PLEASE DO NOT CITE WITHOUT PERMISSION
COMMENTS WELCOME

Evaluating Job Training Programs: What have we learned?

Haeil Jung and Maureen A. Pirog

School of Public and Environmental Affairs
Indiana University Bloomington

Oct, 2009
First Draft: Oct, 2009

I. Introduction

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 addition to their hope for effects on unemployment, a larger skilled labor force can reduce inflationary pressures and also mitigate poverty for some groups. In the United States, training or retraining programs oftentimes have been accompanied by evaluations. This paper 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, the Manpower Development and Training Act (MDTA) (1962-1972) was followed by the Comprehensive Employment and Training Act (CETA) (1973-1982), the Job Training Partnership Act (1982-1998) and eventually the Workforce Investment Act (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. The 1982 JTPA further devolved responsibility to the states. Later WIA consolidated a number of Department of Labor (DOL) 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 ages 18 and older as well as dislocated workers and disadvantaged youth ages 14-21, whereas previous programs primarily targeted economically disadvantaged adults and youths.[1]

The early evaluations of MDTA were non-experimental (Perry et. al., 1975) and largely rudimentary (Barnow and Smith, 2007). Similarly, the CETA evaluations were non-experimental. These evaluations all relied on the CETA Longitudinal Manpower Survey which combined random samples of CETA participants with non-experimental 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, lacked local labor market data as well as recent labor market and program participation histories. Even more sophisticated matching procedures have failed to consistently replicate experimental findings (Pirog et. al., 2008; Barnow and Smith, 2007) and the absence of data on local labor markets, work and program participation choices have been found to be important in arriving at unbiased treatment effects (Card and Sullivan, 1988; Heckman and Vytlacil, 2007; Dolton et al., 2006).

The widely varying findings from the CETA evaluations led to the US DOL decision to evaluate the JTPA as a randomized experiment. Doolittle and Traeger (1990) describe the experiment which took place in 16 of over 600 local JPTA 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 (LaLonde, 1995; Friedlander, Greenberg and Robins, 1997; Heckman, LaLonde and Smith, 1999; Greenberg, Michalopoulous and Robins, 2003). Overall these authors report somewhat disappointing results. Impacts for adults are modest with more positive effects reported for women than men and negligible

[1] http://www.policyalmanac.org/economic/job_training.shtml

impacts for out-of-school youth (Greenberg, Michalopoulos, and Robins, 2006). 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 non-experimental findings. Researchers interested in government programs across the board have been investigating whether and under what circumstances carefully executed quasi-experimental methods can provide robust estimates of treatment effectiveness. In fact, the experimental JPTA 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), Smith and Todd (2005), Friedlander and Robins (1995), Heckman and Hotz (1989), Heckman, Ichamura and Todd (1997), Heckman, Ichamura, Smith and Todd (1996, 1998), Diaz and Handa (2006), 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 (DSW) and the effectiveness of the program or estimated treatment effects were radically different from those determined experimentally.

Later Heckman et al. (1999) rebutted the LaLonde (1986) study in defense of nonexperimental methods noting that social experiments only answer a limited set of policy questions and do not give a full picture of how training programs work. In particular, randomized experiments are effective if we want to know whether or not there are positive treatment effects for program participants. This is known as the effect of the Treatment on the Treated group (TT). Other approaches may be more appropriate if we want to estimate the incremental effects of expanding existing programs (the Local Average Treatment Effect (LATE)) or program effectiveness in light of considerable program dropout (the Intent to Treat (ITT)) or if we are interested in the average effective of randomly assigning a person in the population to a program (Average Treatment Effect (ATE)).

The next section of this paper provides a brief description of the types of parameters we may want to estimate in evaluating and training programs. Conventional selection bias in studies of employment and training programs are discussed in Section III followed by a discussion of “pure” selection bias and the robustness of different estimators that attempt to correct for self-section bias in Section IV. Section V discusses and concludes what we learn.

II. 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 $\Delta = Y_1 - Y_0$.

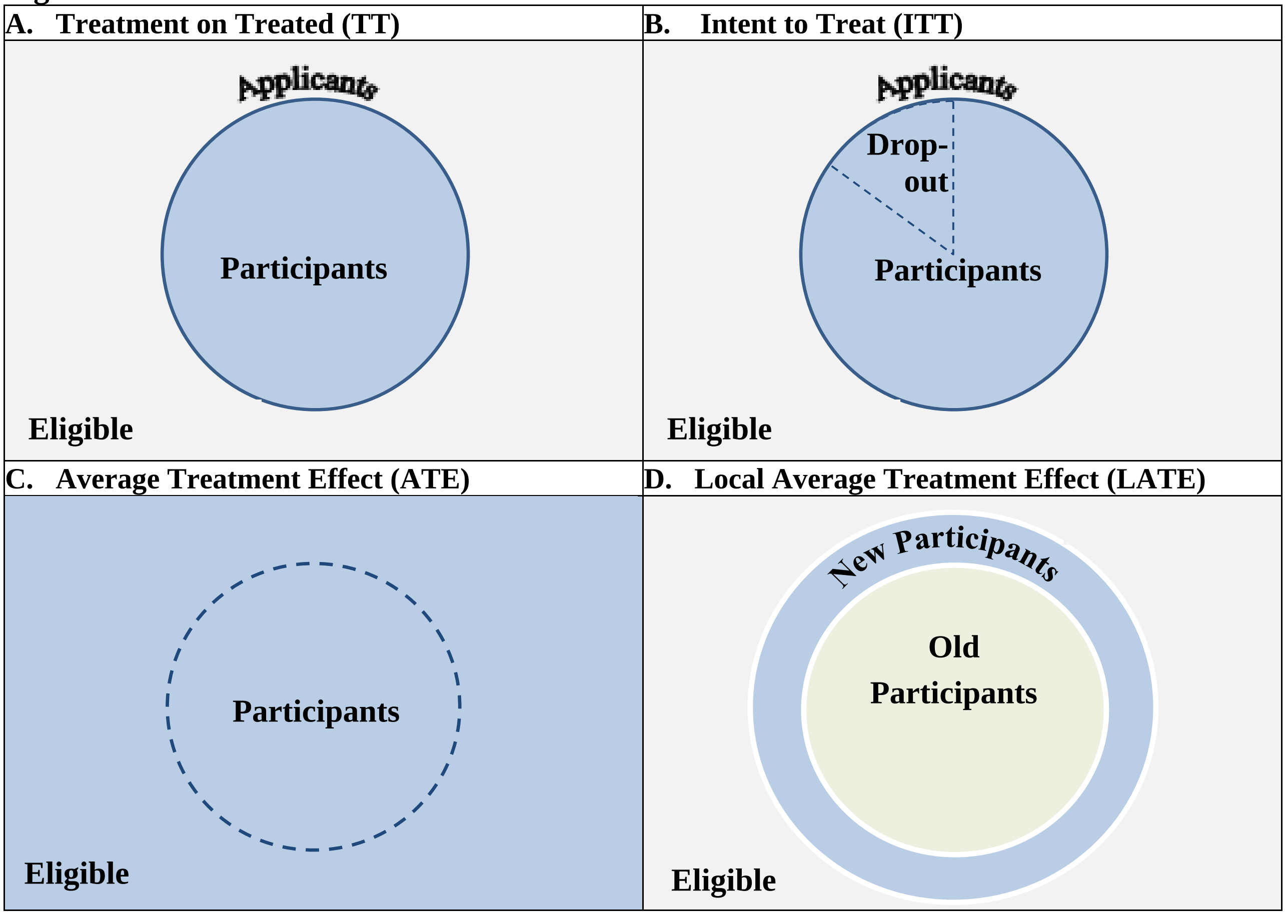
Treatment on Treated. The quantity $\Delta = Y_1 - Y_0$ is hypothetical given that a person cannot occupy two states at the same time. When we want to estimate the effect of a treatment like job training program on participants, the parameter of interest is the effect of treatment on treated (TT), depicted as follows:

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

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

In our example, TT would 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 policy makers with information on whether or not the program should be kept or eliminated. In panel A of Figure 1 (below), TT is depicted as the effect of programs on participants. Random assignment experiments are generally considered the gold standard for obtaining TT estimates.

Figure 1



A. Treatment on Treated (TT)

B. Intent to Treat (ITT)

C. Average Treatment Effect (ATE)

D. Local Average Treatment Effect (LATE)

Intent to Treat. In many social experiments, a significant fraction of the treatment group drops out of the program and does not receive the services being evaluated. This is shown in panel B of figure 1. In general, in the presence of dropping out $E(\Delta \mid X, D=1)$ cannot be identified using $E(Y_1 \mid D=1, R=1) - E(Y_0 \mid D=1, R=0)$, the experimental mean difference, where R = 1 for the treatment group and R = 0 for the control group. This estimator is sometimes called the "intent to treat." For many purposes, this is the policy relevant parameter. It is informative on how the availability of a program affects participant outcomes because attrition is a normal feature of an ongoing program.

Average Treatment Effect. The Average Treatment Effect (ATE) is the mean effect of randomly assigning a person in the eligible population of interest to the program. In panel C of Figure 1, the shaded rectangle constitutes the entire population for which the treatment effect is being estimated, regardless of whether or not they want to participate in the program. It would be particularly interesting in cases where programs are universal for a given population. For example, ATE might compare the average earnings in a world in which all eligible individuals participate in a vocational program versus the earnings in a world in which nobody participates. The ATE is shown mathematically as:

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 randomized regime. It depends on the definition of the population of interest. However, ATE is usually not relevant for policy making because there is a problem of compliance issues as well as ambiguity in the definition of the population of interest. Generally speaking, it is difficult to estimate the effect of randomly assigning a person with characteristics X to go into a program. This is because persons randomized into programs cannot be compelled to participate.

Local Average Treatment Effect. The Local Average Treatment Effect (LATE) is the effect of treatment on persons who were induced to participate by an expansion of the program or increased generosity of a program. See panel D of Figure 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:

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

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 is a one-way flow of compliance to the policy change.

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.

Discussion. Most of evaluation studies focus on estimating the Treatment on the Treated in order to answer the policy question of how the program changes the outcome of participants compared to what they would have experienced if they had not participated. When there is significant attrition by program participants, ITT is more appropriate than TT. The TT and ITT estimators provide information on whether or not programs should be maintained or eliminated.

III. Conventional Selection Bias and Lessons for Program Designs and Data Collection

Before answering the second question of what econometric methods are relevant to answer the policy question, we want to discuss the conventional bias in evaluating job training programs. The problem of job training evaluation originated from the fact that we cannot observe the counterfactual status of participants. Thus, non-experimental methods try to replicate the counterfactual status of participants using nonparticipants. LaLonde (1986) points out that using such econometric methods to answer the policy question can lead to substantial selection bias.

However, Heckman et al. (1996, 1998) argue that LaLonde's non-experimental comparison groups were constructed from various non-comparable data sources. The comparison groups were located in different labor markets from program participants and had their earnings measured in different ways. In addition, they argued that he lacked information on recent preprogram labor market outcomes which are important predictors of participation in training.

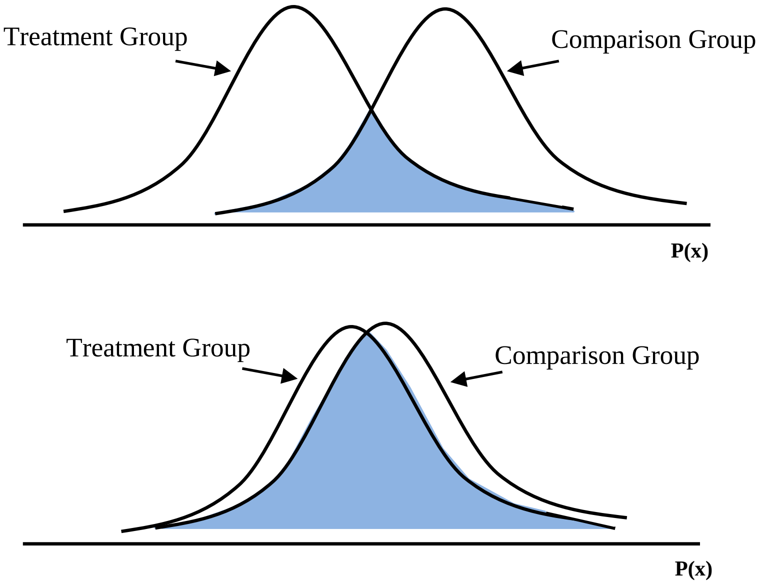
A major conclusion of the analysis of Heckman et al. (1998) is that a substantial portion of the bias and sensitivity reported by LaLonde is due to his failure to compare comparable people and to weight them appropriately. Further, using comparison group members from different labor markets and different questionnaires for data sources are also likely important sources of the selection bias measured in LaLonde's study. Overall, the available evidence indicates that simple parametric econometric models applied to bad data do not eliminate selection bias. Instead, better data, including a rich array of independent (X) variables for use in constructing the propensity score P(X) representing the probability of program participation, as well as more appropriate comparison groups, are crucial to eliminate the sensitivity problems raised in LaLonde's (1986) study.

Heckman et al. (1997) compared estimates using experimental data and different comparison groups for the treatment group to show that the bias estimates of treatment effects are not necessarily from the self-selection behaviors of treatment and comparison group members. The bias found in estimates of treatment effectiveness can be decomposed into three sources.

The first source of bias arises when there are differences in the values of the same observed characteristics in the treatment and comparison groups. This could occur, for example, if the treatment group included individuals ages 20-60 and the comparison group only included individuals ages 30-40. This is not only pertinent for individual characteristics but also for P(X), the propensity score, which is 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.

Figure 2.



The top panel of Figure 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 2.

The third source of bias in estimated treatment effects is from the "pure" self-selection on unobservables. This is the bias caused by the individuals' self-selection behavior based on information that researchers cannot observe and details of which are discussed in Section IV.

Propensity score matching can moderate bias from the first two sources of bias. Re-weighting comparison group members over the common values of observed characteristics so that the distribution of the comparison group's P(X) more closely resembles that of the treatment group can reduce bias from theses two sources.

Heckman et al. (1997) used data from a randomized control group, the no-shows from the treatment group, the eligible but non-participating group, and the Survey of Income and Program Participation (SIPP). They demonstrate that the data used to form a comparison group and the matching procedures utilized are keys to reducing the conventional bias.

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 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, randomized into the treatment group, but do not show up for 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 non-participating group as a comparison group. Individuals in the eligible but non-participating group are those who are located in the same labor market, 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 eligible nonparticipating participants (ENPs) and the treatment group members. By using propensity score matching and reweighting observations, it was 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 Treatment on the Treated.

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 Survey of Income and Program Participation (SIPP) or other large surveys. This was done by using SIPP in the study by Heckman et al. (1997). Two problems stem from 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 pre-existing 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 when using nonrandomized comparison groups. First, it is important to use the same questionnaire or data sources to obtain individual labor market outcomes and demographic information. Second, draw individuals for the treatment and comparison groups from the same local labor markets. Third, use comparison group members whose observed characteristics largely overlap with those of the participants. If this third condition does not hold, then the estimated treatment effects are only relevant for those treatment and comparison group members whose characteristics overlap.

Restricting analyses to treatment and comparison group members with similar characteristics and using propensity score matching can reduce the first and second components of the conventional selection 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 (Pirog et al., 2009). Matching is a non-parametric 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.

IV. Different Sources of "Pure" Self-Selection Bias and Empirical Methods

Different Sources of "Pure" Self-Selection Bias. There are a variety of 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 foregone earnings are low at the time of program entrance (time constant individual heterogeneity in a fixed effect model);

3) individuals' earnings are dependent on previous earnings that are low at the time of program entrance (autocorrelation between earnings in different time periods).

As noted by Heckman et al. (1997), this first type of self-selection implies that individuals with higher returns from the program are more likely to participate in training programs. The second type of self-selection behavior implies that individuals with low opportunity costs or low earnings capacity are likely to participate in training programs. The third type 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 type of self-selection results in overestimates of the effectiveness of employment and training programs while the second and third types of self-selection result in underestimates.

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-difference estimators provide consistent estimates only if self-selection bias is coming from bias source 2. The AR (1) (autoregressive of order one) regression model provides consistent estimates only when self-selection bias is coming from bias source 3 that follows the AR(1) process. The use of the instrumental variables (IV) method and the Heckman-selection correction provides consistent estimates only if bias sources 2 and 3 are present.[2] Thus, understanding which sources of bias we have in the program is critical to choose which empirical method we want to use to best answer the policy question.

In their simulation study, Heckman et al. (1999) found that cross-sectional estimation, Difference-in-Differences, and AR (1) regression estimation work relatively well when all bias sources are present, but it seems that different biases offset one another in these estimations. Also, they show that when bias source 1 is present, these estimation methods working for TT do not work for ATE. They argue that these parameters differ greatly because there is strong selection into the program of persons with high values of individual specific returns. However, it is also about how bias sources 1, 2 and 3 interact when different non-experimental methods are used to estimate TT. It seems that when all three bias sources are present, those three biases offset one another. As mentioned earlier, bias source 1 brings about overestimation of the estimate while bias sources 2 and 3 bring about underestimation of the estimate. As a result, the cross-sectional estimator provides low bias compared to the true estimate. Difference-in-differences and AR (1) regression models also provides similar low bias in estimation. Finally, IV and the Heckman self-selection correction work best when bias sources 2 and 3 are present without bias source 1 as we expected. However, when bias source 1 is present, IV and Heckman correction are the worst methods to use.

[2] The Heckman-selection correction model is also restricted by the distribution assumption of unobservables.

In summary, Difference-in-Differences and AR (1) regression estimators seem robust enough over different bias sources to estimate the Treatment on the Treated. However, it is not clear how offsetting of different bias sources works over different data and programs. Further study is necessary to examine this frontier.

V. New Non-Experimental 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 differences-in-differences extension on matching, regression discontinuity design (RDD), and the marginal treatment effect (MTE) using local instrumental variables (LIV). Our summary of these methods as well as those for "kitchen sink" regression, propensity score matching, difference-in-difference, AR(1) and IV methods is given in Table 1.

Table 1: Data, Methods, Self-Selection Behavior

Methods	Data	Consistency against Self-Selection on Unobservables			Note
		a	b	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	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 (fixed effects model)	Panel Data	No	Yes	No	Sensitive to choosing different times 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 (fixed effects model) 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
Marginal Treatment Effect using Local Instrumental Variables (LIV)	Cross-Sectional Data Repeated Cross-Sectional Data Panel Data	Yes	Yes	Yes	Hard to find local instrumental variables that satisfy LIV assumptions

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

Difference-in-Differences Extension of Matching. As mentioned earlier, propensity score matching can be used to obtain impact estimates for treatment group members whose values of observable characteristics overlap with those of comparison group members. Of course, the impact estimates will only be valid for those individuals whose values of characteristics do overlap. Within the range of overlap of observables, the "comparable" comparison group can also be re-weighted 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 et al. (1997) controls some form of selection on unobservables: it eliminates time-invariant sources of bias that may arise when program participants and non-participants are geographically mismatched or have differences in the survey questionnaire. Unlike traditional matching, this estimator requires the use of longitudinal data, which uses outcomes before and after intervention.

Regression Discontinuity Design (RDD). RDD became popular because it is easy to use and to present to a general audience. On the other hand, it requires a clear cut participation rule and a large sample around the threshold. It is not easy to find data that satisfies such conditions (Pirog et al., 2009). Under the previous two conditions, however, it works like random assignment. A recent study by Battistin and Rettore (2007) well exercises 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 Marginal Treatment Effect (MTE). MTE is the mean effect of treatment on those with a certain degree of intention to participate in the program. Marginal treatment effect can be different over participants and nonparticipants. Heckman et al. (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. A marginal treatment effect can be understood as a local average treatment effect using infinitesimal policy change as a local instrumental variable. However, it has not been well used in practice because the local instrumental variables that are needed to estimate a full schedule of marginal treatment effects are often not available to researchers.

VI. Discussion and Conclusion

Because of recessions, technological advancements, global trade, and international migration of workers, job training programs in the US 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 paper summarizes 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. When policy makers are interested in average treatment effects (ATE) for universal programs or local average treatment effects (LATE) that occur when program benefits or enticements are made more generous, then nonexperimental methods are appropriate. After discussing the differences in the TT, ITT, ATE and LATE parameters and then focus the rest of the discussion on the tradition 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 policy makers 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 charactersistics 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 may are normally required if one uses propensity score matching to align the range and distributions of P(X) 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, there are three types of bias that typically arise in training programs:

- self-selection by individuals who know they will earn higher incomes after participating in the program;
- self-selection by individuals who enter a training program because their latent or foregone earnings are low at the time of program entrance, and;
- 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 nonexperimental method.

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 you cannot meet the conditions described above, estimated treatment effects from nonexperimental methods can give seriously misleading advice to policy makers. 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 quasi-experiments 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 policy makers.

References

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

Barnow, B.S. and C. King (2005). “The Workforce Investment Act in Eight States.” Report prepared for the Department of Labor, Employment and Training Administration, February 2005, 94pps.

Barnow, B.S. and J.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: The Upjohn Institute for Employment: 165-183.

Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. Journal of Econometrics, 142(2), 715–730.

Card, D. and D. Sullivan (1988). “Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment,” *Econometrica* 56(3): 497-530.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics. 84(1), 151–161.

Diaz, J. J., & Handa, S. (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.

Doolittle, F.C. and L. Traeger. (1990). *Implementing the National JTPA Study* (New York: Manpower Demonstration Research Corporation).

Fraker, T., & Maynard, R. (1987). The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *Journal of Human Resources*, 22(2), 194–227.

Friedlander, D., Greenberg, D.H., & Robins, P.K. (1997). Evaluating government programs for the economically disadvantaged. *Journal of Economic Literature*, 35, 1809-1855.

Friedlander, D. and Robins, P.K. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review*, 85, 923-937.

Greenberg, D., Michalopoulos, C. & Robins, P.K. (2003). A meta-analysis of government-sponsored training programs. *Industrial and Labor Relations Review*, 57, 31-53.

Heckman, J.H., and Hotz, V.J. (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, J., H. Ichimura, J. Smith and P. 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 USA 93 (23): 13416-13420.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998). "Characterizing selection bias using experimental data," *Econometrica* 66: 1017-1098.

Heckman, J., H. Ichimura and P. Todd (1997). "Matching as an econometric evaluation estimator: evidence from evaluating a job training program," *Review of Economic Studies* 64 (4): 605-654.

Heckman, J, R. LaLonde, and J. Smith (1999). "The Economics and Econometrics of Active Labor Market Policies." In *The Handbook of Labor Economics Volume 3,* Eds. Orley Ashenfelter and David Card., 1865-2097. Amsterdam: Elsevier.

Heckman, J. and E. 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): 4875-5148.

Heinrich, C, Mueser, P.R. and K.R. Troske. (2008). "Workforce Investment Act Non-Experimental Net Impact Evaluation Final Report. Report by IMPAQ International, December, 92 pps.

LaLonde, R. (1986), "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review* 76 (4): 604-620.

LaLonde, R. (1995). The promise of public sector-sponsored training programs. *Journal of Economic Perspectives* 9: 149-168.

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, C., Anderson, B., Rowan, R., and H. 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).

Schochet, P. and J.A. Burghardt (2008). "Do Job Corps performance measures track program impacts?" *Journal of Policy Analysis and Management* 27(3): 556-578.

Schochet, P. Burghardt, J. and S. Glazerman (2001). National Job Corps study: The impacts of Job Corps on participants' employment and related outcomes. Princeton, NJ: Mathematica Policy Research, Inc.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1–2), 305-353.

Wilde, E. T., & Hollister, R. (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.