to as having a "diffuse prior"),[12] the probability distribution one should construct for the average result of that procedure based on a single application is centered on the value produced by that application.[13] Furthermore, when the observed estimate is drawn from a normal distribution (and, as before, one starts with a diffuse prior) the probability distribution of possible average values constructed from a single application should also follow a normal distribution, with standard deviation equal to the standard error of the available estimate.

Exhibit 4.1 provides one such set of beliefs regarding the average result from using the experimental method to measure the impact of the AFDC Homemaker-Home Health Aide Demonstrations six years after program entry (i.e., in 1990). The distribution in exhibit 4.1 is centered on the 1990 experimental earnings impact estimate from chapter 3 ($469), with standard deviation equal to the standard error of that estimate ($264).

Exhibit 4.1 Subjective Distribution of the Possible Average Outcomes from Using Experimental Methods to Measure the Sixth-Year Impact of the Homemaker Demonstrations



To summarize, then, by starting with an agnostic view of the possibilities (i.e., with a "diffuse prior") and observing the value and standard error of a single experimental impact estimate, we can formulate a "posterior distribution" of possible average values from that method, such as that shown in exhibit 4.1. The same is true of any nonexperimental impact estimation technique: its possible average result in repeated applications can be represented as a normal distribution centered on the one application achieved. Together, these two distributions provide a basis for assessing the policy reliability of alternative nonexperimental methods.

Before we can turn beliefs about the average outcomes of different estimation techniques into an assessment of the policy reliability of nonexperimental techniques, however, we must first adopt two simplifying assumptions. First, we must assume that experiments and nonexperimental approaches will exhibit an equal degree of sampling variability in future applications to abstract from this confounding factor in our analysis.[14] And we must assume that the future context of interest sufficiently resembles the current context to allow us to translate our results to other settings.[15]

Diagramming the Policy Decision

Bayesian posterior distributions for experimental and nonexperimental impact estimates speak directly to the policy question posed earlier: Does true impact, *I*, exceed or fall short of the critical policy cutoff, *C*? On average across repeated applications, experiments give the true impact, *I*. Thus, beliefs held about the average results of an experiment are also beliefs about true impact.

We can combine this information with beliefs about the average result of a nonexperimental technique, also expressed in the form of a Bayesian posterior distribution. The comparison of the two—posterior distributions for truth and the expected result of a nonexperimental method—give a measure of the risk of relying on the nonexperimental method in making future policy decisions. Exhibit 4.2 shows in graphic form how this measure is derived.

The top panel of exhibit 4.2 combines two specific posterior distributions of interest: one generated by the homemaker experiment in 1990, the other produced using the withdrawal-based nonexperimental method in the same year. The experimental distribution, labeled "True impact" and centered at $469, shows the appropriate subjective assessment of what the true demonstration impact might have been in that year, given the available evidence. The nonexperimental distribution, labeled "Average nonexperimental result" and centered at $2,074, gives the appropriate subjective assessment of the size of measured impact the withdrawal-based approach is likely to produce on average.

This panel has been set up to help us determine, for different critical policy cutoff values, the probability of obtaining "good" or "bad" policy results from the withdrawal-based approach, were there no sampling variability in the data. To see how this is done, consider a particular policy cutoff value, *C**, along the horizontal axis. When the true program impact exceeds *C**, a policy maker who cares about that critical value should be advised to favor the program. Conversely, when the true program impact falls below *C**, that policy maker should oppose the program. What are the odds that the withdrawal-based approach will yield the correct policy prescription in this circumstance, given our subjective assessment of the possible values of true impact and the possible average result of the withdrawal-based approach? These odds can be derived directly from the exhibit for one special subcase and put within certain limits for all cases. We show how in the next two subsections.

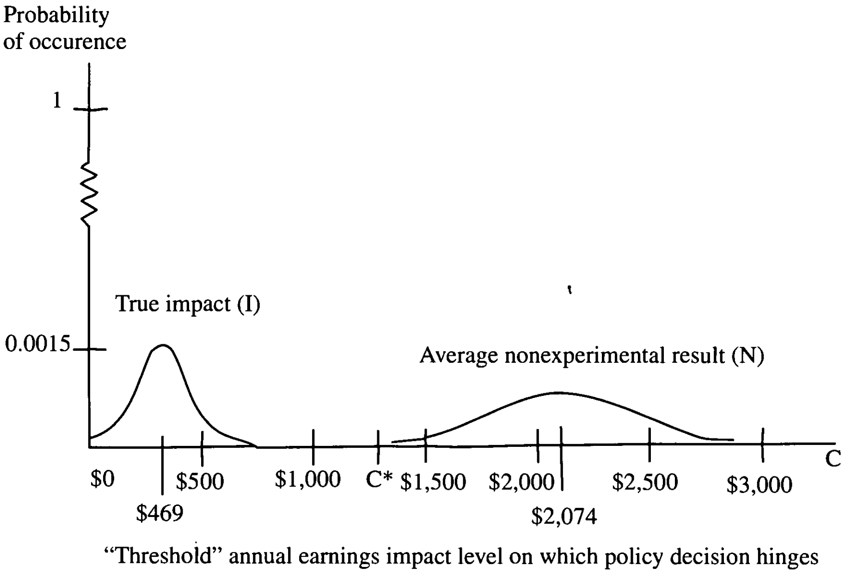
The Probability of "Right" and "Wrong" Policy Decisions

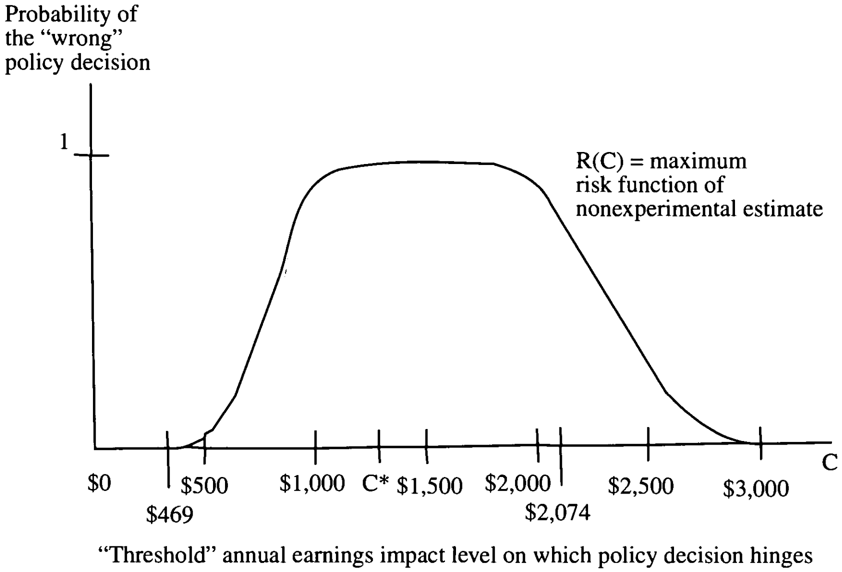
To determine the probability of correct versus incorrect policy decisions using the withdrawal-based estimator, we begin in the top panel of exhibit 4.2. Here, the "True Impact" curve gives our subjective posterior assessment of the probability that true impact lies on either side of a given policy cutoff value, *C**. When the true impact is below *C**, opposing the program is the "right" policy decision; when true impact is above *C**, supporting the program is "right" from the policy maker's perspective. The probabilities of these two events, in our subjective posterior distribution, are determined by the areas under the "True impact" curve on either side of *C**, which can be represented in probability terms as:

$Pr\ (I < C^*)$ = probability that opposing the program is the "right" decision; and

$Pr\ (I > C^*)$ = probability that supporting the program is the "right" decision.

Exhibit 4.2 Subjective Distribution and the "Maximum Risk Function" of Experimental and Withdrawal-Based Impact Estimation Techniques in 1990





The other curve in the upper panel, labeled "Average nonexperimental Result," tells us the odds that the nonexperimental estimator under consideration (the withdrawal-based estimator for 1990) takes on any particular value on average across repeated applications. Absent experimental evidence, the expected value of this procedure (plus some random departure due to statistical variation in the data) would be the only basis for making policy reflective of program impacts. The crucial question, then, is whether policy so formulated is likely to be "right" or "wrong" from the policy maker's point of view.

To make that determination, we must first consider the probability that policy makers relying on the nonexperimental approach would oppose or support the program in relation to a particular cutoff value, C^*. These probabilities, independent of sampling variation, are determined by the area under the "Average nonexperimental result" curve to the left and right of C^*. Using N to represent the nonexperimental estimate, the two areas are:

$Pr(N > C^*)$ = probability that the nonexperimental approach will lead to support of the program, on average; and

$Pr(N < C^*)$ = probability that the nonexperimental approach will lead to opposition to the program, on average.

The risk, R, that the nonexperimental estimator will lead to the wrong decision is:

$$R(\mathrm{C}^*) = Pr\,(N < C^* \text{ and } I > C^*) + Pr\,(N > C^* \text{ and } I < C^*).$$

We call this formula, traced out over a range of C^* values, the "risk function."[16]

In the special case of zero correlation between N and I,[17] this formula reduces to:

$$R(C^*) = Pr\,(N < C^*) \bullet Pr\,(I > C^*) + Pr\,(N > C^*) \bullet Pr\,(I < C^*)\,.$$

All of the terms in this equation can be calculated from the impact estimates presented in chapter 3 and their standard errors. The bottom panel of exhibit 4.2 shows the result of these calculations—the risk run in relying on a nonexperimental impact estimate when making policy decisions, relative to an experiment with the same sampling error—across a range of policy cutoff values, *C*.

Dealing with Correlation in Posterior Beliefs

The risk function shown in exhibit 4.2 was derived under the assumption that our beliefs about the true impact of the program and about the average result of the nonexperimental estimator are uncorrelated. In the most general case, our assessment of the odds of a "correct" policy decision should depend on our beliefs about the joint behavior of the experimental and nonexperimental impact estimates. That is, what matters is not what we believe about each of the two estimates individually, but what we believe about the likely relationship between them—their *joint* posterior distribution. The upper panel in exhibit 4.2 does not show this added aspect of our beliefs, only each set of beliefs in isolation (what statisticians call *marginal* distributions, derived from the joint distribution). If the two sets of beliefs are in fact unrelated, or uncorrelated, the two marginal distributions summarize everything there is to know about our posterior beliefs, individually and in tandem.

More likely, the two sets of beliefs are not unrelated. Rather, a more reasonable set of posterior beliefs would associate relatively large values of the experimental estimator (i.e., large true impacts) with relatively large values of the nonexperimental estimator. We believed at the outset that the experimental estimate gives the true impact on average (see chapter 1) and that the nonexperimental estimate might get us close to that ideal (see chapter 2). Thus, we necessarily believed that the two estimators are positively correlated, and we likely will continue to believe that in forming our joint posterior distribution once we have looked at the data.

While our posterior beliefs are likely to incorporate positive correlation between true impacts and the average nonexperimental result, the zero-correlation special case shown in exhibit 4.2 still proves to be a useful one. It can be shown (see appendix D) that this special case provides an upper bound on the probability of making the "wrong" policy decision when relying on the nonexperimental estimator.[18] Were our beliefs uncorrelated, working through the special case would give the correct probability of policy error; given positive correlation, the correct probability is necessarily below that produced by assuming no correlation.

In fact, an upper bound is all that we can hope for in the general case. To do better—to calculate the exact probability of policy error in the face of correlated beliefs (independent of sampling error)—would require that we posit a specific numerical value for the correlation between true impact and the nonexperimental estimator in our posterior distribution. Each researcher and policy maker seems likely to hold his or her own distinctive beliefs on the subject, making it impossible to formulate a convincing general case.[19] Fortunately, having an upper bound on the probability of incorrect policy decisions can provide the sort of validation findings we need: evidence that a particular nonexperimental estimator holds the consequences of selection bias to a tolerable level. If the upper bound itself seems tolerable, the true probability of policy error must also be tolerable (or better).

Properties of the Maximum Risk Function

The "maximum risk function" traced by $R(C)$ in exhibit 4.2 follows a characteristic shape. As can be seen from the above formula, the maximum risk level associated with any value of C is determined by multiplying two probabilities that sum to 1.0, $Pr\ (N < C)$ and $Pr\ (N > C)$, times two other probabilities that sum to 1.0, $Pr\ (I > C)$ and $Pr\ (I < C)$. A sum of products of this sort is always largest when the two larger pieces are multiplied together, i.e., when both $Pr\ (N < C)$ and $Pr\ (I > C)$ are greater than .5, or when both $Pr\ (N > C)$ and $Pr\ (I < C)$ are greater than .5. This can only happen when the distributions of N and I in the upper panel of exhibit 4.2 lie some ways apart, placing one curve substantially to the left of C* and the other substantially to the right. For the distributions shown, this situation arises for values of C between $500 and $2,000, where most of the distribution of I lies to the left of C and most of the distribution of N to the right. As a result, the maximum risk index, $R(C)$, shown in the bottom panel of the exhibit is quite large in this range, but nowhere else.

Intuitively, this result makes sense in two respects:

- Experimental and nonexperimental distributions far from one another are produced by impact estimates far from one another, implying a large potential risk of making a misguided policy decision if one relies on the nonexperimental estimate; and

- Even when a nonexperimental approach produces a very different impact estimate—and, therefore, a probability distribution far from that of the experiment—it carries a high potential risk of incorrect policy decisions (a risk above, say, .2) only for decisions that hinge on deciding whether true impacts are below or above a "threshold value" in the interval between the two estimates.

The first of these intuitions is not new. In relates to the original reason authors like LaLonde, and Maynard and Fraker, began to worry about using nonexperimental estimators for policy purposes: the observation that experimental and nonexperimental impact estimates often lay far from one another in numerical terms. Indeed, it is the separation of these two estimates that moves the *N* and *I* distributions in the top panel of exhibit 4.2 so far apart, resulting in the sharp, sustained increase in the maximum risk of misguided policy decisions for critical values of *C* between the two estimates.

The second intuition has less precedent. It focuses attention on a point that some earlier studies overlooked: that the reliability of an estimation technique for policy purposes depends critically on the distinctions policy makers need to make. For many purposes (e.g., distinguishing no impact from a positive impact, C=0), a nonexperimental measure far from the experimental benchmark can do little harm. On the other hand, such an estimate may be very detrimental to policy decisions that hinge on the presence or absence of large positive impacts, in the $1,000–$2,000 range (e.g., conclusions from a benefit-cost analysis). Exhibit 4.2 vividly illustrates the dangers of focusing attention on any one such question. When other evaluators adopted such a narrow focus in comparing all impact estimates to $0, they essentially ignored all of the evidence in the exhibit regarding model reliability to the right of $0! As noted earlier—and illustrated dramatically here—such a narrow perspective can lead to highly misleading characterizations of the reliability of a nonexperimental estimation technique in relation to other relevant questions.

A further case not shown in the exhibit also deserves mention: the behavior of the maximum risk function when the experiment and the nonexperimental approach produce the same impact estimate. Here, the risk involved in relying on the nonexperimental approach still exceeds zero over a range of *C* values near that common estimate. This is due to the uncertainty attached to the impact estimates, which forces us to consider that either one could be wrong and, therefore, spreads our probability assessments beyond the matching point estimates. In this sense, a nonexperimental technique that produces a result identical to that of the experiment is not a perfect substitute for an experiment, since we know the experimental estimate is free of selection bias but can only speculate on the chances that the same is true of the nonexperimental estimate. The greater our uncertainty (i.e., the larger the standard errors of the experimental and/or nonexperimental estimates), the broader the band where nontrivial risks of inaccurate policy decisions can arise even in this "best case" scenario. Indeed, when policy makers need to know whether impacts exceed or fall short of a *C* just equal to a common estimated level of impact, the maximum risk function reaches a peak value of .5.[20]

Internal versus External Validity

Before preceding to apply the risk-function approach to model validation, we need to add one final but vital point. Even when a nonexperimental method is shown to carry little "risk" in the sense discussed here, we will have established that method as an acceptable alternative to an experiment only in the particular context where the test was run.[21]

Thus, even if one or more of the applicant-based estimators examined here passes the validation tests we employ, we cannot claim to have found an applicant-based technique that will be as reliable as an experiment in all contexts. Rather, we will have hard evidence of the internal validity of the method: what worked in this particular instance and, therefore, what *might* work in evaluating future programs under similar circumstances. Just how similar those circumstances would have to be cannot be determined until the same technique is tested against the experimental norm in other contexts.[22] We do not know, for example, whether the same estimator would work when evaluating programs involving other types of clients who might exhibit different self-selection behavior, or under different administrative structures where program selection rules might differ.[23] Only when these questions are answered can we begin to gauge the method's external validity or its likelihood of success under a wide range of external circumstances.

Conversely, we should not abandon an approach just because it does not meet our required standards of evidence in the context of the AFDC Homemaker-Home Health Aide Demonstrations. Rather, we should view our results here as the first of a series of possible tests of the applicant-based approach across a variety of different situations.[24] In the meantime, the validation results obtained could still have considerable value to those designing future evaluations. When deciding between alternative research designs, future evaluators can consider how strongly their situation resembles the homemaker demonstrations and how compelling the combined theoretic rationales and empirical findings presented here are with regard to specific nonexperimental methods, and then make an informed judgment about the most appropriate approach. These judgments should be much better founded by virtue of our initial series of tests, just as later judgments and design decisions can be made more reliable by still further testing.

Findings on Model Reliability

To test the policy reliability of the nonexperimental estimators from chapter 3, we use the framework just developed to calculate two sets of maximum risk functions for each of our applicant-based nonexperimental estimation techniques:

- Maximum risk functions for measured effects in the mostly in-program years of 1985 and 1986, using impact estimates for average annual earnings over those years; and
- Maximum risk functions for measured effects in the post-program years of 1987 through 1990, again using impact estimates for average annual earnings.

Impact estimates and standard errors for each year of the follow-up period and for these two multi-year periods appear in exhibit 4.3, with results for the different nonexperimental methods and the experiment shown in separate columns. To hold the number of risk functions examined to a manageable level, we base our model validation tests on the multi-year estimates in the bottom panel. The top panel of year-by-year estimates is repeated here for reference, and includes the same point estimates as exhibits 3.3, 3.4, and 3.5 in chapter 3.

In multiyear format, annual earnings impacts are estimated at \$1,600–\$2,700 for the in-program period and \$500–\$1,200 for the postprogram period, depending on the estimation method used. These estimates do not always exactly equal the average of those for the individual years due to the use of covariates in the impact regressions.[25]

Relative Risks

The maximum risk functions associated with each of the multiperiod nonexperimental impact estimates in exhibit 4.3 appear in exhibit 4.4 (for the in-program period) and exhibit 4.5 (for the postprogram period). The equivalent risk function for the experiment in each instance is just the horizontal axis, since *on average* experimental impact estimates cannot produce incorrect policy decisions due to selection bias, regardless of the policy question asked (the value of C considered).[26]

As can be seen in exhibit 4.4, all of the nonexperimental estimation methods examined—those using withdrawals, screen-outs, and no-shows as their comparison samples—pose substantial risks of poor policy decisions during the in-program period over some range of threshold values that policy makers might care about. Strikingly, however, none poses any risk of misguided policy due to selection bias over other wide bands of policy concerns. For example, if policy makers need to know whether the homemaker demonstrations produced in-program earnings gains above a threshold level of, say, \$1,000 per year, all three nonexperimental techniques carry almost no risk of misguiding policy due to selection bias alone.[27] This conclusion holds for any threshold level below around \$1,200 per year.[28] The same is true of policy decisions that hinge on whether or not in-program impacts exceeded or fell below a threshold of \$3,300 or more per year. But over varying intervals between \$1,200 and \$3,300 per year, all three nonexperimental approaches carry real risks of misguiding policy due to selection bias. Of the three, the no-show-based approach entails by far the narrowest band over which such risks could be substantial.

Exhibit 4.3 Earnings Impact Estimates, by Method, for Individual Years and for the Average of the In-Program and Postprogram Years (standard errors in parentheses)

	Nonexperimental comparison group						Experimental	
Year	Withdrawals		Screen-outs		No-shows		impact estimate	
1984: Mostly preprogram year	$ 121	(259)	$ 183	(144)	$ 26	(145)	$ 116	(86)
1985: In-program year	3,422	(341)	3,401	(257)	2,100	(218)	2,610	(139)
1986: Mostly postprogram year	2,003	(454)	1,802	(295)	1,145	(288)	1,194	(158)
1987: Postprogram year	635	(539)	1,030	(313)	916	(331)	643	(183)
1988: Postprogram year	800	(599)	1,014	(359)	706	(378)	583	(223)
1989: Postprogram year	1,316	(734)	696	(429)	814	(404)	394	(244)
1990: Postprogram year	2,074	(731)	768	(463)	867	(425)	469	(264)

	Nonexperimental comparison group						Experimental	
Period	Withdrawals		Screen-outs		No-shows		impact estimate	
In-program: 1986–86	$2,713	(355)	$2,601	(245)	$1,623	(222)	$1,902	(132)
Postprogram: 1987–90	1,214	(571)	890	(334)	809	(325)	517	(202)

Exhibit 4.4 Maximum Risk Functions of Alternative Nonexperimental Approaches: Annual Earnings in the In-Program Period

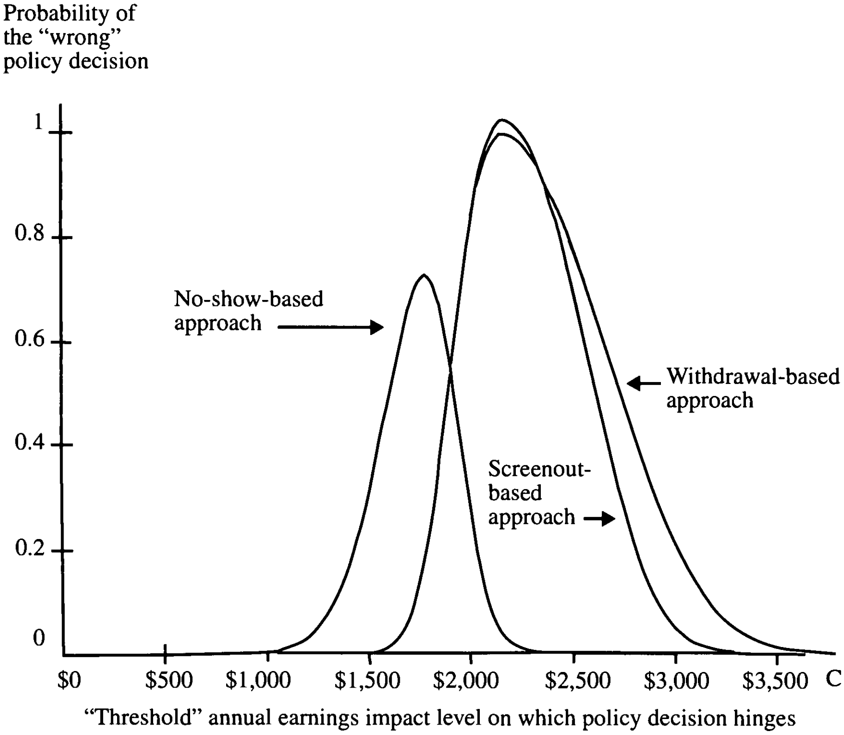
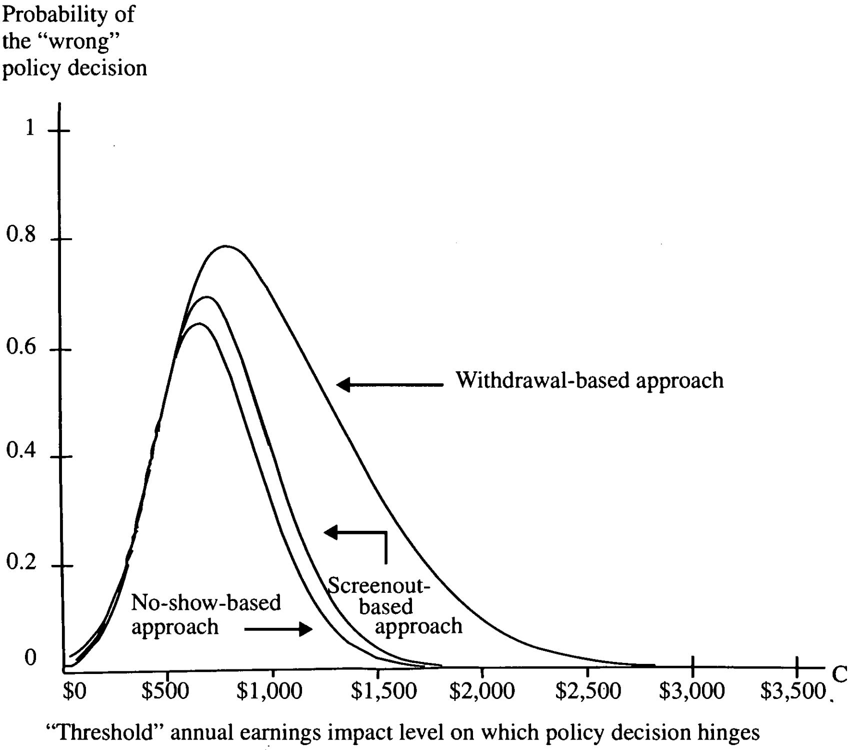


Exhibit 4.5 Maximum Risk Functions of Alternative Nonexperimental Approaches: Annual Earnings in the Postprogram Period



The situation in the postprogram years is much the same, as seen in exhibit 4.5. Several differences are apparent, however. First the interval of nontrivial maximum risk now begins at around $200 of impact per year rather than $1,200, and ends at around $2,200 per year rather than $3,300. Second, it is possible to rank the nonexperimental techniques from best to worst in terms of maximum selection bias risks across all policy questions. For *all* values of *C*, the screen-out-based and no-show-based estimates carry a lower maximum risk of misguided policy due to selection bias than the withdrawal-based estimates. The band of substantial potential risk is much narrower in each instance, particularly for the no-show-based estimates. This result follows from the fact that the no-show-and screen-out-based impact estimates shown in exhibit 4.3 are much closer to the experimental estimate in the postprogram period than is the withdrawal-based estimate. Exhibit 4.3 also indicates a substantially larger standard error for the withdrawal-based estimate over that period, which translates into greater uncertainty in our posterior distribution regarding the expected results of this technique. All other things equal, an impact estimation technique whose average outcome is less certain *a priori* carries with it a greater risk of misguided policy choices if adopted in future studies.

The ranking of methodologies during the in-program years is not so clear-cut. Here, the no-show approach appears to be the riskiest of all approaches for making policy decisions that hinge on threshold levels of $1,000–$1,900 per year. But for thresholds above $1,900, up to almost $3,500, the no-show approach carries less risk of selection bias than either of the other two nonexperimental options. Which approach would have been the better choice absent the experiment depends, then, on what level of impacts policy makers need to be able to discern during the in-program years. As before, the screen-out-based approach carries a lower maximum risk of selection bias error than the withdrawal-based approach across all possible policy questions except for a narrow range between $2,000 and $2,300.

Absolute Risks

While generally decisive with regard to what is knowable about the relative reliability of the three nonexperimental approaches tested, the maximum risk functions in exhibits 4.4 and 4.5 are primarily intended as guides to the *absolute* risks involved in using each of the nonexperimental methodologies. Returning to our most fundamental validation question, then, we again ask:

- Do any of the nonexperimental approaches tested produce policy conclusions of comparable reliability to those of the experiment over a broad range of policy questions?

In one sense, the answer must be "no," since any nonexperimental approach whose average result is uncertain (i.e., which has nonzero variance in the sample from which posterior distributions are formed) will create a serious risk of selection bias for at least some policy questions. Perhaps it is more meaningful to ask whether a particular nonexperimental approach confines this "unavoidable" level of risk to a sufficiently narrow—and, from a policy perspective, remote—band of *C* values.

While necessarily a judgmental decision, we are inclined to view the no-show-based strategy developed here—and, for the postprogram period, the screen-out-based strategy—as approaching, if not quite meeting, this standard, based on the maximum risk functions plotted in exhibits 4.4 and 4.5. Except for policy questions regarding differentiation of impacts in a limited (and possibly critical) band of values—$1,200–$2,200 per year during the in-program period; $200–$1,500 per year in the postprogram period—the no-show-based strategy carries almost no risk of misguiding policy decisions through selection bias. The same is true for the screen-out-based approach in the latter period. But the stated ranges where errors might arise are not trivial. And should sharp discrimination within these bands become essential to policy (e.g., in a benefit-cost analysis where the break-even point for earnings gains lies somewhere in these ranges), none of the nonexperimental methods tested here would have provided an acceptably reliable guide to policy. The experimental estimate, which carries zero risk of policy error due to selection bias across the entire range of possible impacts, would have done so.

We return to the question of the overall reliability of the nonexperimental methods when summarizing our results in chapter 5.

Interpreting Variations in Reliability Across Methods and Over Time

Taken as a whole, three features of our validation results are particularly striking and worthy of further discussion:

- The relatively poor performance of the withdrawal-based approach;
- The relatively strong performance of the no-show-based approach, particularly in relation to the more conceptually appealing screen-out-based approach during the two in-program years; and
- The improvement over time in the screen-out-based approach, moving it from little better than the withdrawal-based approach during the in-program period to little worse than the no-show-based approach during the postprogram period.

Our interpretation of these three key findings draws on the characteristics of the various applicant comparison groups discussed in chapter 2 and the derivation of the nonexperimental impact estimates in chapter 3.

From the outset, we have considered withdrawals as the group of nonparticipating applicants most likely to differ from participants and, therefore, as the group least likely to provide a reliable comparison group. This follows from the fact that withdrawals quickly withdrew their applications to the demonstrations (i.e., prior to a decision to admit or exclude them), and in that respect more closely resembled the AFDC recipients who never applied to the demonstrations than did other applicants. In contrast, both screen-outs and no-shows maintained their involvement in the program at least up to the point of selection by intake staff.

Also, in comparison to screen-outs, we expected the factors that distinguish withdrawals from participants—self-selection on motivation or availability of immediate employment—to be less likely to be controlled for by the selection variables included in our impact models—objective baseline characteristics and intake workers' ratings of suitability for training. The net result, poor relative performance in relation to the experimental benchmark, is apparent in exhibits 4.4 and 4.5.

In terms of the direction of bias, the top panel of exhibit 4.3 shows that in five of the six years after application, use of withdrawals as a comparison group produced impact estimates larger than the experimental estimates. This implies that withdrawals earned less than participants would have in the absence of the demonstration program, suggesting that the decision to drop out of the demonstration intake process had more to do with lack of interest than with alternative employment opportunities. Whatever the explanation, the earnings patterns seen here suggest that there are good reasons to avoid withdrawal-based comparison groups in future evaluations of employment and training programs similar to the AFDC Homemaker-Home Health Aide Demonstrations.

Like withdrawals, no-shows decided to withdraw from the program intake process before services began, though in this case only after they had been approved for services. Hence, we would expect their earnings to more closely approximate the no-program counterfactual earnings of participants. Moreover, the subjective ratings variable is available for virtually all no-shows, but for only a minority of the withdrawals.

We have no direct information on why the no-shows did not enter the program, however. Without this information, we cannot determine how similar or dissimilar their earnings capacities are to those of the controls and participants. On *a priori* grounds, the no-shows may have failed to participate because they had better employment opportunities than participants, or, conversely, they may have simply been less motivated or more pessimistic about their employment potential. These *a priori* reasons apply to the withdrawals as well, but, for reasons noted above, we expect the no-shows to be more similar to the participants than the withdrawals.

Throughout the follow-up period, the earnings of the no-shows are more similar than those of the withdrawals to the counterfactual earnings of participants, as evidenced both by the impact estimates shown in exhibit 4.3 and by the maximum risk functions in exhibits 4.4 and 4.5. No-shows earned more than participants would have during the first year after application (1985), since the no-show-based impact estimate in that year is smaller than the experimental estimate. This initial earnings advantage declined over the next two years, however, and in the last four years of the follow-up period no-shows earned less than participants would have in the absence of the program, resulting in no-show-based impact estimates that were consistently larger than experimental estimates. This pattern suggests that no-shows failed to participate in the program because of superior short-run employment opportunities, but that in the longer run they were likely to earn somewhat less than the participants would have without demonstration assistance.

A key question is why the screen-out approach did not provide a better substitute for the experimental control group during the in-program period, especially in comparison to the no-shows, once we adjusted for the objective and subjective factors that led to exclusion from the program. Returning first to the unadjusted difference-in-means estimates in exhibit 3.4 of chapter 3, we see that throughout the follow-up period screen-outs earned substantially less than controls, leading to a substantial overestimation of demonstration impacts when selected as a comparison group. This shortfall in the unadjusted data is not surprising, since demonstration intake staff consciously excluded applicants who looked less promising as homemaker-home health aides, a factor likely to imply less promise in the labor market generally. Surprisingly, however, the addition of baseline characteristics and intake workers ratings to the model (later columns of exhibit 3.4) did little to improve the model's performance during the in-program period (1985-86).

The particular regression specification used may account for why none of the nonexperimental estimates for 1985 accorded well with the experimental benchmark, as shown in exhibit 4.1. The estimates were derived by regressing earnings on an indicator of treatment status (participant versus comparison group) and a set of covariate measures of baseline characteristics; this is the standard specification for such estimates. If relative earnings within the participant group were determined by a different process during training and subsidized employment from the unsubsidized labor market, the coefficients on the baseline variables for 1985—which were determined largely by the much larger participant sample—may not work well for any of the comparison groups. If this be the case, "equalizing" the two samples by adjusting them to the common set of coefficients may seriously distort the comparison group level and, hence, the nonexperimental impact estimates. It seems likely that the in-program earnings of participants were determined by a different process from the earnings of nonparticipants during the same time period, since participants' earnings were heavily subsidized by the demonstrations. The same hypothesis may also explain the relatively poor performance of the withdrawal and screen-out impact estimates for 1986 (a year that fell partially within the in-program period for some participants), but should not be a factor in 1987 and beyond.

As can also be seen from exhibit 3.4, controlling for standard baseline characteristics and, subsequently, for ratings of suitability as homemaker-home health aides substantially increased the reliability of the screen-out comparison group in the postprogram period, as predictèd in chapter 2. Here, it cut the degree of overestimation in half. Even so, the fully adjusted screen-out-based impact estimates remained substantially above the experimental benchmark throughout the postprogram period, producing a postprogram maximum risk function that rises substantially above zero over a fairly wide range of policy issues. (See exhibit 4.5.)

The failure of the intake workers' ratings to fully capture the effects of the program selection process may be due to two factors:

- The ratings of suitability may not have been applied consistently enough to adequately account for the effects of selective intake on future earnings; or
- Intake staff may have selected among applicants on the basis of factors other than those they used to assign suitability ratings, but factors also related to future earnings.

Both of these possibilities are supported by the pattern of suitability ratings observed among participants and screen-outs. All four ratings levels—excellent, good, fair, and poor—are found in each group, although higher ratings are much more common in the participant sample and lower ratings in the screen-out sample. Even so, for a nontrivial number of cases (the "poors" admitted to the demonstration and the "goods" and "excellents" excluded), intake staff either misapplied the rating scale or consciously based their screen-out decisions on something other than their assessment of future potential as homemaker-home health aides.

While we cannot distinguish between these two hypotheses on the basis of the available data, they both carry similar implications:

- The full potential of screen-outs as a comparison group has not yet been realized in the homemaker data, in terms of the reliability and sophistication of the suitability ratings used; and
- If better measures of suitability or potential can be developed—e.g., a many-valued scale with explicit guidelines for how it should be applied and related to intake decisions—the screen-out approach might provide a more reliable substitute for experiments than in this application.[29]

On the basis of the evidence presented here, none of the applicant groups yielded estimates close enough to the experimental benchmark to justify the claim that it provides an adequate substitute for an experimental control group. For example, even the best applicant-based estimates differed from the experimental norm by more than 50 percent in the critical postprogram period. Nevertheless, there are several reasons for believing that the screen-out and no-show groups could potentially provide a nonexperimental method for evaluating training programs that yields reliable and unbiased impact estimates.

First, both of these applicant groups gave estimates of the program's effect on earnings that were relatively close in absolute terms to those provided by the control group in the postprogram period. Second, the theoretical basis for using screen-outs—that this group allows an unbiased estimate if the selection process is modeled and controlled for—was supported by the finding that methods for measuring selection are feasible and promising. In particular, demonstrations showed that the difficult problem of capturing the subjective determinants of program selection can be addressed, and that controlling for this part of the selection process sharply reduces the selection bias of the screen-outs compared to estimates based on the experimental control group. Finally, with either the screen-outs or no-shows, serious risk of erroneous policy decisions based on the postprogram estimates (risk above, say, 40 percent) was confined to a relatively narrow range of policy thresholds (a band of about $500 in width).

We conclude that further tests of the methodology should be undertaken, using other experimental data sets. Given the relatively strong performance of screen-outs and no-shows in this analysis, we recommend that attention be paid to measuring both the factors underlying selection by program staff and those that lead some applicants who are accepted into the program not to participate. Our results cannot, of course, be considered conclusive, given the limitations of the available data and test procedures and the fact that our results reflect the specific circumstances of a single demonstration project.

NOTES

1. To our knowledge, no evaluator has developed a test methodology that can incorporate the uncertainties involved in measuring program costs (and in projecting future earnings gains beyond the observation period) with the more readily incorporated uncertainties of measuring earnings gains during the observation period to test the statistical significance of "net program benefits" (which equal observed and projected program benefits less program costs).

2. See, for example, Orr et al. (1994), Gueron and Pauly (1991), Orr (1987), Long, Thornton, and Whitebread (1983), and Kemper, Long, and Thornton (1981).

3. See footnote 16 of the Friedlander and Robins paper. As explained there, all types of impact estimates—including the experimental estimate—are less likely to differ significantly from zero in small samples, due to the larger standard error that attaches to small sample estimates. Thus, as sample sizes decrease, experimental and nonexperimental estimates become more and more likely to agree in terms of statistical inference (with both indicating no significant effect), irrespective of the true extent of selection bias.

4. Near the beginning of his analysis, LaLonde states the principle that: "If the econometric model is specified correctly, the nonexperimental estimates should be the same (within sampling error) as the training effects generated from the experimental data, but if there is a significant difference between the experimental and nonexperimental estimates, the econometric model is misspecified" (LaLonde 1986, pp. 610-11). He then applies this criterion to just a few of his numerical estimates: in a series of comparisons between experimental and nonexperimental estimates, the distance between estimates is discussed in relation to their standard errors—the standard measure of sampling error—only twice (pp. 614 and 616), and then only loosely. Fraker and Maynard also refer to "tests of the comparability of the earnings models of the comparison and control samples" (pp. 205 and 212), which could have a similar purpose but are not explained.

5. A second, related problem is the potential for very large samples to reject a nonexperimental estimate because it differs from the experimental finding in *statistical* terms but not enough in *substantive* terms to really matter. Friedlander and Robins' first validation criterion—the magnitude of observed differences—presumably provides some protection against this risk by identifying the statistically significant differences that are of no substantive consequence.

6. In the homemaker demonstration data—as in almost any employment and training program evaluation—limited samples imply a substantial potential for approximation error in both nonexperimental and experimental impact estimates. This potential is reflected in the standard errors of the estimates, which are shown for the experiment in parentheses alongside the experimental impact measures in chapter 3. (Standard errors for the nonexperimental impact measures appear later in this chapter.) The homemaker demonstration data used here provide substantial but not unusually large samples: 1,600 participants and 300 to 900 comparison group members, depending on the comparison group chosen. These samples are somewhat smaller than those available to Friedlander and Robins and roughly equal in size to those used in the earlier validations studies of the National Supported Work Demonstration (except for two comparison groups of 11,000 and 16,000 members, respectively, used by LaLonde 1986). Several more recent employment and training experiments provide substantially larger participant samples, as well as the potential to be

matched to even larger external comparison groups. However, we are aware of only one instance outside of the AFDC Homemaker-Home Health Aide Demonstrations where an internal comparison group of nonparticipating applicants of any size is available: the Corpus Christi site in the National JTPA Study, where data have been collected for 600 participants and 2,600 nonparticipating applicants.

7. Indeed, a null hypothesis this broad almost certainly is true. As DeGroot (1975, p.406) and others have pointed out, the hypothesis that bias (or any other quantity) equals precisely $0 (or any other specific value) is almost certainly false, and hence our particular null hypothesis almost certainly true independent of the data.

8. Note that changing the null hypothesis in this fashion—essentially for statistical reasons—also addresses an important substantive concern in hypothesis testing: the need to ensure that any bias detected by statistical means is also of a magnitude that gives it practical importance. Normally, the criterion of practical significance is applied after the statistical test, by asking whether a statistically significant bias is also large enough to be of concern. Here, we combine both considerations into a single step by looking for conclusive statistical proof of bias at or above some threshold level.

9. DeGroot (1975, pp. 406-407) provides the basis for another approach, in a test of the null hypothesis that a quantity falls within a certain range, against the alternative hypothesis that it is outside that range. A similar test could be constructed with the two hypotheses reversed.

10. This assumes a hypothetical difference between the nonexperimental and experimental estimates of $50 and a hypothetical standard error for that difference of $76.

11. Whether policy decisions do, in fact, emerge from such an exercise is not clear. This formulation of the issue remains appropriate in any case since, if evaluation results are to affect policy, they presumably will do so in this fashion.

12. A diffuse prior attaches equal probability to all points on the real number line.

13. For this result and others cited below, see page 191 of DeGroot (1970) or any of a number of standard textbooks on Bayesian statistics and/or statistical decision theory.

14. In fact, we cannot reasonably anticipate the sample sizes that will be available in future impact studies, nor the tradeoffs in sample sizes to be faced when choosing between experimental and nonexperimental methods. Thus, the only way to incorporate statistical variability in our analysis is to assume that it is neutral across the two methods.

15. See below for a discussion of this "external validity" issue.

16. In general, risk functions in statistical decision theory depend not only on the probability of certain types of mistakes but on the negative consequences, or "loss," associated with each particular mistake. Here, we adopt a simplified framework in which the loss incurred in making policy errors is the same for all possible errors. A more realistic approach, beyond the scope of this monograph, would attach larger loses to larger mistakes, so that favorable (unfavorable) policy decisions for true impacts way below (above) C^* would add more to the risk function than for true impacts only slightly below (above) C^*. It would also be possible to attach greater losses to mistakenly favorable policy decisions than mistakenly unfavorable policy decisions, or vice versa, if one type of policy decision is believed to be more consequential than the other.

17. Zero correlation implies independence in a joint normal distribution. We have assumed normality in the data from the outset, which—with a diffuse prior—implies normality in the posterior distribution, which is where independence is required.

18. Unfortunately, as also shown in appendix D, it is not the case that the special case of perfect (positive) correlation provides a lower bound on the probability of an incorrect policy decision. The only known lower bound is 0.

19. A uniform prior for the correlation coefficient over the [0,1] interval might be considered in this role but would still embody a specific set of beliefs about the reliability of the nonexperimental estimator in relation to truth that could be subject to dispute.

20. When C equals the midpoint of both the I and N distributions, $Pr\ (I < C) = Pr\ (I > C) = Pr\ (N < C) = Pr\ (N > C) = .5$, and $R(C) = (.5) \bullet (.5) + (.5) \bullet (.5) = .5$.

21. None of the papers reviewed earlier emphasize this point. However, Fraker and Maynard (1987) do consider whether the Supported Work setting is likely to be characteristic of other large-scale employment and training programs that might be evaluated using nonexperimental methods similar to those they test and reject.

22. LaLonde (1986) also pointed to the importance of validation work in other settings. With the exception of the Friedlander and Robins (1992) paper and the current monograph, that emphasis has thus far produced only additional analyses of the same program setting considered by LaLonde (the National Supported Work Demonstration). A further extension of the validation literature into new experimental data sets is now possible, based on the many such employment and training data sets produced in recent years. See Greenberg and Shroder (1991) for examples.

23. While not always evident in their conclusions, the same is true of all of the earlier model validation exercises.

24. The approach could, for example, be tested in one of the 16 sites evaluated experimentally as part of the National JTPA Study noted earlier—the one site (Corpus Christi) where baseline and follow-up earnings data are available for a large number of nonparticipating applicants. Unfortunately, none of the other sites in that study provided data on a sufficient number of nonparticipating applicants during the experimental intake period to conduct such a test. Nor are intake workers' ratings of suitability available even in Corpus Christi.

25. Separate regressions were run on average earnings over the multiyear periods, to obtain appropriate standard errors for the multiyear results.

26. Experimental impact estimates can produce incorrect policy decisions due to sampling error, of course, but we are abstracting from that factor (for both the experimental and nonexperimental estimators) in this exercise.

27. Like experimental estimators, they carry some risk of misguiding policy due to sampling and measurement error in the data, which we symmetrically ignore here.

28. In deriving this and other cutoff values noted below, we consider risks of .05 or above to be serious enough to merit mention.

29. Fraker and Maynard (1987) also recommend the collection of better screening data at the conclusion of their article.

5
Summary and Recommendations

In reviewing the literature on nonexperimental estimation of training program impacts in chapter 1, we concluded that there is no generally accepted nonexperimental method available to deal with the problem of selection bias, despite nearly thirty years of efforts to find one. We then returned to an approach tried briefly and abandoned early in the process: the use of nonparticipating applicants as comparison group members.

As demonstrated in chapter 2, there are important *a priori* reasons why applicant-based comparison groups might provide more reliable measures of the impacts of employment and training programs than comparison groups drawn from sources external to the program. In chapters 3 and 4, we tested estimates based on several such "internal" comparison groups against those derived from a true experiment, using a unique data set from the AFDC Homemaker-Home Health Aide Demonstrations. In this chapter, we summarize the *a priori* rationale for the applicant-based approach and the results of our empirical tests. We then close by noting the implications of our work for future research and evaluation practice.

The Rationale for Applicant-Based Comparison Groups

The fundamental challenge in evaluating employment and training programs is to obtain an unbiased estimate of what program participants would have earned absent the program. Past attempts to do so with nonexperimental comparison groups have foundered on the problem of selection bias: the very processes that lead some individuals to participate in an employment and training program also lead to higher or lower earnings even without the program's services.

Past Responses to Selection Bias and the "Preprogram Dip"

The source of this problem was first recognized in connection with early efforts to measure program impacts through pre/post earnings comparisons for participants alone: the "preprogram dip." This dip results from participants entering employment and training programs at a time when earnings are transitorily low. In contrast to these participants, individuals drawn from the general population are, on average, likely to be in steady state in the labor market. Therefore, matching external comparison group members to participants on earnings in the immediate preprogram period is likely to overstate program effects, as participants rebound from their transitorily low earnings levels and comparison group earnings remain relatively stable.[1]

Attempts to control statistically for preexisting differences between participants and comparison group members have generally relied on relatively fixed individual characteristics like age, race, sex, and educational attainment. Because these characteristics are invariant over time, they cannot account for transitory differences between participants and comparison group members.

Panel data became more generally available in the 1970s; this gave analysts a better method of matching the earnings histories of participants and comparison group members based on long-term permanent earnings up to the time of program entry. The problem with this approach is that it assumes participants would regain their previous earnings levels in the absence of the program. There is no guarantee that this is the case. It is quite possible that the earnings loss that triggered program entry represented a permanent break in earnings trends for many individuals. In that case, one cannot project future earnings on the basis of preprogram earnings. In his review of the many studies of the Comprehensive Employment and Training Act employing this technique, Barnow (1987) found that impact estimates were quite sensitive to the specific treatment of the preprogram dip. Later attempts to resolve the problem using preprogram specification tests have also met with mixed success.

Applicant Comparison Groups as a Possible Solution

In chapter 2, we proposed using nonparticipating applicants as comparison groups in order to minimize preexisting differences between participants and comparison group members. Because they apply for services at the same time as participants, nonparticipating applicants can be expected to undergo the same type—though not necessarily the same degree—of dip and recovery (or lack of recovery) in earnings as do participants. Hence, even if the preprogram dip signifies a permanent break in the earnings trend of participants, some version of that break may be evident in the earnings paths of nonparticipating applicants. Alternatively, if the break is transitory for some of the participants, it should also be transitory for some of the nonparticipating applicants. Moreover, by applying to the program, nonparticipating applicants reveal themselves to have some of the same, sometimes difficult-to-measure personal characteristics (e.g., motivation, problem-solving ability) that lead participants to seek help in response to their current economic situation, which often reflected an earnings loss.

The use of applicant-based comparison groups can thus be expected to control for many of the individual characteristics and circumstances (including transitory ones) that lead individuals to apply to training programs. The differences that remain will depend on the reasons comparison group members do not enter the program while other applicants do. For example, some applicants voluntarily drop out of the intake flow before reaching the point of program entry. These groups, which we call withdrawals and no-shows, systematically self-select out of the program. As a result, they seem likely to differ from participants in observable and unobservable attributes (e.g., motivation or alternative employment opportunities) that could affect their future earnings.

Of the two groups, no-shows more closely resemble participants in at least two respects: they remain interested in the program further into the intake process, and they pass the screens for acceptance into the program. While this does not guarantee that they will more closely approximate what the later experience of participants would have been absent the program, it does at least make it more likely to the extent that these same self-selection factors affect future earnings.

In contrast, applicants screened out by intake staff differ from participants primarily in those attributes that determine program selection rather than self-selection. As with any systematically selected group, screen-outs will provide a biased representation of the without-program earnings of participants unless we control for these differences.

Fortunately, the factors leading to program selection—even those not generally measured in evaluation data sets—are more amenable to statistical control than those leading to self-selection. In particular, the factors that cause some program applicants to be rejected while others are accepted, if not totally random, are by definition externally observable at least in the perceptions of the program intake staff who make admission decisions. If these factors can be measured—through intake workers' ratings of applicant potential and more conventional demographic and background variables—they can be controlled for in the analysis to remove the selection bias brought about by the program's intake procedures. By contrast, the use of external comparison groups allows for very little control for program selection and, as noted, virtually no control for self-selection. There is, of course, no guarantee that the use of internal comparison groups and attention to the selection process will eliminate selection bias, but the chances of attaining this ideal should be increased with this approach.

In general, then, comparison groups composed of nonparticipating applicants may be superior to those drawn from the general population, even when the latter are matched to the participants on individual characteristics and prior earnings. Within the population of nonparticipating applicants, there are *a priori* reasons to expect that no-shows will provide better estimates than withdrawals, and that screen-outs will provide the best comparison if we are able to control for the objective and subjective factors intake staff use to select program participants.

Empirical Estimates Using Alternative Applicant-Based Comparison Groups

We tested these theoretical propositions in chapters 3 and 4 by estimating program impacts using applicant-based comparison groups and comparing the results with estimates derived from a controlled experiment. The data for this analysis come from the AFDC Homemaker-Home Health Aide Demonstrations, a voluntary program that provided four to six weeks of training and up to a year of subsidized employment for AFDC recipients in seven states.

The homemaker-home health aide demonstrations provide data that are unusually well-suited to this testing exercise in several respects. First, the sample includes not only participants and a randomly assigned control group, but also a group of applicants who did not participate in the program. Second, as part of the intake process, program staff rated applicants on their potential as homemaker-home health aides, allowing us to extend conventional comparison group adjustments to subjective factors that affect program selection. Third, earnings data are available for both the experimental sample and the nonparticipating applicants for an unusually long follow-up period—up to five years after exit from the program.

Using these data, we estimated three different impact models for each of three applicant-based comparison groups—withdrawals, screen-outs, and no-shows—in each of seven years. The first set of estimates were simply unadjusted differences in mean earnings between participants and the comparison group. The second set of estimates were regression-adjusted to control for differences in observable baseline characteristics, such as demographics and prior earnings. The third set of estimates controlled for both these observable variables and the unobservable screening criteria captured by intake workers' subjective ratings of applicants. Thus, the first model made no attempt to control for selection effects, beyond the use of nonparticipating applicants as a comparison group, while the second used a set of standard nonexperimental controls for self-selection. The third model attempted to control both for self-selection and, through the inclusion of the subjective ratings, for program selection. Each set of estimates was compared against the experimental benchmark.

In general, we found that unadjusted mean differences between participants and comparison group members greatly overstated program effects, especially in the postprogram years, revealing that participant earnings would have exceeded the earnings of other applicants even absent the intervention. The addition of conventional baseline covariates to the model narrowed the gap between the nonexperimental estimates and the experimental estimates for screen-outs and no-shows, although this adjustment actually moved the withdrawal-based estimates further off the mark. Adding the intake workers' subjective ratings narrowed the gap between the nonexperimental and experimental estimates in the post-program years by $150 to $260 for withdrawals and $20 to $160 for screen-outs; it had little effect on the estimates for no-shows.

The fully adjusted nonexperimental estimates for the withdrawal-based sample overestimated program impacts in all years but one, and by a rapidly growing margin (up to $1,600) toward the end of the follow-up period. However, those based on screen-outs and no-shows tended to overstate program effects in the postprogram period by a much smaller and more stable margin, ranging from $120 to $430. On the basis of the impact estimates alone, then, withdrawals do not appear to provide an acceptable comparison group, especially for long-run impacts; screen-outs and no-shows yield much more promising, though still suspect, estimates.

Specifying Validation Tests

Previous comparisons of nonexperimental impact estimates with experimental estimates for the same program have often stopped at this point. Others have focused on whether experimental and nonexperimental estimates yield the same policy implications (i.e., are significantly different from zero in the same direction). These criteria—while emphasizing the right factors—provide an inadequate basis for assessing the performance of nonexperimental estimators, for several reasons.

Problems with Previous Methods

The dividing line between differences that are acceptably small and those that are not, in terms of magnitude alone, is inherently judgmental and arbitrary. Moreover, the comparison of experimental and nonexperimental point estimates does not take account of the fact that each of the estimates is only a single draw from the distribution of estimates that would be yielded by the two approaches in repeated applications. Once the random character of that draw is taken into account, the difference between two estimates only approximates the selection bias in the nonexperimental approach.

This problem can be at least partially addressed by testing whether the difference between the two estimates is statistically significant. The stringency of this test depends, however, on the sample size involved: in small samples, even differences that would be generally regarded as large might not be statistically significant; in very large samples, even inconsequential differences may be statistically significant. And as explained in appendix C, in evaluating voluntary training programs the existence of no-shows makes the test infeasible except through a complex adjustment for the correlation between the two estimates.

Similarly, while focused on an important criterion, past studies that compared nonexperimental and experimental estimates in terms of their policy implications have defined the issue too narrowly. Simply asking whether the estimated impact on earnings was significantly different from zero in the same direction under the two approaches does not address alternative policy questions such as whether program benefits exceeded program costs. Since policy makers are likely to ask a wide range of policy questions in future evaluations, a more robust validation criterion would be highly desirable.

Finally, those previous validation exercises that were based on tests of statistical significance have implicitly presumed each nonexperimental technique to be unbiased unless the data prove otherwise. Given that concerns over selection bias have dominated all past efforts to measure the effects of employment and training programs nonexperimentally, the burden of proof clearly belongs on the other side of the question: nonexperimental methods should be presumed to suffer from selection bias to an unacceptable degree until data prove otherwise.

A More Complete Validation Standard

In chapter 4, we respond to these concerns by proposing a more balanced, comprehensive, and rigorous method for assessing the potential selection bias in any given nonexperimental estimate. We derive a technique that both allows for more meaningful comparisons among such estimates and provides a rigorous basis for judging each measure against the experimental norm. This technique is also more sensitive than past approaches, in that it provides a quantitative measure of how well each nonexperimental estimator performs across a range of policy questions, rather than a global accept/reject decision for only a single policy question.

Our approach begins by recognizing that nonexperimental estimators must be presumed biased until proven otherwise. Unfortunately, proving a nonexperimental approach to be unbiased on the basis of a single estimate drawn from a finite sample is virtually impossible, due to sampling error. As an alternative, we consider whether the approach would yield unbiased policy conclusions, on average (i.e., abstracting from sampling variability), over a range of policy questions. We do this by applying a Bayesian decision theory framework to the results of the experiment to formulate a probability distribution for true program impact and a similar distribution for the average result of a given nonexperimental method. These two distributions are then combined to calculate an upper bound on the odds that the nonexperimental approach will produce a misguided policy conclusion due to systematic selection bias, when the critical policy issue is whether true impact exceeds some threshold level, *C*.

We are then able to map the limits of the probability of a misguided policy decision over all possible policy questions (i.e., all possible *C* values) to arrive at a "maximum risk function" for that particular nonexperimental approach. Abstracting from sampling variation, an experiment produces the correct policy inference regardless of the policy question; therefore, its "maximum risk function" is 0 for all values of *C*. Individual nonexperimental approaches can then be judged acceptable or unacceptable as alternatives to the experiment based on how much, and over which range of policy decisions, their maximum risk exceeds 0. Assessment of the relative performance of alternative nonexperimental estimators can also be made to determine which is the most trustworthy as a guide to policy.

Assessment of Model Performance

The maximum risk function criterion is then applied to the nonexperimental impact estimates derived from comparison samples of withdrawals, screen-outs, and no-shows in chapter 3. Two separate analyses of the homemaker data are conducted, one for the mostly in-program years of 1985-86 and the other for the postprogram period of 1987-90.

The in-program analysis shows an unacceptably high potential risk of incorrect policy conclusions due to selection bias for both the withdrawal-based and screen-out-based impact estimates, over an important range of policy decisions (those that hinge on whether annual impacts exceed or fall short of cutoffs in the $1,700-$3,300 range). By matching much more closely the experimental findings in this interval, the no-show-based approach shows much less risk of misguided policy, and carries a substantial risk only for decisions that hinge on impacts exceeding or falling short of policy thresholds in the range of $1,200-$2,200 per year.

During the postprogram period, both the no-show- and the screen-out-based estimates are judged to be minimally acceptable alternatives to the experiment in this application. Here, if key policy decisions hinge on impacts below $200 per year or over $1,500 a year (and if sampling error is minimized through large sample sizes), both approaches would be expected to provide a highly reliable guide to policy concerning the homemaker intervention. The lower of these two ranges includes questions of whether the homemaker demonstrations had any earnings impact at all in the postprogram period (*C* = $0), and the higher begins just above where impacts become large enough and sustained enough to offset initial program costs (*C* around $1,200 per year).[2] Only if policy makers need to distinguish between annual impacts in the $200-$1,500 range would the risk of incorrect policy conclusions through selection bias preclude the use of comparison groups of no-shows and/or screen-outs. Withdrawal-based estimates were also found to guide policy reliably regarding impacts below $200 per year, but not regarding impacts in the $200-$2,200 range.

Conclusions and Recommendations

While the results in chapters 3 and 4 fall short of identifying a fully acceptable substitute for the experimental approach, we believe that the evidence presented here is generally encouraging with regard to the use of applicant-based impact methods when experiments cannot be implemented. As shown in chapter 4, two of the three families of applicant-based findings from the homemaker demonstrations proved to be reliable guides to policy across a fairly wide range of policy questions in the postprogram period, at least as concerns the core problem of selection bias that experiments are designed to remove. One of these two families—the approach using training no-shows as a comparison group—also proved generally reliable during the in-program period over a range of policy questions. The theoretically preferred screen-out-based approach proved much less reliable than the no-show-based model during the in-program period, but yielded similar results in the postprogram period.

Longer-Term Follow-Up

The improved reliability of the screen-out-based estimates during the postprogram period has a plausible explanation related to the need for longer-term follow-up of the sample. Presumably, program staff based their screening decisions on the most readily detectable signals of future success. In general, indicators of short-run success should have been more visible to staff than precursors of long-run success. If screened on this basis, we would expect any initial differences between screen-outs and participants (after controlling for measured differences between the two groups) to diminish over time. If we were able to follow the sample for a period as long as, say, ten to fifteen years, screen-outs might be virtually equivalent to a randomly selected control group. We would not predict this result for applicants kept out of the program through self-selection (withdrawals and no-shows), since self-selection could be based as much on permanent and unmeasurable individual differences as on short-run differences in observable characteristics. Further follow-up of the homemaker sample would provide a test of these alternative hypotheses.

Over the first six years after random assignment, no-shows performed best among the three comparison groups tested, particularly in the first three or four years following program entry. In the postprogram years, however, which are crucial for measuring how the training program affected the earnings capacities of the trainees in an unsubsidized market and over the long run, there was little difference in the performance of the screen-out and no-show comparison groups. As between these two groups, the theoretical rationale for using screen-outs is more credible. Moreover, the process that generates screen-outs is more amenable to better measurement and modeling than was achieved here (see below). Thus, screen-outs may well provide the comparison group for future nonexperimental evaluations.

Strengthening the Ratings Variable

One of the innovations tested in our analysis is the use of intake workers' subjective ratings to control for the factors influencing program selection among applicants. The addition of these ratings consistently moved the withdrawal- and screen-out-based estimates (though not the no-show-based estimates) closer to the experimental norm. This is particularly encouraging given the relatively crude nature of the rating variable available (a simple four-point scale), the high rate of missing data for this measure in the nonparticipant samples, and the indications in the data that intake workers either applied the ratings inconsistently or made their screening decisions on criteria other than the ratings. We believe that these results justify further research to more fully develop the rating approach to modeling program selection.

In any future application of this approach, we would recommend that the rating scale be improved in several ways. First, more fine-grained ratings—e.g., a ten-point scale—should be used, to break the confounding of scale values with program status that tends to occur when a scale with a small number of discrete values is used.[3] Intake workers might be motivated to use a wider range of values of the scale by instructions that emphasize that the cutoff value for determining selection is not fixed in advance and may vary over time depending on the supply of applicants. Second, care should be taken to see that intake workers provide ratings for as many applicants as possible, to minimize missing data. This can be done by emphasizing the importance of these data when training workers on data collection procedures, and by recording the ratings early in the intake process before large numbers of applicants have dropped out. In training intake workers, it should also be emphasized that the ratings are intended to reflect the criteria used to decide which applicants to accept, however subjective those criteria may be. Finally, attention should be given to maximizing the consistency with which the ratings are applied. This might be done by having all the ratings assigned by the same worker or by a committee of intake workers. At a minimum, it is important to record the identity of the worker who assigned each rating, so that *ex post* adjustments can be made for systematic differences in ratings among workers.

Extensions to Other Data Sets

However one views the results of the present study, it is important to bear in mind that our empirical analysis constitutes only one example of the use of nonparticipating applicants as a comparison group. A single test, conditioned on the specific circumstances of a single set of demonstrations, cannot be viewed as conclusive evidence either for or against any particular method. Many more replications of this approach, using different samples drawn from different programs and environments, will be required to arrive at any definitive conclusions. We hope that this study will encourage others to conduct further tests along these lines, and to revisit the results of earlier validation studies to judge their reliability in light of the new validation methodology presented here.

Regardless of the empirical evidence available at the moment, we hope that the *a priori* arguments we offer for screen-out-based impact analysis will be taken seriously by the evaluation community. As argued in chapter 2, the use of screen-outs as comparison group members is highly feasible in a variety of evaluation contexts and, though not entirely successful in its first application, appears on theoretical grounds to offer the best hope available (outside of an experiment) for resolving the selection bias problem. No-show-based comparison samples also hold promise, on the basis of our empirical results. If other studies show similar results, closer examination of the reasons for nonparticipation among this group should be undertaken in an attempt to model their selection behavior.

Our strongest recommendations are, first, that researchers conducting experimental evaluations of employment and training programs begin to routinely collect subjective ratings from intake workers at baseline, in order to document the selection process. Second, long-term follow-up data should be collected for applicants who do not participate in the program. This information can then serve as the basis for further tests of the applicant-based approach analyzed in this monograph, when experimental data are available. And when experiments are not possible, the same set of inputs can provide a conceptually appealing and empirically promising alternative to conventional nonexperimental methods for measuring the effects of government training programs.

NOTES

1. Throughout this discussion, the term "matching" is intended to encompass both sample selection designed to match comparison group members to participants and the use of statistical methods to control for differences between the two groups.

2. Bell and Orr (1994) project positive net benefits for all five of the homemaker demonstrations with monthly earnings impacts of $100 or more ($1,200 annually) during the second postprogram year.

3. In the extreme, a two-point scale would be useless, because its values would simply correspond to the (observable) decision to accept or reject the applicant. If intake workers use a four-point scale, but rate nearly all applicants in the middle two categories of the scale, the resulting ratings are only slightly more informative than those based on a two-point scale.

Appendix A
Construction of Grouped Data

This appendix describes the construction of the grouped data used in the analysis. The majority of variables are taken from the Basic Information Sheet filled out by all applicants to the AFDC Homemaker-Home Health Aide Demonstrations. These individual data are combined with grouped annual earnings data from the Internal Revenue Service to produce the analysis file used in the monograph.

The Original Evaluation Data

Between 1982 and 1987, Abt Associates collected experimental data from the AFDC Homemaker-Home Health Aide Demonstrations, seven state-run programs that provided training and subsidized employment experience as homemakers and home health aides to selected AFDC recipients. During demonstration intake, prospective applicants filled out Basic Information Sheets providing demographic and educational background, employment experience, caregiving experience, and public program participation. We include only individuals who went through intake between July 1984 and May 1985, as this is the single group for which complete data are available on all demonstration applicants. We have individual-level data on 5,521 program applicants including participants, withdrawals, screen-outs, no-shows, and control group members.

IRS Earnings Data

Abt's original evaluation produced two years of follow-up data on the original experimental sample—participants, no-shows, and controls—for the years 1984 and 1985. In order to include withdrawals and screen-outs in the sample and extend the follow-up period, demonstration data were matched to Internal Revenue Service earnings data (wages plus tips) for 1984 through 1990. To ensure anonymity, the IRS provides means and standard deviations of earnings for groups of 10 to 19 individuals. The nature of the earnings data therefore requires that all subsequent analysis be conducted at the group level.

Formation of Groups

To maximize the efficiency of estimates obtained from the grouped data, we formed groups that were as small and as homogeneous as possible with regard to crucial baseline variables. This procedure minimized within-group variance and preserved as much cross-group variation as possible in grouped analysis.

Individual observations were stratified and grouped according to eight baseline characteristics, including program status, state of residence, race, education, age, number of children, intake workers' ratings of applicants' potential for succeeding in the program, and wage rate on the last job held before intake. Exhibit A.1 shows the variables used to form groups of observations for the experimental and nonexperimental components of the sample, in descending order of priority.[1] The bottom panel summarizes the degree of homogeneity in each set of groups.

Stratification produced 499 groups with an average group size of 11.1 individuals. (See exhibit 3.1 for the average group size of each population.) All experimental groups are completely homogeneous with regard to program status (participant, no-show, control) and state of residence. Due to their smaller number, the nonexperimental groups are homogenous only with regard to program status (withdrawal, screen-out). On average, groups reached the third level of the stratification before experiencing any within-group heterogeneity.

Covariate Means and Missing Data

Exhibit A.2 shows group-level means for all covariates used in the analysis. Most of these measures indicate the average percent of a group in a given category, including a category for missing data. As can be seen, missing data are relatively infrequent. The overall average missing data rate is 1 percent for age, 4 percent for race and number of children, and 12 percent for previous wage.

Group averages for continuous variables such as age exclude individuals with missing or invalid data, except for previous wage and 1984 annual earnings, where those with unknown values are assumed to have earned the average reported value for their applicant group, rating of potential, and race. Categorical variables include missing observations as a distinct category with the exception of race, where missing observations are included in the "other race" category.

NOTE

[1] As shown in exhibit A.1, the grouping of the experimental sample was structured in part to parallel sample divisions used in the original evaluation, resulting in a somewhat different priority order from that of the nonexperimental sample.

Exhibit A.1 Grouping Variables, in Priority Order, and Resulting Degree of Within-Group Homogeneity

Experimental Sample	Nonexperimental Sample
1. Program status (participant, no-show, control)	1. Program status (withdrawal, screen-out)
2. State of residence (states with high impacts in original evaluation, states with low impacts)[1]	2. Rating of potential
3. Rating of potential (4 categories, plus "missing")	3. Race
4. Race (White, Black, Hispanic, other, "missing")	4. Wage rate on last job[2]
5. Wage rate on last job[2]	5. Education
6. Education (H.S. grad, dropout, "missing")	6. Age
7. Age (15-24, 25-34, 35+, "missing")	7. State of residence (7 categories)
8. Number of children (1, 2, 3+, "missing")	8. Number of children
Within-group homogeneity	
All groups homogeneous to Level 2	All groups homogenous to Level 1
Average group homogenous to Level 3	Average group homogenous to Level 3
Maximum degree of homogeneity = Level 8	Maximum degree of homogeneity = Level 7

[1] High-impact states include Arkansas, Kentucky, New Jersey, Ohio, and Texas. Low-impact states are New York and South Carolina.

[2] A two-stage stratification was performed on wage rate. First, the sample was divided between those who had worked previously and those who had not. The latter group was stratified on the remaining variables on the list (education, age, etc.). Those who had worked were ordered by wage rate (with missing wage rates set to the mean for their strata) and divided into groups of 10 consecutive observations to maximize within-group homogeneity on this factor. No further stratification of these cases was possible without disturbing this ordering.

Exhibit A.2 Mean Values of Covariates, by Applicant Group

	Withdrawals[a]	Screen-outs	No-shows	Participants	Controls
Sample Size	82	83	24	144	166
Covariate					
Average age[b]	30	31	29	30	29
% White	30	38	17	28	28
% Black	53	51	58	57	55
% Hispanic	11	7	17	8	10
% Other race[c]	5	4	8	7	7
% Educ. < 12 yrs.	52	50	50	44	46
% Educ. = 12 yrs.	32	30	29	35	33
% Educ. > 12 yrs.	14	20	15	17	15
% Never married	34	32	29	33	33
% Married, spouse present	11	10	4	7	7
% Separated, divorced, widowed	39	44	34	42	42
% Marital status missing	16	14	32	18	18
Average number of dependent children[d]	2.0	2.0	1.9	1.9	1.9
% Ever worked for pay	82	85	86	87	86
% Never worked for pay	18	15	14	13	14
Average maximum previous hourly wage[e]	\$3.00	\$3.17	\$3.37	\$3.33	\$3.33
% Maximum wage ≤ \$25	93	97	100	100	100
% Maximum wage > \$25	7	3	0	0	0
% Educ. missing	1	0	6	5	5
Average 1984 annual earnings[f]	\$823	\$592	\$710	\$840	\$749
% Arkansas	4	11	15	15	15
% Kentucky	17	31	3	13	11
% New Jersey	31	15	14	17	17
% New York	15	14	26	13	13
% Ohio	26	20	4	14	13

Exhibit A.2 (continued)

	Withdrawals[a]	Screen-outs	No-shows	Participants	Controls
% South Carolina	4	4	9	8	9
% Texas	3	4	23	15	17
% State missing	0	0	6	4	5
% Rated potential "excellent"	1	4	13	12	12
% Rated "good"	17	24	56	64	63
% Rated "fair"	12	27	12	12	12
% Rated "poor"	4	9	1	0	0
% Intake rating missing	65	36	18	12	12

a. Of the 82 withdrawal groups, 19 consist of individuals for whom reason for nonparticipation is missing. These individuals are assumed to have withdrawn from the intake process voluntarily without providing a reason to the demonstration staff.

b. Observations with missing or invalid values are excluded from the calculation: 2 percent of the withdrawals, 4 percent of the screen-outs, and 0 percent of the remaining groups.

c. "Other race" includes observations with missing data, from a low of 2 percent of the withdrawals to a high of 7 percent for no-shows. The overall average missing data rate is 4 percent.

d. Observations with missing or invalid values are excluded from the calculations, from a low of 1 percent of the screen-outs to a high of 6 percent of the no-shows. The average overall missing data rate is 4 percent.

e. Includes only individuals whose highest wage rate on previous job was less than or equal to $25 per hour. Those who had never worked for pay are coded zero and included in the average. Those who worked for an unknown wage rate are assumed to have earned the average reported wage rate for their subpopulation (withdrawals, etc.), intake rating, and race.

f. Includes $0 values for those with no earnings reported in the IRS data.

Appendix B
Regression Procedures

In the analysis described in chapters 3 and 4 we use data from the Internal Revenue Service, which provides seven years of earnings information for the AFDC Homemaker-Home Health Aide Demonstration population (1984-90). As described in appendix A, earnings data from the IRS are in the form of means for groups of 10 to 19 individuals. All explanatory variables in the analysis are therefore expressed as group means or as percentages of group totals.

This appendix explains the difficulties associated with using grouped data in regression analyses and describes our use of generalized least-squares (GLS) to surmount these difficulties. It also contains regression results for the final, fully developed equation discussed in chapter 3, for each of three comparison groups and from the experiment itself.

Two problems arise in using ordinary least-squares (OLS) on grouped data to measure program impacts. Both stem from the fact that group-level observations have different variances even when the underlying individual-level data are homoskedastic. Because of this heteroskedasticity, OLS coefficient estimates (though unbiased) are no longer efficient—i.e., the OLS coefficients do not have the minimum variance among all linear unbiased estimators.[1] Also, the usual standard errors for the OLS coefficients will be biased because of heteroskedasticity, invalidating the standard tests of statistical significance.

For our purposes, biased standard errors represent the more important problem.[2] We surmount it by using the usual GLS procedure for dealing with heteroskedastic error terms, weighted least-squares (WLS). More formally, we assume that an OLS regression of earnings on a set of explanatory variables at the individual level would satisfy the classical requirements; namely, that the regression disturbances are not correlated with one another, and that the disturbances all come from the same normal distribution with mean zero and constant variance (i.e., the disturbances are homoskedastic). Letting N_g be the number of individuals in group g (in our case, $10 \leq N_g \leq 19$), each equation estimated in chapter 3 takes the form:

$$y = XB + \varepsilon,$$
$$E[\varepsilon] = 0,$$
$$Cov[\varepsilon] = E[\varepsilon\ \varepsilon'] = \Omega = \sigma^2\Psi$$

where:

$$y = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_G)'$$

$$X = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{1k} \\ \cdot & \\ & \cdot \\ & \cdot \\ \bar{x}_{G1} & \bar{x}_{Gk} \end{bmatrix}$$

$$\varepsilon = (\bar{\varepsilon}_1, \bar{\varepsilon}_2, \ldots, \bar{\varepsilon}_G)'$$

and

$$\bar{y}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} y_{ig}$$

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{ig}$$

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \varepsilon_{ig} \text{ , } \textit{for } g = 1, \ldots, G.$$

The variance of each residual is then:

$$E\left[\bar{\varepsilon}_g^{2}\right] = \frac{1}{N_g^2} E\left[\left(\sum_{i=1}^{N_g} \varepsilon_{ig}\right)^2\right] = \frac{N_g \sigma^2}{N_g^2} = \frac{\sigma^2}{N_g}$$

Thus, the residuals are heteroskedastic, since the variances depend on the number of observations in each group. Observations that correspond to large groups will have smaller variances than observations for small groups.

The GLS procedure weights each observation by the square root of its group size. Since the group sizes N_g are known, we can form the diagonal weighting matrix:

$$\Psi^{-1} = \begin{bmatrix} N_1 & & 0 \\ & \ddots & \\ 0 & & N_G \end{bmatrix} = diag\,(N_1,N_2,\ldots,N_G)$$

The GLS parameter estimates are then just:

$$\begin{aligned} \hat{B} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= \left(X'\frac{1}{\sigma^2}\psi^{-1}X\right)^{-1}X'\frac{1}{\sigma^2}\psi^{-1}y \\ &= (X'\psi^{-1}X)^{-1}X'\psi^{-1}y \end{aligned}$$

When the vector of coefficients is estimated in this way, the estimated standard errors are unbiased.[3]

This method is used to form all of the estimates in chapters 3 and 4, including the experimental impact estimates. Exhibits B.1 through B.7 present individual GLS regression results for each nonexperimental method, together with the experimental benchmark regressions, by year, from the fully specified equation that includes ratings of applicant potential among the explanatory variables. Exhibits B.8 and B.9 do the same for regressions representing the in-program and postprogram periods, where the dependent variable is the sum of the group means for 1985-86 and 1987-90, respectively. In each case, the coefficient on the participant dummy variable provides the impact estimate shown in the text.

NOTES

[1] Intuitively, estimation based on data for which the dependent variable is grouped will be less efficient than estimation using individual data because information is lost about the variation of earnings within groups.

[2] Correcting for the lack of efficiency in the coefficient estimates would require more complex adjustments that take account of the within-group variance of earnings. While the data are available for this purpose, we did not judge the gain in statistical precision (which we expected to be minimal) worth the added complexity.

[3] For a complete discussion of the use of GLS in the context of grouped data, see Kmenta (1986).

Exhibit B.1 1984 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Mean earnings, 1984	$826	$741	$819	$779
Coefficients of independent variables				
Intercept	259 (944)	-723 (715)	-731 (886)	541 (627)
Participant dummy	121 (259)	157 (124)	26 (145)	99 (74)
Average age	30 (21)	41 (18)	44 (23)	6 (16)
% Black	-325 (231)	-274 (201)	-217 (246)	19 (187)
% Hispanic	-186 (321)	-540 (293)	-416 (350)	-261 (254)
% Other race (or unknown)	-674 (443)	-99 (335)	36 (636)	301 (442)
% Education < 12 years	-1,139 (446)	-843 (353)	-1,114 (460)	-397 (347)
% Education = 12 years	-811 (484)	-195 (404)	-252 (500)	-72 (376)
% Education missing	2,072 (2,360)	3,850 (3,589)	4,287 (3,700)	4,129 (2,678)
% Never married	-792 (624)	89 (529)	379 (725)	303 (552)
% Married, spouse present	247 (418)	781 (340)	582 (427)	181 (316)
% Marital status missing	-4,683 (2,720)	-4,941 (3,600)	-6,190 (3,668)	-6,431 (3,093)
Avg. # dependent children	84 (148)	18 (120)	146 (145)	-141 (112)
% Ever worked for pay	-629 (269)	-517 (204)	-538 (270)	-490 (200)

Exhibit B.1 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	164 (51)	125 (44)	146 (56)	165 (42)
% Previous wage > $25	1,840 (3,233)	-2,125 (2,318)	N/A	N/A
% Kentucky	-39 (463)	471 (375)	306 (501)	74 (358)
% New Jersey	972 (385)	591 (327)	703 (389)	425 (289)
% New York	5,077 (2,720)	5,623 (3,575)	6,670 (3,637)	6,704 (3,078)
% Ohio	462 (414)	601 (368)	328 (448)	138 (318)
% South Carolina	189 (447)	471 (375)	510 (432)	524 (304)
% Texas	328 (411)	907 (342)	756 (380)	550 (290)
% State missing	3,608 (2,222)	2,131 (2,480)	2,841 (2,563)	2,103 (2,540)
% Rated "excellent"	189 (195)	183 (162)	150 (185)	252 (137)
% Rated "fair"	25 (168)	-41 (146)	32 (173)	-53 (130)
% Rated "poor"	251 (476)	493 (294)	-1,993 (1,418)	-1,418 (1,159)
% Rating missing	-69 (148)	-49 (130)	-259 (170)	-129 (129)
Adjusted R^2	.18	.22	.24	.13
Sample size	226	251	168	334

Exhibit B.2 1985 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Mean earnings, 1985	$3,205	$2,919	$3,804	$2,643
Coefficients of independent variables				
Intercept	-1,224	-3,177	-1,352	75
	(1,243)	(1,269)	(1,339)	(997)
Participant dummy	3,422	2,914	2,100	2,237
	(341)	(221)	(218)	(119)
Average age	28	62	43	21
	(28)	(31)	(35)	(26)
% Black	-69	68	-334	-80
	(305)	(357)	(371)	(297)
% Hispanic	754	1,101	525	407
	(423)	(524)	(530)	(404)
% Other race (or unknown)	-479	-673	-3,075	-1,021
	(587)	(593)	(958)	(703)
% Education < 12 years	-385	-276	-222	-707
	(597)	(634)	(708)	(552)
% Education = 12 years	-534	666	-57	-333
	(641)	(716)	(754)	(598)
% Education missing	-2,200	-13,183	-5,510	-6,785
	(3,112)	(6,378)	(5,606)	(4,271)
% Never married	-655	-1,272	-1,074	-882
	(825)	(938)	(1,094)	(878)
% Married, spouse present	604	1,017	1,163	452
	(551)	(610)	(648)	(502)
% Marital status missing	218	11,097	5,264	6,633
	(3,606)	(6,408)	(5,587)	(4,948)
Avg. # dependent children	-13	225	227	-79
	(195)	(212)	(219)	(178)
% Ever worked for pay	210	-231	318	97
	(359)	(367)	(412)	(321)

Exhibit B.2 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experiment: controls
Avg. previous wage	-143 (69)	-100 (79)	-186 (86)	-59 (68)
% Previous wage > $25	4,480 (4,258)	2,240 (4,117)	N/A	N/A
Avg. total earnings 1984	.87 (.09)	.85 (.12)	.76 (.13)	.89 (.09)
% Kentucky	1,910 (609)	2,526 (668)	2,625 (757)	1,659 (569)
% New Jersey	1,341 (515)	1,036 (584)	1,386 (593)	639 (460)
% New York	-117 (3,611)	-10,836 (6,371)	-4,470 (5,550)	-6,672 (4,928)
% Ohio	2,370 (547)	3,848 (656)	3,748 (677)	2,345 (505)
% South Carolina	767 (589)	987 (666)	1,313 (655)	950 (486)
% Texas	42 (542)	-145 (615)	328 (581)	277 (464)
% State missing	3,413 (2,944)	4,778 (4,403)	5,974 (3,881)	1,700 (4,040)
% Rated "excellent"	676 (257)	562 (287)	658 (280)	313 (218)
% Rated "fair"	-452 (221)	-364 (258)	-515 (260)	-169 (207)
% Rated "poor"	258 (627)	-1,289 (524)	3,753 (2,154)	1,801 (1,845)
% Rating missing	-55 (195)	-8 (230)	-190 (259)	32 (204)
Adjusted R^2	.75	.67	.66	.63
Sample size	226	251	168	334

Exhibit B.3 1986 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

	Comparison group			
Dependent variable mean	**Withdrawals**	**Screen-outs**	**No-shows**	**Experimental controls**
Mean earnings, 1986	$3,179	$3,026	$3,572	$3,006
Coefficients of independent variables				
Intercept	-727	-1,775	-2,014	177
	(1,652)	(1,453)	(1,765)	(1,133)
Participant dummy	2,003	1,545	1,145	1,023
	(454)	(253)	(288)	(135)
Average age	42	58	89	42
	(37)	(36)	(46)	(30)
% Black	-245	-17	-476	-82
	(406)	(409)	(490)	(338)
% Hispanic	521	954	647	696
	(562)	(600)	(699)	(459)
% Other race (or unknown)	-189	-98	-2,058	-539
	(780)	(679)	(1,264)	(798)
% Education < 12 years	-777	-129	27	-640
	(794)	(726)	(933)	(627)
% Education = 12 years	-325	520	532	138
	(853)	(820)	(995)	(679)
% Education missing	1,959	-470	4,863	2,798
	(4,137)	(7,300)	(7,391)	(4,850)
% Never married	205	-560	423	130
	(1,096)	(1,074)	(1,442)	(997)
% Married, spouse present	265	822	881	256
	(733)	(698)	(855)	(570)
% Marital status missing	-3,260	-3,552	621	-132
	(4,794)	(7,335)	(7,366)	(5,619)
Avg. # dependent children	-51	-110	6	-161
	(259)	(243)	(289)	(202)
% Ever worked for pay	-244	-564	-295	-3
	(477)	(420)	(544)	(365)

Exhibit B.3 (continued)

	Comparison group			
Dependent variable mean	**Withdrawals**	**Screen-outs**	**No-shows**	**Experimental controls**
Avg. previous wage	127	179	137	101
	(91)	(90)	(114)	(77)
% Previous wage > $25	3,237	-7,630	N/A	N/A
	(5,660)	(4,712)		
Avg. total earnings 1984	.77	.78	.75	.91
	(.12)	(.14)	(.17)	(.10)
% Kentucky	1,122	1,666	821	799
	(809)	(764)	(997)	(646)
% New Jersey	1,108	458	217	472
	(685)	(669)	(781)	(522)
% New York	5,063	-2,076	839	633
	(4,800)	(7,293)	(7,317)	(5,596)
% Ohio	2,043	3,448	3,216	1,728
	(727)	(751)	(892)	(574)
% South Carolina	-185	279	-62	225
	(783)	(763)	(863)	(552)
% Texas	479	182	358	98
	(720)	(704)	(765)	(527)
% State missing	2,225	-1,632	-1,927	-2,039
	(3,913)	(5,040)	(5,118)	(4,588)
% Rated “excellent”	466	551	355	68
	(341)	(329)	(369)	(248)
% Rated “fair”	-121	96	-112	-149
	(293)	(296)	(343)	(235)
% Rated “poor”	751	-592	8,092	4,719
	(834)	(599)	(2,940)	(2,095)
% Rating missing	-403	-101	-506	-274
	(259)	(263)	(341)	(232)
Adjusted R^2	.48	.44	.40	.39
Sample size	226	251	168	334

Exhibit B.4 1987 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Mean earnings, 1987	$3,620	$3,418	$3,809	$3,460
Coefficients of independent variables				
Intercept	569	-358	-809	722
	(1,970)	(1,544)	(2,025)	(1,325)
Participant dummy	635	883	916	551
	(539)	(268)	(331)	(157)
Average age	22	31	63	35
	(44)	(38)	(53)	(35)
% Black	8	-316	-728	-837
	(482)	(434)	(561)	(395)
% Hispanic	550	1,225	698	158
	(668)	(635)	(801)	(535)
% Other race (or unknown)	-822	-858	-2,427	-907
	(927)	(719)	(1,446)	(931)
% Education < 12 years	-2,577	-996	-1,344	-1,850
	(943)	(771)	(1,068)	(732)
% Education = 12 years	-1,125	-408	-740	-887
	(1,016)	(869)	(1,144)	(794)
% Education missing	2,357	-3,493	-703	-3,475
	(4,912)	(7,726)	(8,453)	(5,660)
% Never married	-1,093	-867	-2,085	-760
	(1,302)	(1,147)	(1,650)	(1,164)
% Married, spouse present	968	926	1,202	1,316
	(874)	(745)	(982)	(667)
% Marital status missing	1,580	7,347	5,538	8,403
	(5,693)	(7,761)	(8,426)	(6,560)
Avg. # dependent children	624	268	553	382
	(307)	(258)	(331)	(236)
% Ever worked for pay	-244	-316	-53	113
	(567)	(445)	(622)	(426)

Exhibit B.4 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	92	168	97	112
	(109)	(96)	(130)	(90)
% Previous wage > $25	-794	1,069	N/A	N/A
	(6,723)	(5,023)		
Avg. total earnings 1984	.90	1.11	1.19	1.03
	(.15)	(.14)	(.19)	(.12)
% Kentucky	1,729	988	40	361
	(961)	(812)	(1,141)	(754)
% New Jersey	1,954	2,056	1,584	1,467
	(813)	(709)	(894)	(610)
% New York	1,920	-5,289	-3,333	-6,509
	(5,703)	(7,717)	(8,371)	(6,534)
% Ohio	1,276	1,749	991	903
	(863)	(800)	(1,021)	(669)
% South Carolina	-144	677	-154	425
	(931)	(808)	(989)	(644)
% Texas	661	-373	-381	-37
	(856)	(747)	(876)	(615)
% State missing	-1,085	-1,552	-1,202	-3,901
	(4,648)	(5,335)	(5,857)	(5,357)
% Rated "excellent"	449	623	316	104
	(406)	(349)	(423)	(289)
% Rated "fair"	-436	-292	-581	-450
	(348)	(315)	(393)	(274)
% Rated "poor"	745	-1,032	3,174	1,270
	(990)	(636)	(3,248)	(2,445)
% Rating missing	-776	-135	-334	-208
	(308)	(283)	(391)	(271)
Adjusted R^2	.40	.44	.41	.33
Sample size	225	248	167	333

Exhibit B.5 1988 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

	Comparison group			
Dependent variable mean	**Withdrawals**	**Screen-outs**	**No-shows**	**Experimental controls**
Mean earnings, 1988	$4,315	$4,112	$4,489	$4,179
Coefficients of independent variables				
Intercept	797	491	-382	2,277
	(2,180)	(1,772)	(2,315)	(1,605)
Participant dummy	800	869	706	500
	(599)	(308)	(378)	(191)
Average age	26	-5	47	11
	(49)	(44)	(60)	(42)
% Black	-405	-99	-492	-692
	(535)	(498)	(642)	(479)
% Hispanic	190	1,107	581	-34
	(741)	(731)	(917)	(650)
% Other race (or unknown)	-604	-743	-2,040	-966
	(1,029)	(828)	(1,657)	(1,131)
% Education < 12 years	-1,338	87	57	-1,468
	(1,047)	(886)	(1,224)	(889)
% Education = 12 years	-538	483	363	-735
	(1,125)	(1,000)	(1,304)	(962)
% Education missing	6,236	6,462	8,326	-4,942
	(5,458)	(8,902)	(9,693)	(6,872)
% Never married	-936	-578	-1,907	-73
	(1,446)	(1,310)	(1,892)	(1,413)
% Married, spouse present	1,556	1,409	1,927	1,496
	(967)	(851)	(1,121)	(808)
% Marital status missing	-8,610	-3,119	-4,010	6,144
	(6,325)	(8,945)	(9,661)	(7,962)
Avg. # dependent children	472	276	416	235
	(341)	(297)	(379)	(286)
% Ever worked for pay	-634	-652	-360	-399
	(630)	(513)	(713)	(517)

Exhibit B.5 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	274	334	230	200
	(121)	(110)	(149)	(109)
% Previous wage > $25	2,029	4,652	N/A	N/A
	(7,469)	(5,746)		
Avg. total earnings 1984	.99	1.14	1.31	1.09
	(.16)	(.17)	(.22)	(.15)
% Kentucky	1,411	862	-21	716
	(1,068)	(932)	(1,308)	(916)
% New Jersey	1,445	863	334	1,038
	(904)	(816)	(1,025)	(740)
% New York	10,715	4,250	5,057	-4,497
	(6,334)	(8,893)	(9,596)	(7,930)
% Ohio	728	1,816	1,357	1,238
	(959)	(916)	(1,170)	(813)
% South Carolina	-9	671	-163	528
	(1,034)	(930)	(1,132)	(782)
% Texas	-748	-1,567	-1,928	-515
	(950)	(858)	(1,004)	(746)
% State missing	4,681	-381	-509	-332
	(5,163)	(6,146)	(6,712)	(6,501)
% Rated "excellent"	1,373	1,236	1,102	423
	(451)	(401)	(484)	(351)
% Rated "fair"	-814	-800	-952	-654
	(387)	(361)	(450)	(332)
% Rated "poor"	335	-1,989	5,088	2,377
	(1,100)	(731)	(3,724)	(2,969)
% Rating missing	-691	171	-96	-37
	(342)	(320)	(448)	(329)
Adjusted R^2	.39	.40	.38	.26
Sample size	226	251	168	334

Exhibit B.6 1989 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

	Comparison group			
Dependent variable mean	Withdrawals	Screen-outs	No-shows	Experimental controls
Mean earnings, 1989	$4,874	$4,694	$5,018	$4,775
Coefficients of independent variables				
Intercept	3,732 (2,673)	1,988 (2,118)	1,724 (2,477)	3,749 (1,755)
Participant dummy	1,316 (734)	597 (368)	814 (404)	338 (209)
Average age	1 (60)	-13 (52)	35 (64)	33 (46)
% Black	-797 (656)	-176 (596)	-537 (687)	-701 (523)
% Hispanic	-63 (909)	1,355 (874)	752 (981)	215 (711)
% Other race (or unknown)	-1,429 (1,262)	-160 (990)	-1,284 (1,773)	-1,605 (1,236)
% Education < 12 years	-3,154 (1,284)	-147 (1,059)	-577 (1,310)	-2,101 (972)
% Education = 12 years	-1,912 (1,380)	-328 (1,195)	-570 (1,396)	-1,666 (1,052)
% Education missing	5,704 (6,693)	4,140 (10,640)	4,278 (10,371)	-5,265 (7,513)
% Never married	-414 (1,774)	-1,243 (1,566)	-1,962 (2,024)	-944 (1,544)
% Married, spouse present	1,199 (1,186)	780 (1,017)	1,481 (1,200)	1,169 (883)
% Marital status missing	-11,580 (7,755)	-780 (10,691)	-1,975 (10,336)	6,380 (8,704)
Avg. # dependent children	105 (419)	6 (354)	-10 (406)	74 (313)
% Ever worked for pay	-1,007 (772)	-251 (613)	-220 (763)	-113 (565)

Exhibit B.6 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	266	225	112	83
	(148)	(132)	(160)	(119)
% Previous wage > $25	11,991	5,174	N/A	N/A
	(9,157)	(6,869)		
Avg. total earnings 1984	.87	1.23	1.40	1.12
	(.20)	(.20)	(.23)	(.16)
% Kentucky	1,872	1,152	1,671	833
	(1,309)	(1,114)	(1,400)	(1,001)
% New Jersey	2,973	1,630	1,159	1,011
	(1,108)	(975)	(1,096)	(809)
% New York	15,498	2,757	4,164	-4,716
	(7,766)	(10,630)	(10,267)	(8,669)
% Ohio	718	2,603	1,141	1,251
	(1,176)	(1,095)	(1,252)	(888)
% South Carolina	1,068	1,539	789	133
	(1,267)	(1,112)	(1,211)	(855)
% Texas	-270	-1,538	-1,680	-1,187
	(1,165)	(1,026)	(1,074)	(816)
% State missing	7,646	-1,398	5	-630
	(6,331)	(7,346)	(7,181)	(7,107)
% Rated "excellent"	1,396	1,174	1,109	371
	(552)	(479)	(518)	(384)
% Rated "fair"	-660	-711	-886	-856
	(474)	(432)	(481)	(363)
% Rated "poor"	511	-2,813	3,028	2,093
	(1,348)	(873)	(3,985)	(3,246)
% Rating missing	-1,076	-187	-972	-546
	(419)	(383)	(479)	(360)
Adjusted R^2	.34	.33	.40	.25
Sample size	226	251	168	334

Exhibit B.7 1990 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

Dependent variable mean	Comparison group Withdrawals	Screen-outs	No-shows	Experimental controls
Mean earnings, 1990	$5,332	$5,176	$5,469	$5,190
Coefficients of independent variables				
Intercept	4,509	2,370	2,066	4,045
	(2,661)	(2,282)	(2,602)	(1,899)
Participant dummy	2,074	658	867	402
	(731)	(397)	(425)	(226)
Average age	-1	-2	32	25
	(60)	(56)	(68)	(50)
% Black	-708	-55	-308	-517
	(653)	(642)	(722)	(566)
% Hispanic	222	1,827	1,273	503
	(905)	(942)	(1,031)	(769)
% Other race (or unknown)	-2,604 (1,256)	-83 (1,066)	-1,919 (1,863)	-840 (1,338)
% Education < 12 years	-3,348	-1,176	-1,162	-1,805
	(1,278)	(1,140)	(1,376)	(1,051)
% Education = 12 years	-2,576	-152	-670	-487
	(1,373)	(1,287)	(1,466)	(1,138)
% Education missing	8,805	9,449	9,467	3,898
	(6,663)	(11,463)	(10,895)	(8,129)
% Never married	477	-675	461	-641
	(1,766)	(1,687)	(2,126)	(1,671)
% Married, spouse present	943 (1,180)	608 (1,096)	1,535 (1,260)	660 (955)
% Marital status missing	-12,680	-516	-789	4,484
	(7,721)	(11,519)	(10,859)	(9,419)
Avg. # dependent children	-152 (417)	216 (382)	74 (426)	143 (339)
% Ever worked for pay	-1,604	-594	-542	-286
	(769)	(660)	(801)	(611)

Exhibit B.7 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	314 (147)	268 (142)	132 (168)	100 (129)
% Previous wage > $25	23,043 (9,116)	8,868 (7,400)	N/A	N/A
Avg. total earnings 1984	.93 (.20)	1.12 (.21)	1.32 (.25)	1.07 (.17)
% Kentucky	1,854 (1,304)	1,630 (1,200)	1,901 (1,470)	334 (1,083)
% New Jersey	3,239 (1,103)	2,052 (1,051)	1,676 (1,152)	754 (876)
% New York	16,975 (7,732)	3,076 (11,452)	4,005 (10,786)	-3,391 (9,380)
% Ohio	762 (1,171)	1,989 (1,180)	776 (1,315)	1,059 (961)
% South Carolina	931 (1,262)	1,057 (1,198)	373 (1,273)	103 (925)
% Texas	-128 (1,160)	-1,683 (1,105)	-1,633 (1,128)	-1,716 (883)
% State missing	5,760 (6,303)	-6,857 (7,914)	-5,287 (7,544)	-8,466 (7,691)
% Rated "excellent"	1,490 (550)	1,260 (516)	1,215 (544)	808 (415)
% Rated "fair"	-880 (472)	-1,013 (465)	-1,121 (506)	-752 (393)
% Rated "poor"	1,376 (1,342)	-2,170 (941)	4,453 (4,186)	802 (3,512)
% Rating missing	-1,224 (417)	-285 (413)	-1,143 (503)	-549 (389)
Adjusted R^2	.41	.32	.42	.24
Sample size	226	251	168	334

Exhibit B.8 1985–86 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Mean annual earnings, 1985-86	$3,192	$2,971	$3,688	$2,825
Coefficients of independent variables				
Intercept	-976	-2,476	-1,683	126
	(1,292)	(1,208)	(1,360)	(945)
Participant dummy	2,713	2,229	1,623	1,630
	(355)	(210)	(222)	(112)
Average age	35	60	66	32
	(29)	(30)	(35)	(25)
% Black	-157	30	-405	-81
	(317)	(340)	(377)	(281)
% Hispanic	637	1,027	586	551
	(439)	(498)	(539)	(383)
% Other race (or unknown)	-334	-385	-2,567	-780
	(609)	(564)	(974)	(666)
% Education < 12 years	-581	-202	-98	-674
	(620)	(604)	(719)	(523)
% Education = 12 years	-430	593	236	-97
	(666)	(681)	(766)	(566)
% Education missing	-120	-6,827	-323	-1,994
	(3,233)	(6,067)	(5,695)	(4,047)
% Never married	-225	-916	-326	-375
	(857)	(893)	(1,111)	(832)
% Married, spouse present	434	919	1,021	354
	(573)	(580)	(659)	(476)
% Marital status missing	-1,521	7,325	2,942	3,251
	(3,747)	(6,096)	(5,676)	(4,688)
Avg. # dependent children	-32	58	117	-120
	(202)	(202)	(223)	(169)
% Ever worked for pay	-7	-397	12	47
	(373)	(349)	(419)	(304)

Exhibit B.8 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	-8	40	-24	21
	(71)	(75)	(88)	(64)
% Previous wage > $25	3,858	-2,695	N/A	N/A
	(4,424)	(3,916)		
% Kentucky	1,516	2,096	1,723	1,229
	(633)	(635)	(769)	(539)
% New Jersey	1,224	747	801	555
	(535)	(556)	(602)	(436)
% New York	2,473	-6,456	-1,815	-3,019
	(3,752)	(6,061)	(5,638)	(4,669)
% Ohio	2,206	3,648	3,482	2,036
	(568)	(624)	(688)	(479)
% South Carolina	291	633	625	587
	(612)	(634)	(665)	(460)
% Texas	260	19	343	187
	(563)	(585)	(590)	(439)
% State missing	2,819	1,573	2,023	-170
	(3,058)	(4,188)	(3,944)	(3,828)
% Rated "excellent"	571	556	507	190
	(267)	(273)	(284)	(207)
% Rated "fair"	-286	-134	-314	-159
	(229)	(246)	(264)	(196)
% Rated "poor"	504	-940	5,922	3,260
	(651)	(498)	(2,188)	(1,748)
% Rating missing	-229	-54	-347	-121
	(202)	(218)	(263)	(194)
Adjusted R^2	.67	.61	.58	.56
Sample size	226	251	168	334

Exhibit B.9 1987–90 Full-Model Weighted Least-Squares Regression Estimates, by Comparison Group (standard errors in parentheses)

	Comparison group			
Dependent variable mean	**Withdrawals**	**Screen-outs**	**No-shows**	**Experimental controls**
Mean annual earnings, 1987-90	$4,532	$4,340	$4,691	$4,399
Coefficients of independent variables				
Intercept	2,289	1,077	560	2,657
	(2,078)	(1,647)	(1,994)	(1,455)
Participant dummy	1,214	763	809	443
	(571)	(286)	(325)	(173)
Average age	14	3	46	27
	(47)	(41)	(52)	(38)
% Black	-491	-191	-535	-698
	(510)	(463)	(553)	(434)
% Hispanic	241	1,395	845	214
	(706)	(680)	(790)	(589)
% Other race (or unknown)	-1,347	-355	-1,885	-1,075
	(981)	(780)	(1,428)	(1,025)
% Education < 12 years	-2,585	-522	-747	-1,800
	(998)	(823)	(1,055)	(806)
% Education = 12 years	-1.482	94	-339	-915
	(1,072)	(929)	(1,124)	(872)
% Education missing	5,781	4,327	5,380	-2,444
	(5,202)	(8,275)	(8,351)	(6,231)
% Never married	-489	-860	-1,403	-613
	(1,378)	(1,217)	(1,630)	(1,281)
% Married, spouse present	1,219	1,009	1,584	1,183
	(922)	(791)	(966)	(732)
% Marital status missing	-7,749	652	-227	6,443
	(6,028)	(8,315)	(8,323)	(7,219)
Avg. # dependent children	269	183	267	212
	(325)	(276)	(327)	(260)
% Ever worked for pay	-883	-471	-309	-178
	(600)	(476)	(614)	(469)

Exhibit B.9 (continued)

Dependent variable mean	Comparison group			
	Withdrawals	Screen-outs	No-shows	Experimental controls
Avg. previous wage	237 (115)	255 (102)	145 (128)	124 (99)
% Previous wage > $25	9,175 (7,117)	4,666 (5,342)	0 (0)	0 (0)
% Kentucky	1,715 (1,017)	1,145 (866)	906 (1,127)	555 (830)
% New Jersey	2,389 (861)	1,618 (758)	1,176 (883)	1,060 (671)
% New York	11,158 (6,036)	1,216 (8,267)	2,354 (8,267)	-4,883 (7,190)
% Ohio	868 (914)	1,999 (852)	1,081 (1,008)	1,115 (737)
% South Carolina	479 (985)	1,023 (865)	241 (976)	305 (709)
% Texas	-142 (906)	-1,339 (798)	-1,426 (865)	-871 (676)
% State missing	4,185 (4,920)	-2,697 (5,713)	-1,873 (5,782)	-3,417 (5,895)
% Rated "excellent"	1,193 (429)	1,100 (373)	949 (417)	433 (318)
% Rated "fair"	-688 (369)	-709 (336)	-872 (388)	-673 (301)
% Rated "poor"	737 (1,048)	-1,966 (679)	3,900 (3,209)	1,640 (2,692)
% Rating missing	-929 (326)	-86 (298)	-626 (386)	-329 (298)
Adjusted R^2	.44	.44	.49	.32
Sample size	226	251	168	334

Appendix C
Tests for Proven Bias in Nonexperimental Impact Estimates: Voluntary versus Mandatory Programs

In testing for selection bias in nonexperimental measures of the effects of employment and training programs, procedures differ between voluntary and mandatory programs. We explain why in this appendix. As noted in chapter 4, for voluntary programs these tests require more effort than is warranted in the current study. Tests for proven bias may prove worthwhile in other studies, in which case the analysis of alternative test methods provided here takes on heightened significance.

We begin with a general discussion of bias test procedures that readers with statistical backgrounds may wish to skip. We then deescribe our approaches to proving bias for mandatory and voluntary programs and critique an alternative voluntary approach from the literature.

The General Structure of Tests for Proven Bias

To prove that a nonexperimental impact estimate suffers from selection bias, one must first assume that an impact estimate from a randomized experiment does not. Under this assumption, and ignoring other possible sources of bias (e.g., a nonrepresentative sample), the experiment estimates true program impact without bias. To be unbiased, a nonexperimental estimator must estimate the same quantity. This proposition—that the experimental and nonexperimental estimators estimate the same quantity—serves as the null hypothesis in testing for proven bias. As in any statistical test, we look for conclusive evidence that the null hypothesis is wrong. Should we find such evidence, we will have proven that selection bias exists.

A standard method for determining whether two estimates estimate the same quantity is a test of the statistical significance of their difference—whether that difference differs significantly from zero. A test of this sort shows whether the difference between the two estimates could have arisen by chance alone while estimating a common quantity. If so, the null hypothesis of no difference is allowed to stand. If not, we reject the null hypothesis and conclude that the two estimates estimate two *different* quantities and, hence, that selection bias is present in the nonexperimental estimate.

Bias Tests Involving Overlapping Samples: The Case of Mandatory Programs

Tests for selection bias based on comparisons of experimental and nonexperimental impact estimates are complicated by the fact that the two estimates come from overlapping samples. In evaluating employment and training programs, data on program participants are necessarily included in both the experimental and nonexperimental estimates, creating an automatic correlation between the two measures which must be taken into account in conducting the test.

Fortunately, for mandatory training programs there is a way to eliminate the correlation. In this instance, the desired experimental estimate of impacts on participants is just the difference between the average treatment group outcome and the average control group outcome, since all treatment group members participate in the program to some extent (if only through the threat of sanctions for nonparticipation).[1] Each nonexperimental estimate is the difference between the average outcome for the treatment (i.e., participant) group and a "without-program" representation of that group derived from the comparison sample. Thus, the two estimates differ only in the use of a control group mean or a comparison group mean, measures drawn from nonoverlapping samples. The comparison of average outcomes for these two groups—which are uncorrelated—can become the focus of the test, implicitly "cancelling out" the participants who play a parallel role in both of the original impact estimates.

Though not indicated in their text, Friedlander and Robins (1992) presumably followed a procedure of this sort when testing for bias among nonexperimental estimates of the effects of mandatory AFDC employment and training programs.

Bias Tests Involving Overlapping Samples: The Case of Voluntary Programs

The cancellation strategy just described does not work when evaluating voluntary training programs such as the AFDC Homemaker-Home Health Aide Demonstrations or the Job Training Partnership Act (JTPA) program. Here, the desired experimental impact estimate is not simply the difference in average outcomes between participants and control group members. Rather, it is the broader treatment-control group difference adjusted for no-shows—individuals assigned to receive treatment who fail to participate in the program. As explained in chapter 3, an experiment estimates the average effect on participants by dividing the average effect on the treatment group by the treatment group participation rate.[2]

Because of the no-show adjustment, participants in voluntary programs are not compared to members of the control group in a way that allows for their "cancellation" when contrasting experimental and nonexperimental impact estimates. Nor can the entire treatment group be "cancelled" in a such a fashion, since the nonexperimental estimate does not use this group.[3] An alternative is to abandon the cancellation approach, calculate the covariance between the two impact estimates, and take it into account in testing for significant differences between the two. As explained in chapter 4, we view this added effort as unneeded in the current context, where we wish to prove the *absence* of bias rather than its presence. It may become essential in other applications, however.[4]

Removing Overlap by Cancelling Out the Entire Treatment Group

Another means of testing for selection bias in voluntary programs appears in the literature. Hotz (1991) suggests comparing control and comparison groups as a way of validating nonexperimental impact estimates from the National JTPA Study, a voluntary program with high no-show rates. Heckman and Roselius (1994) apply this technique to JTPA data on female youths, where the no-show rate among treatment group members is 33 percent. In identifying a comparison group that matches the female youth control group, they have in effect found a valid substitute for the experimental control group when estimating effects *on the entire treatment group*. It must be noted, however, that this is not the same as finding a valid counterfactual for participants alone when a substantial number of no-shows appear in the experimental treatment group.[5]

To use such a comparison group, one has to compare it to a combined sample of participants and no-shows. This raises several issues when thinking about future applications of the technique:

- To produce a reasonable estimate of a program's average impact on participants, the comparison of participants and no-shows to the comparison group would need to be followed by the no-show adjustment, just as in the experiment. As a result, future nonexperimental analyses would have to adopt the assumption of no effects on no-shows when, in principle, the no-shows need not enter into the analysis at all.

- The full technique has little intuitive appeal, proposing first to derive a nonexperimental estimate of program effects on the combined participant-plus-no-show sample, and then to recover the desired effect on participants-only by making the no-show adjustment. This eventual "removal" of no-shows from the analysis begs the question of why they were included in the first place.

- The need for a participant-plus-no-show sample—and the subsequent no-show adjustment—might well be overlooked in future applications of the method. Instead, studies might inappropriately adopt the comparison group as an appropriate counterfactual to participants alone.

- Correct use of the technique requires additional data collection for the no-show sample, a group that most conventional nonexperimental analyses ignore. No-shows may be difficult to identify in the field (if program operators are reluctant or unable to identify all of the applicants admitted to their programs) and will add to the sample for which baseline and follow-up data are needed.

In light of these difficulties, it would seem more useful to focus validation research for voluntary programs on comparisons of alternative impact estimators, not on comparisons of comparison and control groups.[6]

NOTES

[1] This is true of any program where treatment group members are subject to some type of intervention regardless of their own decisions. In such circumstances, all treatment group members are participants.

[2] As explained by Bloom (1984b), this procedure assumes no effect on nonparticipants, a standard assumption in experimental evaluations of voluntary employment and training programs. If r represents the share of treatment group members who participate in the program, the average effect of the program on the treatment group as a whole (*T*) can be decomposed into a weighted average of the average effect on participants (*P*) and a zero effect on no-shows: $T = r \bullet P + (1-r) \bullet 0$. Solving for *P*, we get $P = T/r$ as the no-show-adjusted impact estimate for participants.

[3] The nonexperimental estimate uses participants but not no-shows. We explore below the possibility of redefining nonexperimental estimates to include participants and no-shows.

[4] The problem of correlation between experimental and nonexperimental estimators could still be handled through a modification of the "cancellation" strategy, were it not for the inclusion of baseline variables in the standard impact model. (A description of the modified cancellation approach can be obtained from the first coauthor.) With the inclusion of baseline variables, the strategy fails because experimental and nonexperimental estimators are typically designed to relate baseline variables to participant outcomes *in different ways.* In deriving an experimental estimate, the relationship between baseline characteristics and outcomes is generally assumed to be the same for participants and no-shows, and usually for controls as well (the three groups included in the experimental analysis). In contrast, the relationship between baseline characteristics and outcomes is generally assumed to be the same for participants and *the comparison group* when deriving nonexperimental estimators. (The equations in chapter 3 illustrate these points.) It thus becomes impossible to "net out" the participant group from both estimators, since it enters in different ways in the two contexts.

[5] This assumes that participants differ from the treatment group as a whole due to self-selection following random assignment.

[6] There is one instance in which a comparison/control group emphasis could still provide a valid short-cut in testing for selection bias: if there are strong *a priori* reasons to expect a particular comparison group to match up well with the combined participant-plus-no-show sample (but not with the participant sample alone). This is true of one of the three applicant-based comparison groups tested in this monograph, screen-outs. For this comparison group approach, it would be possible to "cancel out" the participant-plus-no-show sample from the experimental and nonexperimental estimators to focus on the remaining contrast between the control and comparison groups. We do not take this approach here, however, since this option is not generally available in nonexperimental analyses of voluntary programs and cannot be used for the other two comparison groups covered in this monograph (withdrawals and no-shows).

Appendix D
Upper and Lower Bounds on the Risk Function

This appendix provides proofs that:

- The special case of a posterior distribution with *zero correlation* between the expected value of a nonexperimental impact estimator and true program impact *provides an upper bound* on the risk function defined in chapter 4; and
- The special case of a posterior distribution with *perfect positive correlation* (correlation coefficient = 1.00) between the expected value of a nonexperimental impact estimator and true program impact *does not provide a lower bound* on the risk function defined in chapter 4.

Proof that the Zero-Correlation Special Case Provides an Upper Bound on Risk

The general equation for risk from chapter 4 of the text states the probability of making an erroneous policy decision when using a nonexperimental impact estimator, N, to determine whether true impact, I, exceeds some cutoff value, C^*:

$$(1)\quad R(C^*) = Pr\,(N < C^* \text{ and } I > C^*) + Pr\,(N > C^* \text{ and } I < C^*)\,,$$

where the probabilities involved come from the joint posterior distribution of N and I, which is assumed to be normal. Equation (1) can be restated in terms of the conditional and marginal distributions of N and I, respectively, as:

$$(2)\quad R(C^*) = Pr\,(N < C^* \mid I > C^*) \bullet Pr\,(I > C^*) + Pr\,(N > C^* \mid I < C^*) \bullet Pr\,(I < C^*)\,.$$

When N and I are positively correlated,

$$(3)\quad Pr\,(N < C^* \mid I > C^*) < Pr\,(N < C^*)\,,$$

since knowing that true impact is larger than some threshold value ($I > C^*$) reduces the probability that the nonexperimental estimator is smaller than that value ($N < C^*$). By similar reasoning, positive correlation also implies that

$$(4)\quad Pr\,(N > C^* \mid I < C^*) < Pr\,(N > C^*)\,.$$

If we now substitute the larger right-hand expressions from equations (3) and (4) in place of the smaller lefthand expressions where they appear in equation 2, we get:

$$(5)\; R(C^*) < Pr\,(N < C^*) \bullet Pr\,(I > C^*) + \\ Pr\,(N > C^*) \bullet Pr\,(I < C^*)\,.$$

The right-hand side of this equation is precisely the expression for $R(C^*)$ obtained in the text (from equation (1)) when N and I are assumed to be uncorrelated (and, hence, through normality, independent). It follows, then, that risk derived under the assumption of zero correlation exceeds true risk whenever correlation is believed to be positive. In other words, the risk derived under the assumption of zero correlation provides an upper bound on true risk for any non-negative true correlation.

Proof that the Perfect Positive Correlation Special Case Does Not Provide a Lower Bound on Risk

Perfect positive correlation between N and I allows us to express N as a linear transformation of I. We do so in this section, in order to show that this special case does *not* automatically minimize the risk of policy error in comparison to cases where the relationship between N and I is not so tightly constrained.

In general, one can stipulate an entire family of linear transformations of I which each have the same marginal distribution as N:

$$(6)\quad F = E(N) - E(I) \cdot \sqrt{[\mathrm{Var}(N) - \mathrm{Var}(e)]/\mathrm{Var}(I)} + \\ \sqrt{[\mathrm{Var}(N) - \mathrm{Var}(e)]/\mathrm{Var}(I)} \cdot I + e\,,$$

where e is a normal random variable with mean 0 and variance Var(e) (< Var(N)) uncorrelated with N and I. Different values for Var(e) give different members of the family.

In general, equation (6) transforms two independent normal random variables, I and e, into a third normal random variable, F. The mean and variance of this new variable are $E(N)$ and Var(N), respectively; thus, N is a member of the family. The covariance and correlation of these variables with I are given by the formulas:

$$(7)\;\; \mathrm{Cov}\,(F, I) = \sqrt{[\mathrm{Var}(N) - \mathrm{Var}(e)]/\mathrm{Var}(I)} \cdot \mathrm{Var}(I)\text{ , and}$$

$$(8)\quad \mathrm{Cor}\,(F, I) \;=\; \mathrm{Cov}\,(F^{*}, I) \,/\, \sqrt{\mathrm{Var}\,(F) \cdot \mathrm{Var}\,(I)}$$

$$= \sqrt{[\mathrm{Var}\,(N) - \mathrm{Var}\,(e)\,] \,/\, \mathrm{Var}\,(N)}\ .$$

What distinguishes N from other variables in the family is its perfect positive correlation with I. To achieve $\mathrm{Cor}(F,I) = 1$, $\mathrm{Var}(e)$ must be 0. In this special case, equation (6) defines the particular family member N:

$$(9)\quad N \;=\; E\,(N) - E\,(I) \cdot \sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \;+$$

$$\sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \cdot I\,,$$

since e identically equals 0 whenever $\mathrm{Var}(e) = 0$.

We can now rewrite equation (1) replacing N with the linear transformation of I given in the right-hand expression of equation (9). This produces a risk function for the perfect positive correlation subcase:

$$(10)\quad R\,(C^{*}) \;=\; Pr\,\{E\,(N) - E\,(I) \cdot \sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \;+$$

$$\sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \cdot I < C^{*}, \text{ and } I > C^{*}\ \} \;+$$

$$Pr\,\{E\,(N) - E\,(I)\,\sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \;+$$

$$\sqrt{\mathrm{Var}\,(N) \,/\, \mathrm{Var}\,(I)} \cdot I > C^{*}, \text{ and } I < C^{*}\}\ .$$

Solving each probability term for I, this equation reduces to:

$$(11)\quad R\,(C^{*}) \;=\; Pr\,\{\,C^{*} < I < [\,C^{*} - E\,(N)\,] \cdot \sqrt{Var\,(I) \,/\, Var\,(N)} \;+$$

$$E\,(I)\,\} \;+ Pr\,\{\,[\,C^{*} - E\,(N)\,] \cdot \sqrt{Var\,(I) \,/\, Var\,(N)} \;+$$

$$E\,(I) < I < C^{*}\}\ .$$

One of the two terms in this expression is always 0 for a given C^{*}, since C^{*} cannot simultaneously be both below and above

$$[\,C^{*} - E\,(N)\,] \cdot \sqrt{\mathrm{Var}\,(I) \,/\, \mathrm{Var}\,(N)} + E\,(I)\ .$$

Without loss of generality, assume C^* exceeds this expression so that equation (11) reduces to:

$$(12)\quad R(C^*) = Pr\{[C^* - E(N)] \cdot \sqrt{\mathrm{Var}(I)/\mathrm{Var}(N)} + E(I) < I < C^*\}.$$

Exhibit D.1 shows this probability as the shaded area under the marginal posterior distribution of I. We need to consider whether other risk functions—those implied by posteriors without perfect positive correlation between N and I—would encompass a smaller or larger probability of error (i.e., area). Whatever its other properties, we know that in general N has mean $E(N)$ and variance $\mathrm{Var}(N)$ and so must be a member of the family F. When N does not correlate perfectly with I, $\mathrm{Var}(e)$ is not 0 and equation (6), rather than equation (9), becomes the starting point for deriving a risk index. Here, equation (12) becomes:

$$(13)\quad R(C^*) = Pr\{[C^* - E(N) - e] \cdot \sqrt{\mathrm{Var}(I)/\mathrm{Var}(N) - \mathrm{Var}(e)} + E(I) < I < C^*\}.$$

This change from equation (12) alters the left-hand boundary of the shaded region in exhibit D.1 in two ways:

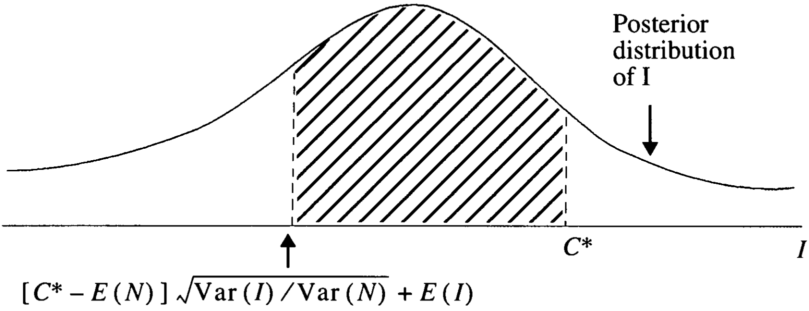
- By shifting it to the left or right, as the square-root factor declines in magnitude—right if C^*-$E(N)$ is positive, left if C^*-$E(N)$ is negative; and
- By adding variation to the boundary unrelated to the distribution of I or N, through the addition of the stocastic "-e" term.

Once e enters the equation, the risk index at C^* becomes an average of a range of shaded regions generated by this variability in the (now shifted) left-hand boundary of the region. Each resulting shaded region is weighted according to the posterior probabilities of various possible values of e.

The critical question is whether these two changes necessarily increase the value of the risk index at C^*, for all possible C^*s. If so, the perfect positive correlation case is a lower bound on risk; if not, it is not. In fact, each of the changes noted above can at times reduce the risk index.

Taking the changes one at a time, consider first a scenario where a general shift in the boundary is the only change. When $C^* > E(N)$, as it necessarily will be for some C^*, this shift moves the boundary to the right, reducing the shaded area under the curve, and hence the risk.

Exhibit D.1 Risk with Perfect Correlation Between N and I



Now suppose the boundary does not shift, but that the addition of the "-*e*" term causes it to vary across different possible values of *e*. Since *e* is normally distributed with a mean of 0, we derive the expected risk by attaching the greatest weight to the shaded region actually shown, where $e = 0$. This was the only region that received any weight in computing risks when *N* and *I* were perfectly correlated. It will now count for less than 100 percent and other regions for more than 0 percent. Each of these other regions can be smaller or larger than the region shown, depending on the sign of *e*.

Since normal distributions are symmetric, variations d units to the left of $[C^* - E(N)] \cdot \sqrt{\mathrm{Var}(I)/\mathrm{Var}(N)} + E(I)$are as likely as variations *d* units to the right of $[C^* - E(N)] \cdot \sqrt{\mathrm{Var}(I)/\mathrm{Var}(N)} + E(I)$. These two variations, and their associated shaded regions, will therefore get equal weights in our weighted average risk formula. They will not, however, add and subtract equal areas to the shaded region; whether the addition or subtraction is larger or smaller will depend on the position of the distribution of *I* relative to $[C^* - E(N)] \cdot \sqrt{\mathrm{Var}(I)/\mathrm{Var}(N)} + E(I)$. As pictured, each addition will be smaller than the associated subtraction for a given *d*, reducing the risk index.

When these two scenarios combine, the risk index unambiguously declines as the assumption of perfect positive correlation is removed. It follows, then, that the special case of perfect positive correlation does *not* provide a lower bound on risk.

References

Anderson, Kathryn H., Richard V. Burkhauser, and Jennie E. Raymond. 1992. "Reality or Illusion: The Importance of Creaming on Job Placement Rates in Job Training Partnership Act Title II-A Programs." Unpublished paper at Vanderbilt University.

Angrist, Joshua D., and Guido W. Imbens. 1991. "Sources of Identifying Information in Evaluation Models." NBER Technical Working Paper 117.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics* 60 (1): 47-57.

Ashenfelter, Orley, and David Card. 1985. "Using Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics* 67 (November): 648-660.

Ashenfelter, Orley, and G. E. Johnson. 1972. "Unionism, Relative Wages, and Labor Quality in U.S. Manufacturing Industries," *International Economic Review* 13: 488-508.

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22 (2): 157-193.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias," *Evaluation Studies* 5: 42-59.

Bassi, Laurie J. 1983. "The Effect of CETA on the Postprogram Earnings of Participants," *Journal of Human Resources* 18 (Fall): 539-556.

______ . 1984. "Estimating the Effect of Training Programs with Non-Random Selection," *Review of Economics and Statistics* 66 (February): 36-43.

Bassi, Laurie J., Margaret C. Simms, Lynn C. Burbridge, and Charles L. Betsey. 1984. *Measuring the Effects of CETA on Youth and the Economically Disadvantaged.* Washington. DC: Urban Institute.

Bell, Stephen H., Nancy R. Burstein, and Larry L. Orr. 1987. *The AFDC Homemaker—Home Health Aide Demonstrations: Overview of Evaluation Results.* Bethesda, MD: Abt Associates.

Bell, Stephen H., Jane Kulik, John Blomquist, Michelle L. Wood, Valerie Leiter, and Stephen D. Kennedy. 1993. *Project NetWork Evaluation: Research Design Report.* Bethesda, MD: Abt Associates.

Bell, Stephen H., Glen Cain, Larry Orr, and Winston Lin. 1993. "Measuring Employment and Training Program Impacts with Data on Program Applicants." Paper presented to the American Economic Association Annual Research Conference. Bethesda, MD: Abt Associates.

Bell, Stephen H., John H. Enns, and Kathleen L. Flanagan. 1987. *The AFDC Homemaker-Home Health Aide Demonstrations: Trainee Employment and Earnings*. Bethesda, MD: Abt Associates.

Bell, Stephen H., and Larry L. Orr. 1988. "Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker-Home Health Aide Demonstration." Paper presented to the Association for Public Policy Analysis and Management Annual Research Conference. Bethesda, MD: Abt Associates.

______ . 1994. "Is Subsidized Employment Cost-Effective for Welfare Recipients? Experimental Evidence from Seven State Demonstrations," *Journal of Human Resources* 29 (1): 42-61.

Bell, Stephen H., and Cilla J. Reesman. 1987. *The AFDC Homemaker-Home Health Aide Demonstrations: Trainee Potential and Performance.* Bethesda, MD: Abt Associates.

Benus, Jacob M., and Rhonda M. Byrnes. 1993. *The St. Louis Metropolitan Re-Employment Project: An Impact Evaluation.* Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Berger, Mark C., and Dan A. Black. 1992. "Child Care Subsidies, Quality of Care, and the Labor Supply of Low-Income, Single Mothers," *Review of Economics and Statistics* 74: 635-642.

Bloch, Farrell, ed. 1979. *Evaluating Manpower Training Programs.* Supplement 1 to *Research in Labor Economics.* Greenwich, CT: JAI Press.

Bloom, Howard S. 1984a. "Estimating the Effect of Job-Training Programs, Using Longitudinal Data: Ashenfelter's Findings Reconsidered," *Journal of Human Resources* 19 (Fall): 544-556.

______ . 1984b. "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review* 8 (April): 225-246.

Bloom, Howard S., and M. A. McLaughlin. 1982. *CETA Training Programs: Do They Work For Adults?* Washington DC: Congressional Budget Office and National Commission for Employment Policy.

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, and Fred Doolittle. 1993. *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months.* Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Bloom, Howard S., Larry L. Orr, Fred Doolittle, Joseph V. Hotz, and Burt S. Barnow. 1988. *Design of the National JTPA Study.* Bethesda, MD: Abt Associates and Manpower Demonstration Research Corporation.

Bloom, Howard S., and Neil M. Singer. 1979. "Determining the Cost-Effectiveness of Correctional Programs: The Case of Patuxent Institution," *Evaluation Quarterly* 3: 609-628.

Borus, Michael E. 1964. "A Benefit Cost Analysis of the Economic Effectiveness of Retraining the Unemployed," *Yale Economic Essays* 4 (Fall): 371-430.

Borus, Michael E., John P. Brennan, Sidney Rosen. 1970. "A Benefit-Cost Analysis of the Neighborhood Youth Corps: the Out-of-School Program in Indiana," *Journal of Human Resources* 5 (Spring): 139-159.

Borus, Michael E., and Charles G. Buntz. 1972. "Problems and Issues in the Evaluation of Manpower Programs," *Industrial and Labor Relations Review*: 234-245.

Bound, John. 1989. "The Health and Earnings of Rejected Disability Insurance Applicants," *American Economic Review* 79 (June): 482-503.

______ . 1991. "The Health and Earnings of Rejected Disability Insurance Applicants: Reply," *American Economic Review* 81 (December): 1427-1434.

Bryant, Edward C., and Kalman Rupp. 1987. "Evaluating the Impact of CETA on Participant Earnings," *Evaluation Review* 11 (August): 473-492.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources* 21 (4): 606-639.

Cain, Glen G. 1968. "Benefit Cost Estimates for Job Corps." Discussion Paper-68. Institute for Research on Poverty, University of Wisconsin—Madison.

______. 1975. "Regression and Selection Models to Improve Nonexperimental Comparisons." In *Evaluation and Experiment*, C.A. Bennett and A.A. Lumsdaine, eds. New York: Academic Press.

Cain, Glen G., Stephen Bell, Larry Orr, and Winston Lin. 1993. "Using Data on Applicants to Training Programs to Measure the Program's Effects on Earnings." Discussion Paper 1015-93, Institute for Research on Poverty, University of Wisconsin—Madison.

Cain, Glen G., and Robinson G. Hollister. 1969. "The Methodology of Evaluating Social Action Programs." In *Public-Private Manpower Policies,* F.H. Cassell, W. Ginsburg, A.R. Weber, eds., Madison, WI: Industrial Relations Research Association, pp. 5-33.

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

Card, David, and Daniel G. Sullivan. 1988. "Measuring the Effect of the CETA Program on Movements In and Out of Employment," *Econometrica* 56: 497-530.

Cave, George. 1988. "Noncompliance Bias in Evaluation Research: Assignment, Participation, and Impacts." Unpublished paper at the Manpower Demonstration Research Corporation.

Chamberlain, Gary. 1986. "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics* 32: 189-218.

Collignon, Fredrick C., Sydelle Raffe, Mary Vencill, Laurie Glass, and Reed Grier. 1989. *Use of the Social Security Data-Link for Assessing the Impact* *of the Federal-State Vocational Rehabilitation Program.* Berkeley, CA: Berkeley Planning Associates.

Cooley, Thomas, Timothy W. McGuire, and Edward C. Prescott. 1979. "Earnings and Employment Dynamics of Manpower Trainees: An Exploratory Econometric Analysis." In *Evaluating Manpower Training Programs*, Farrell E. Bloch, ed. Greenwich, CT: JAI Press.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally.

Couch, Kenneth A. 1992. "Long-Term Effects of the National Supported Work Experiment, and Parametric and Nonparametric Tests of Model Specification and the Estimation of Treatment Effects." Unpublished Ph.D. dissertation, University of Wisconsin—Madison.

Dean, David H., and Robert C. Dolan. 1991. "Fixed-Effects Estimates of Earnings Impacts for the Vocational Rehabilitation Program," *Journal of Human Resources* 26 (2): 381-391.

DeGroot, Morris H. 1970. *Optimal Statistical Decisions.* New York: McGraw-Hill.

______ . 1975. *Probability and Statistics.* Reading, MA: Addison-Wesley.

Dickinson, Katherine P., David Drury, Marlene Franks, Deborah Kogan, Laura Schlichtmann, Richard W. West, and Mary Vencill. 1988. *Evaluation of the Effects of Performance Standards on Clients, Services, and Costs.* Washington DC: National Commission for Employment Policy.

Dickinson, Katherine P., Terry R. Johnson, and Richard W. West. 1984. *An Analysis of the Impact of CETA Programs on Participants' Earnings.* Menlo Park, CA: SRI International.

______ . 1986. "An Analysis of the Impact of CETA Programs on Participants' Earnings," *Journal of Human Resources* 21: 64-91.

Director, Steven M. 1979. "Underadjustment Bias in the Evaluation of Manpower Training," *Evaluation Quarterly* 3 (May): 190-218.

Farber, David. 1972. "An Analysis of Changes in Earnings of Participants in Manpower Training Programs." Internal staff paper of the Manpower Administration, U.S. Department of Labor.

Fishman, Michael E., and Daniel H. Weinberg. 1991. "The Role of Evaluation in State Welfare Reform 'Waiver' Demonstrations." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources* 22 (2): 194-227.

Friedlander, Daniel, and Gary Burtless. 1992. "Job Losses and Return to Welfare: The Impacts of Welfare Employment Programs." Paper presented to the Association for Public Policy Analysis and Management Annual Research Conference. New York: Manpower Demonstration Research Corporation.

Friedlander, Daniel, and Philip K. Robins. 1992. "Estimating the Effects of Employment and Training Programs: An Assessment of Some Nonexperimental Techniques." Paper presented to the American Economic Association Annual Research Conference. New York: Manpower Demonstration Research Corporation.

Garfinkel, Irwin, Charles F. Manski, and Charles Michalopoulos. 1991. "Are Micro-Experiments Always Best? Randomization of Individuals or Sites," in *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Geraci, Vincent J. 1984. "Short-Term Indicators of Job Training Program Effects on Long-Term Participant Earnings." Discussion paper, Center for Economic Research, University of Texas-Austin.

Gibbard, Harold A., and Gerald G. Somers. 1968. "Government Retraining of the Unemployed in West Virginia." In *Retraining the Unemployed*, Gerald G. Somers, ed. Madison, WI: University of Wisconsin.

Ginsburg, Paul B. 1985. "Macroexperiments versus Microexperiments for Health Policy: Comment." In *Social Experimentation,* Jerry A. Hausman and David A. Wise, eds. Chicago: University of Chicago Press.

Goldberger, Arthur. 1972. "Selection Bias in Evaluating Treatment Effects." Discussion Paper 123-72, Institute for Research on Poverty, University of

______ . 1983. "Abnormal Selection Bias." In *Studies in Econometrics, Time Series and Multivariate Statistics*, T. Amemiya and I. Olkin, eds. Orlando: Academic Press.

Goldfarb, Robert S. 1969. "The Evaluation of Government Programs: The Case of New Haven's Manpower Training Activities," *Yale Economic Essays* (Fall): 59-104.

Goldstein, Jon H. 1972. *The Effectiveness of Manpower Training Programs: A Review of Research on the Impact on the Poor.* Staff study, Subcommittee on Fiscal Policy, Joint Economic Committee. Washington DC: Government Printing Office.

Greenberg, David, and Marvin Kosters. 1973. "Income Guarantees and the Working Poor." In *Income Maintenance and Labor Supply,* G. G. Cain and G. W. Watts, eds. Chicago: Rand McNally.

Greenberg, David, Robert H. Meyer, and Michael Wiseman. 1992. "Prying the Lid from the Black Box: Plotting Evaluation Strategy for Employment and Training Programs." Paper presented to the Association for Public Policy Analysis and Management Annual Research Conference. Madison, WI: University of Wisconsin—Madison.

Greenberg, David, and Mark Shroder. 1991. "Digest of the Social Experiments." Institute for Research on Poverty Special Report No. 52. Madison, WI: University of Wisconsin—Madison.

Greenberg, David, and Michael Wiseman. 1992. "What Did the Work-Welfare Demonstrations Do?" Discussion Paper 969-92, Institute for Research on Poverty, University of Wisconsin—Madison.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work.* New York: Russell Sage Foundation.

Hardin, Einar. 1969. "Benefit-Cost Analyses of Occupational Training Programs: A Comparison of Recent Studies." In *Cost-Benefit Analysis of Manpower Policies*, Gerald G. Somers and W. Donald Wood, eds. Kingston, Ontario: Industrial Relations Centre.

Hardin, Einar, and Michael Borus. 1971. *Economic Benefits and Costs of Retraining Courses in Michigan.* Lexington, MA: D.C. Heath.

Harris, Jeffery E. 1985. "Macroexperiments versus Microexperiments for Health Policy." In *Social Experimentation*, Jerry A. Hausman and David A. Wise, eds. Chicago: University of Chicago Press.

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply," *Econometrica* 42 (July): 679-694.

______ . 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5 (4): 475-492.

______ . 1979. "Sample Selection Bias as a Specification Error," *Econometrica* 47: 153-161.

______ . 1990. "Varieties of Selection Bias," *American Economic Review* 80(2): 313-318.

______ . 1991. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs," *Journal of the American Statistical Association* 84: 862-874.

Heckman, James J., V. Joseph Hotz, and M. Dabos. 1987. "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" *Evaluation Review* 11: 395-427.

Heckman, James J., and Richard R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, J. Heckman and B. Singer, eds. Cambridge, MA: Cambridge University Press.

______ . 1986. "Alternative Identifying Assumptions in Econometric Models of Selection Bias." In *Advances in Econometrics: Innovations in Quantitative Economics, Essays in Honor of Robert L. Basmann*, D. Slottje, ed. Greenwich, CT: JAI Press.

Heckman, James J., and Rebecca L. Roselius. 1994. "Evaluating the Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help a Lot." Unpublished mimeograph, Harris School of Public Policy, University of Chicago.

Heckman, James J., and Jeffery Smith. 1993. "Assessing the Case for Randomized Evaluation of Social Programs." Paper presented to the Danish Presidency Conference on Effects and Measuring Effects of Labour Market Policy Initiatives. Chicago: Center for Social Program Evaluation.

Heckman, James, Jeffrey Smith, and Christopher Taber. 1994. "Accounting for Dropouts in Evaluations of Social Experiments." Technical Working Paper No. 166, National Bureau of Economic Research, Cambridge, Massachusetts.

Hollister, Robinson G. Jr., Peter Kemper, and Rebecca A. Maynard. 1984. *The National Supported Work Demonstration.* Madison, WI: University of Wisconsin Press.

Horowitz, Joel L., and George R. Neumann. 1987. "Semiparametric Estimation of Employment Duration Models," *Econometric Reviews* 6 (1): 1-40.

Hotz, V. Joseph. 1991. "Recent Experience in Designing Evaluations of Social Programs: The Case of the National JTPA Study." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Imbens, Guido W., and Joshua D. Angrist. 1992. "Identification and Estimation of Local Average Treatment Effects." Discussion Paper No. 1617, Harvard Institute of Economic Research, Harvard University.

Johnston, Willliam B., Arnold E. Packer, Matthew P. Jaffe, Marylin Chou, Philip Delaty, Maurice Ernst, Adrienne Kearney, Jane Newitt, David Reed, Ernest Schneider, and John Thomas. 1987. *Workforce 2000: Work and Workers for the 21st Century.* Indianapolis, IN: Hudson Institute.

Kemper, Peter, David Long, and Craig Thornton. 1981. *The Supported Work Demonstration: Final Benefit-Cost Analysis.* New York: Manpower Demonstration Research Corporation.

Kiefer, Nicholas. 1979. "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Education," *Journal of Political Economy* 87 (October): S213-S226.

Kmenta, Jan. 1986. *Elements of Econometrics.* New York: Macmillian.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (4): 604-620.

LaLonde, Robert J., and R. Maynard. 1987. "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment," *Evaluation Studies* 11:

428-451.

Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Non-Experimental Data*. New York: Wiley.

______ . 1982. "Sets of Posterior Means with Bounded Variance Priors," *Econometrica* 50: 725-736.

Lee, Lung-Fei. 1978. "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review* 19:

415-434.

Long, David, Craig Thornton, and Christine Whitebread. 1983. *An Examination of the Benefits and Costs of the Employment Opportunities Pilot Project*. Princeton, NJ: Mathematica Policy Research.

Long, Sharon K., and Douglas A. Wissoker. 1992. "The Washington State Family Independence Program: Welfare Reform at Two Years." Paper presented to the American Economic Association Annual Research Conference. Washington DC: Urban Institute.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variable in Econometrics*. Cambridge, MA: Cambridge University Press.

Main, Earl D. 1968. *A Nationwide Evaluation of MDTA-Institutional Job Training*. Chicago: National Opinion Research Center.

Mallar, Charles, David Long, Stewart Kerachsky, and Craig Thornton. 1982. *Evaluation of the Impact of the Job Corps Program. Third Follow-up Report*. Princeton, NJ.: Mathematica Policy Research.

Manpower Demonstration Research Corporation. 1983. *Summary and Findings of the National Supported Work Demonstration*. Cambridge, MA: Ballinger.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects," *American Economic Review* 80 (May): 319-323.

______ . 1989. "Anatomy of the Selection Problem," *Journal of Human Resources* 24 (Summer): 343-360.

Manski, Charles F., and Irwin Garfinkel. 1991. "Issues in the Evaluation of Welfare and Training Programs." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Manski, Charles F., Gary D. Sandefur, Sara McLanahan, and Daniel Powers. 1992. "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation," *Journal of the American Statistical Association* 87 (March):

25-37.

Miller, Leslie, and Patricia Buckley. 1993. *Developing a High Performance Workforce*. Washington, DC: Manufacturers' Alliance for Productivity and Innovation.

Moffitt, Robert. 1987. "Symposium on the Econometric Evaluation of Manpower Training Programs," *Journal of Human Resources* 22 (2): 149-156.

______. 1991. "Program Evaluation with Nonexperimental Data," *Evaluation Review* 15: 291-314.

National Academy of Science. 1974. "Final Report of the Panel on Manpower Training Evaluation—The Use of Social Security Earning Data for Assessing the Impact of Manpower Training Programs." Washington DC: National Academy of Sciences.

Newey, Whitney K., James L. Powell, and James R. Walker. 1990. "The Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review* 80 (May): 324-328.

O'Neill, Dave M. 1973. *The Federal Government and Manpower: A Critical Look at the MDTA-Institutional and Job Corps Programs*. Washington DC: American Enterprise Institute for Public Policy Research.

Orr, Larry L. 1985. "Macroexperiments versus Microexperiments for Health Policy: Comment." In *Social Experimentation*, Jerry A. Hausman and David A. Wise, eds. Chicago: The University of Chicago Press.

______. 1987. *Evaluation of the AFDC Homemaker-Home Health Aide Demonstrations: Benefits and Costs*. Bethesda, MD: Abt Associates.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, and Fred Doolittle. 1994. *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A*. Bethesda, MD: Abt Associates.

Parsons, Donald O. 1991. "The Health and Earnings of Rejected Disability Insurance Applicants: Comment," *American Economic Review* 81 (December): 1419-1426.

Perry, Charles, R., Bernard E. Anderson, Richard L. Rowan, and Herbert R. Northrup. 1975. *The Impact of Government Manpower Programs: In General, and on Minorities and Women*. Manpower and Human Resource Studies No. 4. Industrial Research Unit, Wharton School, University of Pennsylvania.

Prescott, Edward, and T. F. Cooley. 1972. *Evaluating the Impact of MDTA Programs on Earnings Under Varying Labor Market Conditions*. Philadelphia, PA: University of Pennsylvania.

Puma, Michael J., Nancy R. Burstein, Katie Merrell, and Gary Silverstein. 1990. *Evaluation of the Food Stamp Employment and Training Program Final Report: Volume I.* Bethesda, MD: Abt Associates.

Reich, Robert B. 1983. *The Next American Frontier.* New York: Times Books.

Reichardt, Charles S., William M. K. Trochim, and Joseph C. Cappelleri. 1992. "Reports of the Death of Regression Discontinuity Analysis Are Greatly Exaggerated." Unpublished paper at the University of Denver.

Robin, Gerald G. 1969. *An Assessment of the In-Public School Neighborhood Youth Corps Projects in Cincinnati and Detroit, with Special Reference to Summer-Only and Year-Around Enrollees.* Philadelphia, PA: National Analysts.

Rupp, Kalman, Edward C. Bryant, and Richard E. Mantovani. 1983. *Factors Affecting the Participation of Older Americans in Employment and Training Programs.* Washington, DC: National Commission for Employment Policy.

Rupp, Kalman, Edward Bryant, Richard Mantovani, and Michael Rhoads. 1987. "Government Employment and Training Programs, and Older Americans." In *The Problem Isn't Age: Work and Older Americans,* Steven H. Sandell, ed. New York: Praeger.

Sandell, Steven H., and Kalman Rupp. 1988. *Who is Served in JTPA Programs: Patterns of Participation and Intergroup Equity.* Washington DC: National Commission for Employment Policy.

Smith, Ralph Ely. 1970. "An Analysis of the Efficiency and Equity of Manpower Programs." Unpublished Ph.D. dissertation, Georgetown University.

Solie, Richard J. 1968. "An Evaluation of the Effects of Retraining in Tennessee." In *Retraining the Unemployed,* Gerald G. Somers, ed. Madison, WI: University of Wisconsin.

Somers, Gerald G., ed. 1968. *Retraining the Unemployed.* Madison, WI: University of Wisconsin Press.

Stanley, T. D. 1991. " 'Regression-Discontinuity Design' by Any Other Name Might Be Less Problematic," *Evaluation Review* 15 (5): 605-624.

Stromsdorfer, Ernst W. 1968. "Determinants of Economic Success in Retraining the Unemployed: The West Virginia Experience," *Journal of Human Resources* 3 (Spring): 139-158.

Stromsdorfer, Ernst, Howard Bloom, Robert Boruch, Michael Borus, Judith Gueron, Alan Gustman, Peter Rossi, Fritz Scheuren, Marshall Smith, and Frank Stafford. 1985. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel.* Washington, DC: U.S. Department of Labor.

Thistlethwaite, D. L., and Campbell, D. T. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology* 51: 309-317.

Trochim, William M. K. 1984. *Research Design for Program Evaluation: The Regression Discontinuity Approach.* Beverly Hills, CA: Sage.

Westat, Inc. 1981. *Continuous Longitudinal Manpower Survey Net Impact Report No. 1: Impact on 1977 Earnings of New FY 1976 CETA Enrollees in Selected Program Activities.* Rockville, MD: Westat.

______ . 1984. *Continuous Longitudinal Manpower Survey: Summary of the Net Impact Results.* Rockville, MD: Westat.

Wiseman, Michael. 1991. "Research and Policy: A Symposium on the Family Support Act of 1988," *Journal of Policy Analysis and Management* 10: 588-666.

______ . 1993. "Welfare Reform in the States: The Bush Legacy," *Focus* 15 (Spring): 18-36.

U.S. Congress, Office of Technology Assessment. 1990. *Worker Training: Competing in the New International Economy.* Washington DC: Government Printing Office.

U.S. Department of Labor. 1970. "The Influence of MDTA Training on Earnings." Manpower Administration Evaluation Report No. 8.

U.S. General Accounting Office. 1992. "Feds Pay for 125 Programs, GAO Finds," *Employment and Training Reporter* 23 (August): 997.

INDEX

About the Institute

The W.E. Upjohn Institute for Employment Research is a nonprofit research organization devoted to finding and promoting solutions to employment-related problems at the national, state, and local level. It is an activity of the W.E. Upjohn Unemployment Trustee Corporation, which was established in 1932 to administer a fund set aside by the late Dr. W.E. Upjohn, founder of The Upjohn Company, to seek ways to counteract the loss of employment income during economic downturns.

The Institute is funded largely by income from the W.E. Upjohn Unemployment Trust, supplemented by outside grants, contracts, and sales of publications. Activities of the Institute are comprised of the following elements: (1) a research program conducted by a resident staff of professional social scientists; (2) a competitive grant program, which expands and complements the internal research program by providing financial support to researchers outside the Institute; (3) a publications program, which provides the major vehicle for the dissemination of research by staff and grantees, as well as other selected work in the field; and (4) an Employment Management Services division, which manages most of the publicly funded employment and training programs in the local area.

The broad objectives of the Institute's research, grant, and publication programs are to: (1) promote scholarship and experimentation on issues of public and private employment and unemployment policy; and (2) make knowledge and scholarship relevant and useful to policymakers in their pursuit of solutions to employment and unemployment problems.

Current areas of concentration for these programs include: causes, consequences, and measures to alleviate unemployment; social insurance and income maintenance programs; compensation; workforce quality; work arrangements; family labor issues; labor-management relations; and regional economic development and local labor markets.

W.E. UPJOHN
INSTITUTE
FOR EMPLOYMENT RESEARCH

0-88099-157-7