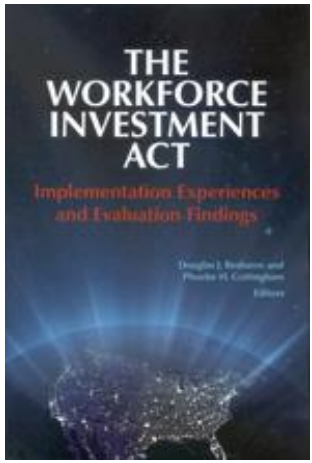
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Nonexperimental Impact Evaluations



Haeil Jung
Indiana University

Maureen A. Pirog
Indiana University

Chapter 14 (pp. 407-430) in:
The Workforce Investment Act: Implementation Experiences and Evaluation Findings
Douglas J. Besharov and Phoebe H. Cottingham, eds.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2011
DOI: 10.17848/9780880994026.ch14

The Workforce Investment Act

Implementation Experiences and Evaluation Findings

Douglas J. Besharov
Phoebe H. Cottingham
Editors

2011

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

Library of Congress Cataloging-in-Publication Data

The Workforce Investment Act : implementation experiences and evaluation findings / Douglas J. Besharov, Phoebe H. Cottingham, editors.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-88099-370-8 (pbk. : alk. paper)
ISBN-10: 0-88099-370-7 (pbk. : alk. paper)
ISBN-13: 978-0-88099-371-5 (hardcover : alk. paper)
ISBN-10: 0-88099-371-5 (hardcover : alk. paper)
1. Occupational training—Government policy—United States. 2. Occupational training—Government policy—United States—Evaluation. 3. Occupational training—Law and legislation—United States. 4. Employees—Training of—Law and legislation—United States. 5. Vocational guidance—Law and legislation—United States. 6. United States. Workforce Investment Act of 1998. I. Besharov, Douglas J. II. Cottingham, Phoebe H.

HD5715.2.W666 2011
370.1130973—dc22

2011010981

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

Cover design by Alcorn Publication Design.
Index prepared by Diane Worden.
Printed in the United States of America.
Printed on recycled paper.

14
Nonexperimental Impact Evaluations

Haeil Jung
Maureen A. Pirog
Indiana University

Job training for transitional workers and disadvantaged individuals is of keen interest for governments across the globe. Advancements in technology and globalized trade make some jobs obsolete or move them to lesser developed countries. Such structural transitions mean a sizable number of workers can lose their jobs. Also, inevitable business downturns lead to cyclical unemployment, which disproportionately affects disadvantaged workers with low human capital. In light of structural and cyclical changes in the labor markets, governments in industrialized nations have tried to support disadvantaged adults by retraining them. In the United States, training or retraining programs oftentimes have been accompanied by evaluations. This chapter briefly discusses what we have learned from these evaluations and then focuses on the related evaluation methods literature that informs how we can best design such evaluations in the future.

In the United States, there have been several major shifts in the goals, organization, groups targeted, and funding of employment and training programs. After the employment programs of the Great Depression, MDTA (1962–1972) was followed by CETA (1973–1982), JTPA (1982–1998), and eventually WIA (1998–present). CETA transformed a number of population-specific job training programs into block grants, which were then given to the states. This marked the first step in a devolutionary process that saw increased responsibility for job training delegated to states and localities. JTPA further devolved responsibility to the states. Later, WIA consolidated a number of USDOL job training programs and created One-Stop centers for job seekers negotiating their way through an otherwise bewildering system of federal

job training programs. WIA includes all adults aged 18 and older, as well as dislocated workers and disadvantaged youth aged 14–21.

The early evaluations of MDTA were nonexperimental (Perry et al. 1975) and largely rudimentary (Barnow and Smith 2009). Similarly, the CETA evaluations were nonexperimental. These evaluations all relied on the CETA Longitudinal Manpower Survey, which combined random samples of CETA participants with comparison group data constructed from the Current Population Survey. Barnow's 1987 review of the CETA evaluations concludes that they relied on crude matching estimators, and lacked local labor market data and recent labor market and program participation histories. Even more sophisticated matching procedures have failed to consistently replicate experimental findings (Barnow and Smith 2009; Pirog et al. 2009), and the absence of data on local labor markets, work, and program participation choices has been important in arriving at unbiased treatment effects (Card and Sullivan 1988; Dolton et al. 2006; Heckman and Vytlacil 2007).

The widely varying findings from the CETA evaluations led to the USDOL decision to evaluate JTPA as a randomized experiment. Doolittle and Traeger (1990) describe the experiment which took place in 16 of over 600 local JTPA sites, while Bloom et al. (1997) and Orr et al. (1996) describe the experimental impact results. A variety of authors have synthesized numerous evaluations of employment and training programs (Friedlander, Greenberg, and Robins 1997; Greenberg, Michalopoulos, and Robins 2003; Heckman, LaLonde, and Smith 1999; LaLonde 1995). Overall, these authors report somewhat disappointing results. Impacts for adults are modest, with more positive effects reported for women than men and negligible impacts for out-of-school youth (Greenberg, Michalopoulos, and Robins 2003). The limited effectiveness of job training programs is hardly surprising when we consider participants' overwhelmingly low human capital levels and relatively small amount of job training investment.

Within the related literature on program evaluation methodologies, there has been a hot debate over the accuracy of these largely nonexperimental findings. Researchers interested in government programs across the board have been investigating whether and under what circumstances carefully executed nonexperimental methods can provide robust estimates of treatment effectiveness. In fact, the experimental JTPA study provided data for a variety of studies that constructed

nonrandomized comparison groups and used various econometric corrections for self-selection bias to determine how effectively they work compared to the experimental results.

The approach of using experimental data to provide a benchmark against nonexperimental findings was used initially by LaLonde (1986) and Fraker and Maynard (1987). Both of these studies relied on data from the National Supported Work Demonstration. Other related studies of this type included Dehejia and Wahba (1999, 2002), Diaz and Handa (2006), Friedlander and Robins (1995), Heckman et al. (1996, 1998), Heckman and Hotz (1989), Heckman, Ichimura, and Todd (1997), Smith and Todd (2005), and Wilde and Hollister (2007).

LaLonde's 1986 study was particularly influential. He demonstrated that many self-selection correction procedures do not replicate estimated treatment effects in randomized experiments. In fact, nonexperimental methods were not robust to model specification changes in his study of the National Supported Work Demonstration, and the effectiveness of the program or estimated treatment effects were radically different from those determined experimentally. Later, Heckman, LaLonde, and Smith (1999) rebutted the LaLonde (1986) study in defense of nonexperimental methods, noting that each estimator is associated with testable assumptions and that by systematically testing them, the range of results resembles those originating from experimental methods.

The next section of this chapter provides a brief description of the types of parameters we may want to estimate in evaluating employment and training programs. After that we discuss conventional selection bias in studies of employment and training programs, followed by a discussion of pure selection bias and the robustness of different estimators that attempt to correct for self-section bias. The final section discusses what we have learned from previous studies.

FITTING THE METHODOLOGY TO THE POLICY QUESTION

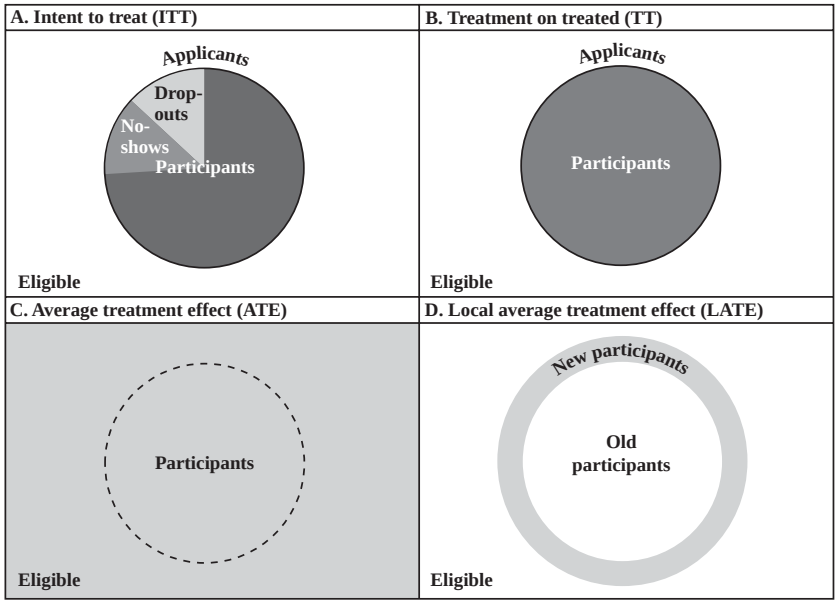
When evaluating the impacts of any program, researchers should ask two questions. First, what policy question do we need to answer? Second, what research designs and econometric methods are best suited

to answer the question? In employment and training programs, income (Y) is a typical outcome variable, although researchers have looked at a myriad of other possible outcomes, such as weeks worked, labor force attachment, and reliance on government cash assistance programs or poverty. Regardless of the outcome variable chosen (and for the purposes of this discussion, we focus on income), we need to establish a counterfactual. For example, we want to know the incomes of individuals given that they participated in a training program (Y_1) in order to compare it to the income of the same individuals without the benefit of the program (Y_0). In theory, a person can occupy either of these two potential states (treated or untreated), but in reality only one state is realized for a given individual. If people could occupy both states at the same time, then the problem of program evaluation would be easy and the treatment effect could be depicted as $\Delta = Y_1 - Y_0$. Four commonly discussed variants of treatment effects estimates are shown in Figure 14.1.

In practice, most randomized social experiments are designed to obtain intent to treat (ITT) estimates (Panel A of Figure 14.1). Eligible participants, frequently identified through administrative data, or those who have applied for services are randomly assigned to the treatment after which they comply with program requirements to some extent: some complete, others drop out, while still others are no-shows. When all individuals randomly assigned to treatment are compared to the randomized control group, the ITT estimates can be interpreted as the average impact over a sample of applicants, some of whom comply to some degree with the program. However, program administrators and supporters have often raised concerns with ITT estimates, arguing that they unfairly bias downward positive treatment effects by including the no-shows and even dropouts in the treatment group. After all, no-shows and dropouts either received no program services or only partial services. As such, no-shows and dropouts should not be expected to benefit either at all or fully from the program.

Largely in response to these concerns, experimenters created the treatment on the treated (TT) estimates. Individuals who started but dropped out at some point are typically, but not always, included with completers in these estimates. Viewed from this perspective, TT estimates are derivatives of ITT estimates—mechanical approximations with known properties and assumptions.

Figure 14.1 Variants of Treatment Effects



While most experimentors focus on ITT or TT estimation, it would be relatively straightforward to design a randomized experiment to estimate the impacts of program expansions (the local average treatment effect [LATE]). However, it is likely to be difficult to obtain good average treatment effects (ATE) estimates because randomly assigned individuals from an eligible population may well fail to comply with the treatment protocols. Moreover, unless treatment is mandated by court order or another mechanism, the usefulness of such estimates is rather limited. Each of these four types of estimators is discussed below.

ITT. This estimator is depicted in panel A of Figure 14.1. In this case,

$$\text{ITT} = E(Y_1 \mid D = 1, R = 1) - E(Y_0 \mid D = 1, R = 0),$$

including the no-shows and dropouts in the treatment group, where D = 1 if eligible individuals apply to the program and D = 0 if they do not,

and $R = 1$ for the treatment group members and $R = 0$ for the controls. Under many circumstances this is an interesting and policy-relevant parameter that reflects how the availability of a program affects participant outcomes when participation in the program is incomplete.

TT. When we want to estimate the effect of a treatment like a job training program on actual participants, the parameter of interest is the effect of TT, depicted as follows:

$$\text{TT} = E(\Delta \mid D = 1, X) = E(Y_1 - Y_0 \mid D = 1, X),$$

where X is a vector of individual characteristics, $D = 1$ if an individual participates in the program, and $D = 0$ if they do not.

In our example, TT could compare the earnings of vocational program participants with what they would earn if they did not participate in the program. This is the information required for an "all or nothing" evaluation of a program and provides policymakers with information on whether or not the program generates positive outcomes. In panel B of Figure 14.1, the TT is depicted as the effect of treatment on participants. Social experiments randomly assigning eligible applicants to the treatment and control groups are generally considered the gold standard for obtaining ITT and TT estimates.[1]

ATE. This is the average impact that results from randomly assigning a person from the eligible population to a treatment. In panel C of Figure 14.1, the shaded rectangle constitutes the entire population for which the treatment effect is being estimated, regardless of whether or not they chose to participate in the program. The ATE is shown mathematically as

$$\text{ATE} = E(\Delta \mid X) = E(Y_1 - Y_0 \mid X).$$

Neither component of this mean has a sample analogue unless there is universal participation or nonparticipation in the program, or if participation is randomly determined and there is full compliance with the random assignment. As such, the ATE can be difficult, sometimes impossible, to compute. More importantly, however, this estimator is typically uninteresting to policymakers, who are typically loath to force randomly selected individuals to participate in programs.

LATE. This is the effect of treatment on persons who were induced to participate by an expansion or increased generosity of a program (see panel D of Figure 14.1). For example, LATE could measure the effect of a change in a policy (Z) of providing a new stipend or a more generous stipend to vocational program participants on those induced to attend the program because of the new policy. LATE is shown as follows:

$$\text{LATE} = E(Y_1 - Y_0 \mid D(z) = 1, D(z') = 0) = E(Y_1 - Y_0 \mid D(z) - D(z') = 1)$$

where $D(z)$ is the conditional random variable D, given $Z = z$, and where z' is distinct from z, so $z \neq z'$. Two assumptions are required to identify LATE. First, Z does not directly affect the outcome and program participation is correlated with Z controlling for other factors. This is a typical assumption for IV estimation. Second, there must be compliance with the policy change such that there are no individuals who refuse to participate if eligible and want to participate if not.[2]

Because it is defined by variation in an instrumental variable that is external to the outcome equation, different instruments define different parameters. When the instruments are indicator variables that denote different policy regimes, LATE has a natural interpretation as the response to policy changes for those who change participation status in response to the new policy. For any given instrument, LATE is defined on an unidentified hypothetical population—persons who would certainly change from 0 to 1 if Z is changed. For different values of Z and for different instruments, the LATE parameter changes, and the population for which it is defined changes. In other words, when we estimate the LATE parameter, we need to make sure who is possibly affected by the policy change from z' to z and how to interpret the estimated value in terms of relevant policy changes.

Most randomized experiments focus on estimating the ITT or TT in order to answer the policy question of how a program changes the outcomes of eligibles or eligible applicants and actual program participants compared to what they would have experienced if they had not participated. The ATE estimator is infrequently used largely because most researchers and policymakers are reluctant to force program participation. Finally, when programs became more generous or eligibility

is expanded, the LATE estimator can be used to obtain the incremental effect of the policy change.

While random assignment studies are considered the gold standard for obtaining program impact estimates, the reality is that the vast majority of published evaluations are nonexperimental, with perhaps the exception of the randomized clinical trials in the medical literature (Pirog 2007). Thus, it is imperative to understand the issues relating to selection bias and the construction of a reasonable counterfactual. It is also important to follow closely the emerging literature on the nonexperimental designs, estimators and statistical approaches that give rise to estimates of treatment effects that better approximate those that would be found using random assignment studies. These issues are discussed in the next three sections of this chapter.

CONVENTIONAL SELECTION BIAS AND LESSONS FOR PROGRAM DESIGNS AND DATA COLLECTION

Before addressing which econometric methods are relevant to answer the policy question, we want to discuss the selection bias that occurs when participation in job training programs is not randomized. Randomization should result in statistically equivalent groups of treatment and control group members in terms of both their observed *and* unobserved characteristics. This is not the case with nonexperimental studies, which often rely on propensity score matching, instrumental variable approaches, difference-in-difference techniques, and other statistical corrections to attempt to create a reasonable counterfactual or comparison group.

Early in the still ongoing debate on the relative merits of experimental versus nonexperimental evaluation, LaLonde (1986) pointed out that the use of nonrandomized comparison groups in evaluations can lead to substantial selection bias. Heckman et al. (1996, 1998) countered that LaLonde reached his conclusions incorrectly by constructing his comparison groups from noncomparable data sources. LaLonde's comparison groups were located in different labor markets from program participants, and their earnings were measured using different questionnaires. Heckman also noted that LaLonde lacked

information on recent preprogram labor market outcomes, which are important predictors of participation in training. In sum, Heckman et al. (1998) concluded that simple parametric econometric models applied to bad data should not be expected to eliminate selection bias. In 1997, Heckman, Ichimura, and Todd (1997) showed how the bias found in estimates of treatment effectiveness can be decomposed into three sources. This analysis is still relevant for labor market researchers today who wish to construct a counterfactual or comparison group without the benefit of randomization.

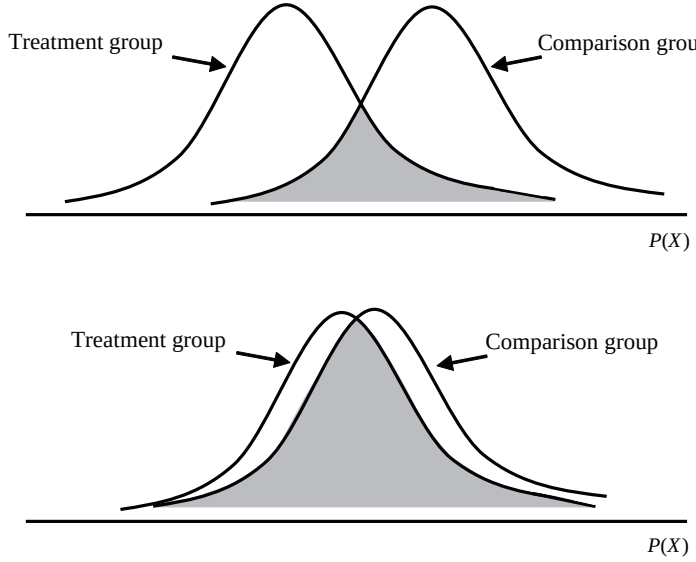
The first source of bias that can occur when using a nonrandomized comparison group relates to differences in the values of the same observed characteristics in the treatment and comparison groups. This would occur, for example, if the treatment group included individuals aged 20–60 and the comparison group only included individuals aged 30–40.

When we have many observed characteristics, *X*'s, they can be represented as *P*(*X*), the propensity score, which is the probability of participation in a program based on a vector of observed individual characteristics. The second source of bias occurs when propensity scores obtained by matching on observable characteristics have different distributions over the same range.

The top panel of Figure 14.2 depicts a situation where both sources of bias are serious. In the top panel, the treatment and comparison groups have a modest overlap in their propensity scores, *P*(*X*). In fact, no comparison group members are in the left tail of the distribution for the treatment group, and conversely, no treatment group members are in the right tail of the distribution for the comparison group. This difference reflects the first source of bias. In the top panel, you can also see that the distributions of propensity scores over the same range are different. This reflects the second source of bias. Both sources of bias are mitigated in the bottom panel of Figure 14.2.

The third source of bias in estimated treatment effects is from the pure self-selection on unobservables such as motivation. This would exist, for example, if the treatment group members of a job training program were highly motivated in contrast to comparison group members who lacked drive or motivation. This is the bias caused by the individuals' self-selection behavior based on information that researchers cannot observe and details of which are discussed later in the chapter.

Figure 14.2 Two Conventional Sources of Bias



Propensity score matching can moderate bias from the first two sources of bias. Reweighting comparison group members so that the distribution of the comparison group's $P(X)$ more closely resembles that of the treatment group can further reduce bias from the second source. Because much of the bias attributed to selection by LaLonde (1986) was actually due to the first two sources described above, Heckman, LaLonde, and Smith (1999) continue to make arguments in favor of nonexperimental evaluations.

CONSTRUCTING A COUNTERFACTUAL

The characteristics of different types of comparison groups, including the randomized control group, are described below. The conclusion that the quality of data used to form a comparison group and the matching procedures utilized are keys to reducing the conventional bias is based on Heckman, Ichimura, and Todd (1997), who used data from

a randomized control group, the no-shows from the treatment group, the eligible but nonparticipating group, and the Survey of Income and Program Participation (SIPP) in order to analyze the quality of the comparisons achieved.

Randomized Control Group as an Ideal Comparison Group

After applying to a program and being deemed eligible, individuals are randomly assigned to a control group. Data from the control and treatment groups should have nearly the same distribution of observed and unobserved characteristics. Because eligible applicants from the same local labor markets are randomly assigned to the treatment and control groups, and the same survey instruments were used with both groups, all three sources of bias should be controlled.

No-Shows from the Treatment Group as a Comparison Group

No-shows include individuals who are accepted to the program and randomized into the treatment group but who do not participate in the program. The simple mean difference between the treatment group and the no-show group without matching demonstrates that no-shows have similar characteristics as well as overlapping distributions of $P(X)$. The main source of bias is from selection on unobservables.

The Eligible but Nonparticipants (ENPs) as a Comparison Group

Individuals in the eligible but nonparticipating group are those who are located in the same labor market, and are eligible for the program but do not apply for the program. These individuals' information is collected by using the same questionnaire as for the treatment group. There were some clear differences in the characteristics and distribution of $P(X)$ between the ENPs and the treatment group members. By using propensity score matching and reweighting observations, it is possible to reduce the first two sources of bias as well as rigorously defined self-selection bias. While improvements in the estimated treatment effectiveness were obtained, the estimated treatment effect was still not equivalent to the TT estimate.

A Comparison Group from SIPP or Other Data Sources

To construct a comparison group, it is also possible to apply the eligibility criteria for a program to survey respondents in the SIPP or other large surveys. Two problems arise from using this approach. First, local labor market conditions are likely to be different for comparison and treatment group members when the comparison group members are selected from preexisting survey data. Second, data collected from the treatment and comparison groups are likely to come from different surveys or sources of measurement. In models comparing the treatment group with the SIPP comparison group, there was some discrepancy in observed characteristics and $P(X)$. They found that the first and second sources of bias were close to those found when using the ENPs for a comparison group. The discrepancies in the local labor markets and the questionnaires contributed to bias stemming from selection on unobservables; the third component of the selection bias is larger than that when they use ENPs.

Discussion

When we design training programs and collect information on participants to evaluate program effectiveness using nonexperimental methods, we need to consider how to develop comparison groups. Several factors are critical in reducing bias in our estimates of treatment effects: use the same questionnaire or data sources to obtain individual labor market outcomes and demographic information, draw individuals for the treatment and comparison groups from the same local labor markets, and use comparison group members whose observed characteristics largely overlap with those of the participants.

Restricting analyses to treatment and comparison group members with similar characteristics and using propensity score matching can reduce the first and second components of conventional bias, even though the characteristics of the parameter that we want to estimate can change. However, propensity score matching has its own limitations: it cannot control for self-selection on unobservables. Its uses and limitations are discussed with related empirical studies surveyed by Pirog et al. (2009). This study points out that matching is a nonparametric method that is flexible to any functional relationships between

outcomes and programs. However, it needs a large sample size and is sensitive to various matching methods. There is no clear guidance for superior matching procedures.

SOURCES OF PURE SELF-SELECTION BIAS AND EMPIRICAL METHODS

Different Sources of Pure Self-Selection Bias

There are three reasons why individuals might self-select into an employment and training program:

1) they know they will earn higher incomes after participating in the program (heterogeneous response to the program in a random coefficient model);
2) individuals select into the program because their latent or forgone earnings are low at the time of program entrance (time constant individual heterogeneity in a fixed effect model); and
3) individuals' earnings are dependent on previous earnings that are low at the time of program entrance (autocorrelation between earnings in different time periods).

The first source of self-selection implies that individuals with higher returns from the program are more likely to participate in training programs. The second source of self-selection behavior implies that individuals with low opportunity costs or low earnings capacity are likely to participate in training programs. The third source of self-selection behavior implies that the low earnings capacity that encourages program participation at the time of participation is positively associated with earnings after program. Thus, the first source of self-selection results in overestimates of the effectiveness of employment and training programs while the second and third sources of self-selection result in underestimates. In the employment and training literature, it is understood that all three sources of bias contribute to the phenomenon known as "Ashenfelter's dip"; the fact that participants in employment and training programs often have earnings that are temporarily low at

the time of program entry but that their earnings usually rebound (even in the absence of program participation) (Ashenfelter 1978).

Different empirical techniques appear to work better or worse depending on which sources of bias are operating. Theoretically, we expect that cross-sectional estimators provide consistent estimates only if there is no bias. Difference-in-differences estimators provide consistent estimates only if self-selection bias is coming from bias source 2. The AR (1) (autoregressive of order one) regression models provide consistent estimates only when self-selection bias is coming from bias source 3. The use of the instrumental variables method and the Heckman-selection correction provides consistent estimates only if bias sources 2 and 3 are present.[3] Thus, understanding which sources of bias we have in the program is critical in choosing which empirical method to use to best answer the policy question.

In simulations, cross-sectional estimation, difference-in-differences, and AR (1) regression estimation work relatively well when all three sources of bias are present, but it appears that they work well because the different biases offset one another (Heckman, LaLonde, and Smith 1999). Also, when bias source 1 is present, the estimation methods working for TT do not work for ATE. The authors argue that these parameters differ greatly because there is strong selection into the program by persons with high values of individual specific returns. However, they are not clear about how bias sources 1, 2, and 3 interact when different nonexperimental methods estimate ATE and TT. It seems that when all three bias sources are present, those three biases might offset one another. Difference-in-differences and AR (1) regression models also provide a similarly low bias in estimation. Finally, instrumental variables and the Heckman self-selection correction work best when bias sources 2 and 3 are present without bias source 1. However, when bias source 1 is present, IV and Heckman correction are the worst methods to use.

In sum, difference-in-differences and AR (1) regression estimators seem robust enough over different bias sources to estimate the TT. However, this does not mean that they are superior nonexperimental methods to others. In addition, it is not clear how offsetting of different bias sources works over different data and programs. Further research is needed.

NEW NONEXPERIMENTAL METHODS

Since the Heckman/LaLonde debate, a number of econometric methods have become more popular, and they relate directly to the issues of how best to estimate treatment effects for employment and training programs in the absence of random assignment. These additional methods include the difference-in-differences extension on matching, regression discontinuity design, and the marginal treatment effect (MTE) using local instrumental variables. Table 14.1 presents our summary of these methods as well as those for "kitchen sink" regression, propensity score matching, difference-in-differences, AR (1), and instrumental variables methods.

Difference-in-Differences Extension of Matching

As mentioned earlier, propensity score matching can be used to obtain impact estimates for treatment group members whose observable characteristics overlap with those of comparison group members. Of course, the impact estimates will only be valid for those individuals whose characteristics do overlap. Within the range of overlap of observables, the "comparable" comparison group can also be reweighted to better represent the distribution of observed treatment group characteristics, further reducing bias from different distributions of observables between treatment and comparison group members. Neither of these adjustments, however, controls for selection on unobservables.

Difference-in-differences extension of matching, introduced in Heckman, Ichimura, and Todd (1997), controls for some forms of selection on unobservables: it eliminates time-invariant sources of bias that may arise when program participants and nonparticipants are geographically mismatched or have differences in their survey questionnaire. Unlike traditional matching, this estimator requires the use of longitudinal data, which uses outcomes before and after intervention.

Regression Discontinuity Design

Regression discontinuity design became popular because it is easy to use and easy to present to a general audience. On the other hand, it

Table 14.1 Data, Methods, Self-Selection Behavior

Methods	Data	Consistency against self-selection on unobservables			Note
		(1)[a]	(2)[b]	(3)[c]	
"Kitchen sink" regression estimator	Cross-sectional data Repeated cross-sectional data Panel data	No	No	No	Strict parametric assumption on a control function.
Propensity score matching	Cross-sectional data Repeated cross-sectional data Panel data (Large sample is required)	No	No	No	Flexible nonparametric method but large sample is required. Good at moderating the bias from the mismatched observed characteristics between the treatment and the comparison, and the bias from the mismatched distribution in the common values of observed characteristics.
Difference-in-differences	Panel data	No	Yes	No	Sensitive to choosing different time points before and after the treatment period.
AR (1) regression estimator	Panel data	No	No	Yes	It does not need to have outcome before the program; outcomes of two periods after the program is enough. AR (1) process assumption itself can be restrictive to represent the earnings dependency in practice.

Instrumental variable method	Cross-sectional data Repeated cross-sectional data Panel data	No	Yes	Yes	Hard to find a valid instrument variable.
Difference-in-differences extension of matching	Panel data	No	Yes	No	Flexible nonparametric method but large sample is required. Good at moderating the bias from the mismatched supports between the treatment and the comparison, and the bias from the mismatched distribution in the common support.
Regression discontinuity design	Cross-sectional data Repeated cross-sectional data Panel data	Yes	Yes	Yes	Hard to find a clear-cut participation rule and a large sample around the threshold; requires an assumption about the functional form of the dependence of the outcome on the assignment criterion variable.
Estimation using marginal treatment effect (MTE)	Cross-sectional data Repeated cross-sectional data Panel data	Yes	Yes	Yes	Hard-to-find valid and powerful instrumental variables that are needed to estimate a full schedule of marginal treatment effects.

[a] Individuals select into the program because they know they will earn higher returns from the program.
[b] Individuals select into the program because their latent or forgone earnings are low at the time of program entrance.
[c] Individuals' earnings are depending on previous earnings that are low at the time of program entrance.

requires a clear-cut participation rule and a large sample around the threshold. It also requires an assumption about the functional form of the dependence of the outcome on the assignment criterion variable. It is not easy to find data that satisfy such conditions (Pirog et al. 2009). Under the previous conditions, however, it works like random assignment. A recent study by Battistin and Rettore (2008) uses this method and discusses its weaknesses and strengths. They also warn that effects are obtained only for individuals around the threshold for participation. Thus, if there is a serious heterogeneous response across the population of interest, it is hard to generalize the estimates.

Estimation Using MTE

The MTE is the mean effect of treatment on those with a particular degree of intention to participate in the program. It can vary over different participation rates of participants and nonparticipants, and can be understood as a local average treatment effect using instrumental variables. Heckman and Vytlacil (2007) analyze how we can estimate different policy parameters as weighted averages of the MTE. It is attractive in the sense that we can estimate the different policy questions only using the MTE. However, it has its own limitation because the valid and powerful instrumental variables that are needed to estimate a full schedule of marginal treatment effects are often not available to researchers (Moffitt 2008).

DISCUSSION AND CONCLUSION

Because of recessions, technological advancements, global trade, and international migration of workers, job training programs in the United States have become more inclusive, pushing beyond their initial clientele of disadvantaged workers to additionally include more mainstream segments of the labor force. WIA clearly reflects this trend in training programs. Given the expanded scope of WIA, program evaluation has become more important and far more challenging given the highly heterogeneous nature of the target population.

This chapter summarizes the previous literature related to the methodology of evaluating training programs. We begin by noting that it is necessary to understand the policy question being posed so that the evaluation design can be tailored to answer that question. If policymakers are interested in ATEs for universal programs or LATEs that occur when program benefits or enticements are made more generous, then nonexperimental methods can be appropriate. After discussing the differences in the TT, ITT, ATE, and LATE parameters, the rest of the discussion focuses on the traditional question of program evaluation which requires estimation of the TT. This question is, how does the program change the outcomes of participants compared to what they would have experienced if they had not participated? The estimated treatment effect for program participants allows policymakers to answer whether or not a program should be retained.

Despite considerable debate in the literature, random assignment experiments are still considered the gold standard for such evaluations. If random assignment is not possible, we have learned that

- comparison groups should be drawn from the same local labor markets, and
- the same instrumentation should be used to collect data from the treatment and comparison groups.

Following these practices will reduce bias in estimated treatment effects. Unfortunately, this is not enough. To provide better nonexperimental estimates of treatment effects, the comparison group members should

- have observed characteristics that span the same range of values as members of the treatment group, and
- even if the observed characteristics span the same range, the distributions of these characteristics should also be the same.

Finding a comparison group that meets all of these criteria may well be onerous. For example, large, even very large, sample sizes are normally required if one uses propensity score matching to align the range and distributions of $P(X)$ that represents observed characteristics, X's, of the treatment and comparison groups.

Even if all of the above criteria can be met, it is also critically important to understand the sources of selection bias so that an econometric

estimator can be used to correct for that particular type or combination of types of bias. Recall that there are three types of bias that typically arise in training programs:

1) self-selection by individuals who know they will earn higher incomes after participating in the program;
2) self-selection by individuals who enter a training program because their latent or forgone earnings are low at the time of program entrance; and
3) self-selection by individuals whose earnings are dependent on previous earnings that are low at the time of program entrance.

How to tease out the relative importance these sources of bias *a priori* is neither obvious nor easy. Nonetheless, it is clear that understanding how these sources of bias operate in any given evaluation of training programs is critical to choosing the most appropriate statistical methods.

Overall, we conclude that the choices made by evaluators regarding their data sources, the composition of their comparison groups, and the specification of their econometric models will have important impacts on the estimated effects of training. If a researcher cannot meet the conditions described above, estimated treatment effects from nonexperimental methods can give seriously misleading advice to policymakers. It has sometimes been argued that randomized experiments are impractical, take too long to implement, and are costly. However, the time and financial costs associated with collecting high-quality (usually longitudinal) data for nonexperiments will likely offset any extra time or financial costs of randomization. At the end of this exercise, we are forced to conclude that the logistical difficulties encountered in implementing a random assignment experiment must be weighed against the likelihood of giving bad advice to policymakers.

Notes

1. Social experiments assigning eligible applicants to the treatment and control groups to estimate the TT often have substitutes in the control group. Substitutes are individuals that have similar services from other programs even if they are assigned to the control group. When there are only no-shows, a Bloom-estimator is used to estimate the TT. When there are no-shows and substitutes, a Wald-estimator is used to estimate the TT. For further discussions of technical details and assumptions, refer to Bloom (1984) and Heckman et al. (1999, pp. 1903–1905).
2. There are four types of individuals in the program participation: 1) those who are induced to participate in the program if eligible, 2) those who will participate in the program whether or not they are eligible, 3) those who refuse to participate in the program whether or not they are eligible, 4) those who refuse to participate if eligible and want to participate if not. This second assumption for LATE eliminates the fourth type of individuals.
3. The Heckman-selection correction model is also restricted by the distribution assumption of unobservables.

References

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

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., and Jeffrey A. Smith. 2009. "What We Know about the Impacts of Workforce Investment Programs." In *Strategies for Improving the Economic Mobility of Workers: Bridging Research and Practice*, Maude Toussaint-Comeau and Bruce D. Meyer, eds. Kalamazoo, MI: Upjohn Institute for Employment Research, pp. 165–183.

Battistin, Erich, and Enrico Rettore. 2008. "Ineligibles and Eligible Non-participants as a Double Comparison Group in Regression-Discontinuity Designs." *Journal of Econometrics* 142(2): 715–730.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8(2): 225–246.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act." *Journal of Human Resources* 32(3): 549–576.

Card, David, and Daniel Sullivan. 1988. "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment." *Econometrica* 56(3): 497–530.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–1062.

———. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*. 84(1): 151–161.

Diaz, Juan J., and Sudhanshu Handa. 2006. "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program" (Programa de Educacion, Salud y Alimentacion). *Journal of Human Resources* 41(2): 319–345.

Dolton, Peter, João Pedro Azevedo, and Jeffrey Smith. 2006. *The Economic Evaluation of the New Deal for Lone Parents*. Research Report No. 356. Leeds, West Yorkshire, England: UK Department for Work and Pensions.

Doolittle, Fred C., and Linda Traeger. 1990. *Implementing the National JTPA Study.* New York: Manpower Demonstration Research Corporation.

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, David H. Greenberg, and Philip K. Robins. 1997. "Evaluating Government Programs for the Economically Disadvantaged." *Journal of Economic Literature* 35: 1809–1855.

Friedlander, Daniel, and Philip K. Robins. 1995. Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review* 85: 923–937.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robins. 2003. "A Meta-Analysis of Government-Sponsored Training Programs." *Industrial and Labor Relations Review* 57: 31–53.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84: 862–874.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra E. Todd. 1996. "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method." *Proceedings of the National Academy of Sciences* 93(23): 13416–13420.

———. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66: 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4): 605–654.

Heckman, James J., Robert J. LaLonde, and Jeffrey Smith. 1999. "The Eco-

nomics and Econometrics of Active Labor Market Policies." In *The Handbook of Labor Economics, Volume 3*, Orley Ashenfelter and David Card, eds. Amsterdam: Elsevier, pp. 1865–2097.

Heckman, James J., and Edward Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II." In *The Handbook of Econometrics, Volume 6B*, James J. Heckman and Edward E. Leamer, eds. Amsterdam: Elsevier, pp. 4875–5148.

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

———. 1995. "The Promise of Public Sector–Sponsored Training Programs." *Journal of Economic Perspectives* 9: 149–168.

Moffitt, Robert. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." Working paper. Baltimore, MD: Johns Hopkins University.

Orr, Larry L., Harold S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study.* Washington, DC: Urban Institute Press.

Pirog, Maureen A. 2007. "Trends in Public Program Evaluation in the US: Ramifications for Russia and China." *Chinese Public Affairs Quarterly* 3(1): 1–40.

Pirog, Maureen A., Anne L. Buffardi, Colleen K. Chrisinger, Pradeep Singh, and John Briney. 2009. "Are the Alternatives to Randomized Assignment Nearly as Good? Statistical Corrections to Non-randomized Evaluations." (A response to the Nathan-Hollister debate.) *Journal of Policy Analysis and Management*. 28(1): 169–172.

Perry, Charles, Bernard Anderson, Richard Rowan, and Herbert Northrup. 1975. *The Impact of Government Manpower Programs in General, and on Minorities and Women.* Philadelphia, PA: Industrial Research Unit, the Wharton School, University of Pennsylvania.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1–2): 305–353.

Wilde, Elizabeth T., and Robinson Hollister. 2007. "How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management* 26(3): 455–477.