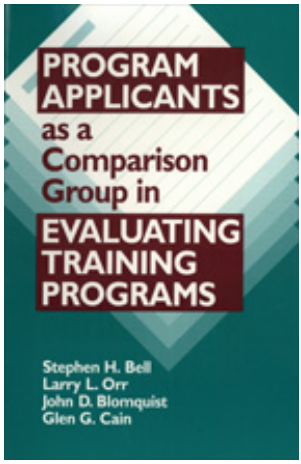
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Methods Used to Evaluate Employment and Training Programs in the Past



Stephen H. Bell
Abt Associates

Larry L. Orr
Abt Associates

John D. Blomquist
Abt Associates

Glen G. Cain
University of Wisconsin - Madison

Chapter 1 (pp. 1-19) in:
Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test
Stephen H. Bell, Larry L. Orr, John D. Blomquist, and Glen G. Cain
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1995

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 gov-

ernment 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) summa-

rizes 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 imple-

mented 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 gen-

eral 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.