# Lessons from the United States: Evaluating Employment Services is Important but Neither Easy Nor Cheap

**Sheena McConnell,* Peter Schochet,* and Alberto Martini****

***Mathematica Policy Research and ****Universitá del Piemonte Orientale**

**November 16, 2009**

## I. Introduction

Title I of the 1998 Workforce Investment Act (WIA) is the largest source of federally-funded employment services in the U.S. Its purpose is to increase the employment, job retention, and earnings of its participants. WIA funds the dislocated workers, low-income disadvantaged adults, and youth programs as well as Job Corps—a primarily residential training program for disadvantaged youth—and specific programs for Native Americans, migrant and seasonal farm-workers, and veterans. In fiscal year 2008, $4.5 billion was spent on WIA programs.

The European Social Fund (ESF) provides funding to promote employment in the 27 member states of the European Union (EU). Over the seven years of the current funding cycle (2007-2013), ESF will fund $114 billion in services, accounting for about 10 percent of the total EU budget. ESF has many important similarities to WIA. They are both large and decentralized. WIA allows state and local workforce investment areas to shape their programs. ESF funds are allocated to member states, which funnel the funds to one or more operational programs, which in turn have the ability to fund a wide variety of programs and services at the local level. A similar wide range of services are funded by both WIA and ESF including counseling, job search assistance, basic education, vocational training, support services, retention services, and

entrepreneurial assistance. Services under both WIA and ESF are provided by both government and nongovernment agencies including small community-based organizations.

Given the considerable amount spent on employment services in both the U.S. and Europe, policymakers, participants, taxpayers, and program administrators on both continents want to know which services are effective. For more than three decades, the U.S. Department of Labor (DOL) has invested heavily in conducting rigorous impact evaluations of its employment programs. In the past decade alone, it has funded experimental evaluations of Job Corps, approaches to administering training vouchers, entrepreneurial services, and prisoner reentry programs.[1] It has also funded nonexperimental evaluations of the WIA adult and dislocated worker programs and the Trade and Adjustment Assistance (TAA) program.[2] Recently, DOL funded a nationally-representative experimental evaluation of the WIA adult, dislocated worker, and youth programs that is in its design phase.[3] Although the EU does sponsor evaluations of its operational programs, much less emphasis is placed on impact evaluations. And as noted by Greenberg and Shroder (2004), very few experimental evaluations have been conducted on employment programs outside the U.S.

The purpose of this paper is to inform EU officials about some of the lessons learned from conducting impact evaluations of employment programs in the U.S. It begins by describing the role evaluations have played in decisions about employment policy and programs in the U.S. It then discusses the three key main steps in any evaluation: (1) choosing the policy-relevant

[1]Schochet et al. (2008), McConnell et al. (2006), and Benus et al. (2008). The experimental evaluation of prisoner reentry programs is being conducted by Social Policy Research Associates and MDRC.

[2]Heinrich et al. (2008). The nonexperimental evaluation of TAA is being conducted by Social Policy Research Associates and Mathematica Policy Research.

[3]DOL contracted with Mathematica Policy Research to conduct the National Evaluation of WIA.

evaluation questions; (2) choosing the best design; and (3) collecting data. The paper concludes with a summary of our recommendations.

## II. Evaluation Can Affect Policy and Programmatic Decisions

Information on the effectiveness of employment services is needed for three main reasons. First, as a considerable amount of government funds is invested in employment services, taxpayers need information on the investment's return. Second, most people in need of employment services are vulnerable and disadvantaged, so it is particularly important that the services offered to them are helpful. Third, a workforce with the skills required by employers is critical for the continued growth of the economy. As discussed below, evidence on service effectiveness has led U.S. Congress to fund new programs, expand existing programs, and reduce funding for others. Evaluation findings have also been used by program administrators to improve programs.

An example of an evaluation that led to a new program is the New Jersey UI Reemployment Demonstration sponsored by DOL in the 1980s (Corson et al. 1989). The demonstration involved targeting UI recipients who were likely to have difficulty becoming employed and randomly assigning them to four groups: (1) a treatment group that received job search assistance; (2) a treatment group that received job search assistance and training or relocation assistance; (3) a treatment group that received job search assistance with a cash bonus for early reemployment; and (4) a control group that received no services or bonuses. The evaluation of the demonstration found that compared to the control, all three treatments led to increased earnings and employment and to benefits to society and claimants that outweighed their costs. As a result of this evaluation, in 1993 Congress required all states to establish a Worker Profiling and Reemployment Services (WPRS) system which: (1) identifies UI recipients who are likely to

exhaust their benefits before they find employment, and (2) requires these UI recipients to receive reemployment services (Reich 1997).

Another example of the funding of a program based on research evidence occurred at about the same time. In the late 1980s, DOL funded the UI Self-Employment Demonstration in Massachusetts and Washington to help UI recipients start their own businesses by offering financial assistance and workshops on issues related to business start-up. The generally positive findings from an evaluation of these demonstrations (Benus et al. 1995) led to the 1993 legislation to establish the Self-Employment Assistance program for UI recipients.

Congress has also expanded funding for existing programs found to be effective. A nonexperimental evaluation of Job Corps conducted in the 1970s found that the program increased employment and earnings and was cost-effective for society and for the participants (Mallar et al. 1982). Following these findings, Congress increased funding for Job Corps.

While program designers and administrators nearly always ardently believe their programs are effective, rigorous evaluations have sometimes found that they are wrong. For example, an experimental evaluation of the youth program under Job Training Partnership Act (JTPA)—the predecessor to WIA—found that overall the program had no significant impact on earnings for youth and may even have had negative impacts on male youth who had been arrested prior to random assignment (Bloom et al. 1997). The findings from this study led Congress to reduce funding for the JTPA youth program and subsequently require major changes in the youth program when JTPA was replaced with WIA.

Evaluation findings have also been used by program administrators to improve programs. The Job Corps program examined the services it provided Hispanic youth after the National Job Corps Study found that the program did not increase earnings for this population of youth (Schochet et al. 2008). A study of different approaches to providing training vouchers, or

individual training accounts, found that, contrary to the fears of program staff, the recipients of the vouchers made similar training and employment choices irrespective of whether they were required to be counseled by an employment counselor at the one-stop career center (McConnell et al. 2006). This has direct implications for the administration of vouchers.

## III. Careful Development of Evaluation Questions

The first step in any evaluation is to carefully specify what policymakers want to learn from the evaluation. Although most evaluations involve considerable exploratory analysis, an evaluation can usually only address a few questions *rigorously*. Hence, it is important to design the evaluation so that the questions it does ask are the ones that are most helpful to policy makers.

### 1. Evaluating the Entire Program or Program Components

In many cases, the most policy relevant question is not whether an entire program is effective but rather which program components are effective. Evaluating an entire program is appropriate if policymakers are considering *whether* to fund the program or the program consists of only a few key components. The U.S. Congress has asked for evaluations of entire programs, such as the Job Corps and JTPA programs. When the programs are large and comprised of many diverse components, such as WIA and ESF, policymakers are unlikely to stop funding the entire program, but do want to know which components of the program are effective. In these cases, evaluating specific program components is more informative. For example, DOL's nonexperimental WIA evaluation did not attempt to evaluate the entire program but focused on evaluating just the adult and dislocated worker programs, which are large but not the only programs funded by WIA (Heinrich et al. 2008).

If individuals choose which service component to receive, care must be taken in interpreting impacts by program component, however. The impact estimates pertain only to the people who chose that component and not to all study participants. During the design phase of the National Job Corps Study, program administrators expressed interest in not only the effectiveness of the entire program but also in the effectiveness of the nonresidential component of the program. Most participating youth live at a Job Corps center, but some youth choose to live at home and commute to the center (and are referred to as nonresidents). The study found that both the residential and nonresidential components of Job Corps had positive impacts (Schochet et al. 2008; Schochet and Burghardt 2007). However, as nonresidential and residential youth differ, it cannot be concluded that the nonresidential program is effective for those youth who chose the residential component.

## 2. Determining for Whom the Program is Effective

Some programs and policies are effective for some people but not for others. In the design phase of the evaluation, policymakers should specify which target populations are of policy interest. The JTPA evaluation, for example, focused on four groups with different employment needs—adult women, adult men, young women, and young men. The National Job Corps Study estimated the impacts for youth in three different age groups—16-17, 18-19, and 20-24. The choice of estimating impacts for youth by age was motivated by conversations with Job Corps staff who viewed younger participants much more difficult to serve than the older youth.

It is important to decide on the target populations that are of policy interest *prior* to conducting the evaluation for two reasons. First, the size of the target populations will affect the required sample size. Estimating impacts for subgroups requires a larger sample; and the required sample is larger the smaller the subgroup. Second, it avoids the temptation to estimate impacts for numerous subgroups and interpret any significant impact as a true program effect. If

a large number of subgroup impacts are estimated, the estimate of the program impact for some subgroup is likely to be significantly positive *by chance* and may not reflect a true positive program impact (Schochet 2009a). Statistical adjustments can be made to account for estimating multiple subgroups but these adjustments result in a loss of statistical power, with the loss increasing with the number of subgroups.

## 3. Determining the Counterfactual

Perhaps the greatest challenge in designing evaluation questions is to determine the counterfactual—the scenario against which the intervention is tested. Evaluations in which the counterfactual is the absence of all employment services are rare or nonexistent. WIA is not the only source of employment services in the U.S.—people can receive training at a community college funded by a Pell grant, for example. Similarly, the ESF is not the only source of employment services in European countries. Hence, if people do not receive employment services from WIA or the ESF, they may receive services from other sources. In the National JTPA Study, for example, about 40 percent of the control group received some employment services not funded by JTPA (Bloom et al. 1997).

It may be that a counterfactual in which other employment services can be received is the more appropriate one. Such an evaluation provides policymakers information about the effectiveness of additional WIA funding in the real world, a world in which other services exist. The estimated impact of employment services in these cases is likely to be smaller because it is based on the impact of *additional* services not the impact of receiving services versus no services. Hence the estimated impact of the JTPA services was not the impact of receiving the services versus no services, but the impact of more treatment group members receiving services. Correct interpretation of the impacts requires information about the receipt of services by both the treatment and control/comparison groups.

## III. Impact Evaluation Design: Constructing a Comparison Group

An ideal evaluation of employment services would compare the outcomes of people who receive the services with the outcomes of the same people who do not. As this is impossible, the challenge is to choose another set of people—a comparison group—who are as similar as possible to the people who receive the services. Under an experiment, this comparison group is determined randomly and is referred to as a control group. In nonexperimental evaluations, other approaches are used to construct a comparison group. Below, we describe the considerations in choosing an evaluation design.

### 1. Experiments

Experiments involve randomly assigning individuals to two or more groups, with each group offered a different set of services. When implemented carefully, random assignment creates groups of individuals that, on average, have identical observable and nonobservable characteristics prior to the intervention, differing only in the program services they are offered. As a result, the great advantage of experimental designs is that differences in average outcomes between the groups can be *causally* attributed to the specific interventions under investigation. Under other designs, there is always a concern that the differences in outcomes are a result of differences in the underlying characteristics between the group receiving the intervention and the comparison group (or between the groups receiving different interventions).

The fundamental and unavoidable challenge of experiments is that they require that some people are offered more or different services than others. This may be politically challenging and often is resisted by program administrators. Yet, numerous social service experiments have been

conducted successfully in the U.S. and developing countries.[4] To be successful, the evaluator needs to obtain political support for the study and minimize the burden on the program and study participants.

Experiments are often more acceptable politically and to program administrators when they are used to evaluate a demonstration or a pilot of an intervention rather than an existing program. In these cases, control group members receive the services they would in the absence of the experiment and treatment group members receive *more* services. DOL has supported numerous experimental evaluations of demonstrations including the National Supported Work (NSW) Demonstration (Maynard et al. 1979), a series of UI job search assistance and bonus experiments (Woodbury and Spiegelman 1987; Corson et al. 1989; Corson et al. 1992; and Spiegelman et al. 1992), the Individual Training Account Experiment (McConnell et al. 2006), and the recent evaluation of Project GATE (Growing America Through Entrepreneurship) (Benus et al. 2008).

If the roll out of new programs takes place over time, an experiment can be conducted if the order at which potential sites receives the program is determined randomly. In this case early implementation sites are the treatment sites and the later implementation sites are the control sites, at least until program implementation. This design requires a large number of sites to ensure enough statistical power due to the clustering of individuals within sites. While we do not know of an example of this design in evaluating employment service, it has been used extensively in education evaluations—schools have been randomly assigned to either receive funding for an intervention immediately or receive future funding for the intervention (see, for example, Glazerman et al. 2007).

[4] The Poverty Action Lab at MIT (www.povertyactionlab.org/papers) has conducted numerous experiments in developing countries.

Evaluating existing programs experimentally is more difficult because the experiments lead to some people not participating or receiving fewer services than they would in the absence of the evaluation. The control group may also lead to empty slots at the program. The best conditions for an experiment occur when there is excess demand for the program. With a surplus of people wanting to participate in the program, the existence of a control group could affect who receives the intervention but not the number of people who received the intervention, and thus, the program would not suffer from empty slots. This was the case in an evaluation of Upward Bound, a program to assist disadvantaged youth to prepare for, enter, and succeed in college (Seftor et al. 2009). The program recruited enough students that the treatment group could fill all programs slots and the control group was placed on a waiting list. If any openings in the program occurred, they were filled by selecting students randomly off the waiting list.

Experiments are also more acceptable when the research groups are offered different treatments, so that all study participants receive some services. In an evaluation of individual training accounts, or training vouchers, people who were found eligible for the vouchers were assigned to three groups that varied in the extent to which counseling was required and the role the counselor played in setting the amount of the voucher (McConnell et al. 2006). No one was denied a voucher, and anyone could receive counseling by requesting it, even if they were in the group for which counseling was not mandatory.

Randomized encouragement is another experimental evaluation approach that does not involve denial of services. Under this design, both treatment and control group members can receive the intervention, but the treatment group is given additional encouragement to receive the intervention. This encouragement can take the form of information, financial, or other incentives, but the encouragement must not directly affect the outcomes of interest. While we know of no study of employment services that have used randomized encouragement, it has been used to

evaluate the effectiveness of health interventions such as the influenza vaccine (Hirano et al. 2000).

Cooperation from program staff is a prerequisite for a careful implementation of an experiment, and so evaluators need to obtain support for the study from program staff at all levels, and then train and monitor them. Most program staff will support an evaluation if they understand that the findings will be used to inform the development of effective employment services. Staff must also understand the rationale behind an experiment and the drawbacks of alternative designs.

Evaluators should work with program staff to find ways to reduce the burden of the experiment to the program and participants. The web-based random assignment systems used in recent experimental evaluations (such as the evaluation of a relationship-skills program, Building Strong Families) mean that program staff can learn the research assignment of a program applicant almost instantaneously rather than having to wait a few days before knowing the assignment. Another way to reduce the burden on program and participants is to have small control groups. The National Job Corps Study, for example, had control groups that were only 7 percent of all eligible Job Corps applicants (Schochet et al. 2008).

It can be challenging to estimate the impact of service components in an experiment because of a lack of information on which services the control group would receive. It is sometimes possible to ask program staff to predict prior to random assignment which services each sample member would receive if they were assigned to the treatment group. If the predictions are accurate, an estimate of the impact can be obtained by comparing the outcomes of those members of the treatment and control groups who are predicted to receive the services. This approach was used successfully in the National Job Corps Study to estimate the separate impacts of the residential and nonresidential services (Schochet et al. 2008).

A major drawback of experiments is that they cannot provide policymakers quick answers. The National Job Corps Study began in 1993; the last evaluation report was published over a decade later in 2006. It takes considerable time for an experiment to provide findings for three reasons. First, it takes some time to obtain political and program support for the evaluation. Second, it takes time for enough eligible people to request the services and be randomly assigned. Typical sample intake periods are one or two years. Third, as many programs are designed to have long-term effects, follow-up data collection needs to occur for a lengthy period after participants enter the program. The total follow-up period for the participants of in the Job Corps study was 48 months for survey data and 8 to 10 years for administrative data.

It is often said that experiments are more expensive than other evaluation designs (Levitan 1992). Some costs that are incurred for experiments but not nonexperimenal evaluations include recruiting sites, training staff, conducting random assignment, and monitoring. In practice, experiments can be very expensive—some have cost millions of dollars. However, it is not clear that this is because they are experiments or because experiments often involve surveys while many nonexperimental evaluations rely only on less costly administrative data. Yet, the type of data collected is unrelated to the design—experiments can be conducted with administrative data and nonexperimental evaluations can include survey data collection. Rigorous nonexperimental evaluations require more detailed baseline data. More research is needed to compare the costs of experimental and nonexperimental designs, holding constant data collection costs.

## 2. Nonexperimental Designs

It is not always possible to conduct experiments. Experiments are typically not feasible for evaluating entitlement programs (because program services cannot be denied to eligible program applicants, thereby making it impossible to create control groups), and may not be appropriate

for evaluating existing employment-related programs that are undersubscribed. It may also not be feasible to create control groups if there is no way of restricting program services (for example, re-employment services that are obtained by computer in one's home). Furthermore, experiments cannot be conducted using retrospective treatment samples (that is, past program participants who are identified using administrative program data) or treatment samples selected using secondary data (for example, using large national survey data). Finally, even if random assignment is feasible, program staff may refuse to participate in the experiment, because of ethical concerns about restricting services to program applicants and the extra burden associated with implementing random assignment procedures (such as obtaining study consent forms, collecting additional customer information that is required for random assignment, notifying customers about random assignment results, and so on).

Consequently, researchers often use nonexperimental methods to estimate program impacts. In this section, we briefly discuss key features of two nonexperimental methods that are becoming increasingly popular for evaluating employment and training programs: (1) regression discontinuity (RD) methods, and (2) propensity score matching methods. We do not discuss pre-post designs where the outcomes of program participants are compared before and after program participation, because of obvious confounding factors that could bias the impact estimates (such as changes in economic conditions or participant's health status). In addition, we do not discuss instrumental variables (IV) methods, because it is often difficult to find defensible instruments that are strongly correlated with the decision to participate in an employment or training program, but that are uncorrelated with the disturbance terms that influence key postprogram

outcomes (such as employment and earnings).[5] We conclude this section with a discussion of the available evidence on the validity of these methods.

## 1. Regression Discontinuity Designs

RD designs are increasingly used by researchers to obtain unbiased estimates of intervention effects in the social policy area (see, for example, Cook 2008, Schochet 2009b, Imbens and Lemieux 2008 for reviews). These designs are applicable when a continuous "scoring" rule is used to assign the program, policy, or other intervention to people or other study units (for example, one-stop career centers). People or units with scores above a pre-set cutoff value are assigned to the treatment group and units with scores below the cutoff value are assigned to the comparison group, or vice versa. For example, Black et al (2007) estimated the impacts of the WPRS system in the state of Kentucky using the rule that UI recipients are required to receive reemployment services if their model-based UI profiling scores are larger than a cutoff value. As another example, the effects of providing competitive grants to workforce investment areas for one-stop career center innovations could be estimated using grant application scores and collecting data on a random sample of workers in both the winning and losing grantee sites.

Under an RD design, the effect of an intervention can be estimated as the difference in mean outcomes between treatment and comparison group units, adjusting statistically for the relationship between the outcomes and the variable used to assign people or other units to the intervention, typically referred to as the "forcing" variable. A regression line (or curve) is fit for the treatment group and similarly for the comparison group, and the difference in average outcomes between these lines at the cutoff value of the forcing variable is the estimate of the

[5] Though IV methods are important in experiments when members of the treatment group do not receive the treatment or when control group members receive the intervention being tested (Heckman et al. 1998).

effect of the intervention; an impact occurs if there is a "discontinuity" between the two regression lines at the cutoff.

RD designs generate unbiased estimates of the effect of an intervention if the relationship between the outcome and forcing variable can be modeled correctly (using parametric, local linear, or other nonparametric methods, and using appropriate score bandwidths), and the forcing variable was not systematically manipulated to influence treatment assignments. Furthermore, the forcing variable must be reasonably continuous, and should not be binary (such as gender) or categorical with no natural ordering (like race). In addition, the cutoff value for the forcing variable must not be used to assign people or other units to interventions other than the one being tested. This requirement is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions.

Well-designed RD designs can yield unbiased impact estimates, and may be easier to sell to program staff and participants than experimental designs because treatment assignments are determined by rules developed by program staff or policymakers rather than randomly. However, RD designs cannot necessarily be viewed as a substitute for experimental designs. Sample sizes typically need to be about three to four times *larger* under RD than experimental designs to achieve impact estimates with the same levels of precision (Schochet 2009b). The estimate of the impact under the RD design typically pertains to a narrower population (those with scores near the cutoff) than under an experimental design (those with a broader range of scores). Furthermore, the RD design requires critical modeling assumptions that are not required under the experimental design.

## 2. Propensity Score Matching Methods

Propensity score methods involve matching program participants to a comparison sample of people using available data on demographic characteristics, earnings histories, and local area characteristics. The best data source for selecting comparison samples will depend on the specific application and study research questions, but options often include (1) administrative records (such as UI claims data); (2) program data on ineligible program applicants or eligible applicants who decide not to participate in the studied program; (3) program data for workers who are eligible for a related, but less-intensive program to the one under investigation; and (4) national surveys that cover the same time period as the treatment sample data and that include comparable matching variables. In all cases, the outcomes of the comparison group are intended to represent the outcomes of the treatment group had they not received the program services under investigation. The relevant counterfactual for the study, however, will often depend on the specific data source.

Under comparison-group designs, assumptions and statistical models must eliminate differences between the treatment and comparison group samples that could result from sources other than the intervention. If these efforts are successful, remaining differences can be attributed to the intervention, possibly with some measure of statistical confidence. However, if sources of unmeasured differences exist, this approach could produce impact estimates that suffer from sample selection biases.

Rosenbaum and Rubin (1983) developed a statistical procedure—propensity scoring—to select a matched comparison group. A propensity score is the probability that a worker with a given set of characteristics receives the treatment. Rosenbaum and Rubin (1983) proved the key result that individuals with the same propensity score will also have the same distribution of the matching variables.

Several methods can be used to perform the matching, such as nearest neighbor, caliper, or kernel methods. Smith and Todd (2005) and Imbens and Wooldridge (2008) conclude that with sufficient sample overlap in the propensity scores and well-balanced matching variable distributions, impact estimates are relatively insensitive to the choice of matching methods. It is critical that the adequacy of the matching process be assessed, for example, by comparing the distribution of the matching variables and propensity scores of treatment and selected comparison group members within propensity score classes.

Several recent large-scale evaluations of employment and training programs have used propensity score matching methods that were structured to satisfy the conditions discussed above for obtaining credible impact estimates. For example, Heinrich et al. (2008) estimated the impacts of WIA on the (1) combined effects of core and intensive services relative to no WIA services; and (2) the incremental effect of WIA-funded training relative to WIA participants who did not receive training. The comparison group for their analysis was drawn from UI claimants or from U.S. Employment Service (ES) participants in the 12 study states. The data used for propensity score matching were obtained from UI claims data, ES data, and WIA program data, and included employment histories, labor force status at the time of program entry, demographic characteristics (gender, age, race and ethnicity, education attained, veteran's status, and welfare receipt), and local labor market characteristics.

As another example, a national evaluation of the TAA program is employing a propensity score matching design (Social Policy Research et al. 2004). The large TAA program provides training, extended UI benefits, and other employment-related services to workers who are displaced from their jobs due to trade-related reasons. A random assignment design was not feasible for the evaluation—because TAA services cannot be denied to eligible workers and so under program rules, it would not be possible to construct a control group. Furthermore, it was

not feasible to randomly assign participants to different service groups, because TAA services are voluntary and are tailored to meet the needs of individual clients. Consequently, the evaluation is employing a comparison group design to obtain estimated impacts, where the comparison group was selected using UI claims data from the 26 study states, and using similar matching variables to those described above for the Heinrich et al. study.

## 3. The Validity of Nonexperimental Methods

There is a long-standing debate in the literature about whether social programs can be reliably evaluated using nonexperimental methods. To investigate their validity, data from experiments have been used to try to replicate the experimental estimates—the "gold-standard" estimates—using nonexperimental methods.

In an influential study, LaLonde (1986) found that the impact results from the experimental NSW Demonstration could not be replicated using a comparison group design. He estimated program impacts using a number of standard nonexperimental evaluation econometric methods, including simple regression methods, difference-in-difference methods, instrumental variable procedures, and the two-step estimator of Heckman (1979), and found that the alternative estimators produced very different impact results. Fraker and Maynard (1987) came to similarly pessimistic conclusions using a slightly different comparison sample. Similarly, Peikes et al. (2008) found that matching methods produced incorrect impact estimates when compared with a randomized design for the State Partnership Initiative employment promotion program.

Using the same data as LaLonde, however, Heckman and Hotz (1989) used a broader set of specification tests to help select among nonexperimental estimators, and found that their tests could exclude those estimators that produced impact results that differed from the experimental ones. A key specification test that they used was that a credible estimator should yield no

differences between the treatment and comparison groups in their mean outcomes pertaining to the *pre-intervention* period.

In an influential study, Deheija and Wahba (1999) reexamined LaLonde's data using propensity scoring—to find matched comparison group members for the NSW treatment group; their resulting impact estimates were similar to the experimental ones. A key contribution of their study was the careful use of model specification tests that yielded treatment and comparison groups with similar distributions of the matching variables and propensity scores. Mueser el al. (2007) also concluded using JTPA data that matching methods may be effective in evaluating job training programs. Smith and Todd (2005a and 2005b) cautioned, however, that the Deheija and Wahba results are not robust to alternative analysis samples and matching variables included in their models.

Glazerman et al. (2003) surveyed sixteen studies that each used nonexperimental methods to try to replicate impact findings from a social experiment. Their systematic review was intended to shed light on the conditions under which nonexperimental methods most closely approximate impact results from well-designed and well-executed experimental studies. They found that nonexperimental methods occasionally replicate the findings from experimental impact evaluations, but in ways that are not easy to predict. However, they identified several factors that lead to more successful replications. These factors, which are similar to the ones that Heckman et al. (1997, 1998) found in trying to replicate experimental results from the National JTPA Study, are as follows: (1) the data should include a rich set of matching variables relevant to modeling the program participation decision, and in particular, preprogram earnings histories; (2) the same data sources should be used for the treatment and comparison groups; and (3) the treatment and comparison samples should be from the same geographic areas. Bloom et al. (2005) identify

similar criteria for increasing the chances that nonexperimental methods can produce credible impact estimates.

Studies have shown that the RD approach has promise for evaluating employment and training programs when experimental methods are not viable. Cook et al. (2008) provide empirical evidence that impact estimates based on RD designs can replicate experimental estimates in a range of settings. The advantage of the RD approach relative to the propensity score comparison group approach is that the selection rule for receiving the treatment is *fully* known under the RD approach and can be used to obtain unbiased estimates if the outcome-score relationship can be modeled correctly. In contrast, the propensity score approach assumes that the program participation decision can be adequately modeled using observable baseline data, which is typically very difficult to test, suggesting that one never knows for sure whether unobservable factors bias the impact findings.

## IV. Collecting the Necessary Data

Data on outcomes need to be collected for both the treatment and control/comparison groups. These data can be obtained from surveys or from administrative records. Much more complete and detailed information can be obtained from surveys than is typically available from administrative databases. Surveys can also collect details that may suggest a job's quality, such as the receipt of fringe benefits, union status, and wage rates. Data on criminal activity, substance abuse, and receipt of a wide range of services are often not available from sources other than surveys.

On the other hand, administrative data do not suffer from recall error or nonresponse bias. And because they are much cheaper than survey data to collect, they can provide data on many more study participants over a longer period of time. However, they are more limited in the

variables they include and may miss some jobs. In the U.S., state UI agencies collect quarterly earnings from all people covered by UI and these data are often used to evaluate employment programs. These data, however, do not cover federal employment, jobs not covered by UI (such as self-employment or agricultural jobs), or any jobs that employees or employers do not want reported. Hotz and Scholz (2001) estimate that these data may understate employment by about 13 percent. In the U.S., Social Security data are another potential source of administrative data on earnings, which are sometimes used in impact evaluations. These data do cover federal and self-employed workers and cover all states, but are annual rather than quarterly.

Baseline data—or data collected prior to the receipt of the intervention—are essential for implementing nonexperimental designs. For example, detailed data on the baseline characteristics of both participants and nonparticipants is required to construct a matched comparison group design. While baseline data are not essential for experiments, they are useful for ensuring that random assignment created research groups with similar baseline characteristics. Irrespective of the design, baseline data are also necessary for defining subgroups of interest, adjusting for baseline differences in the treatment and control/comparison groups due to sampling error, and testing and adjusting for survey nonresponse bias. Finally, baseline data on program participants are useful for describing those who receive the intervention.

Baseline data can be collected from administrative records, application forms, or surveys. In some studies, study specific forms are administered to study participants. In experiments, participants typically need to be administered a consent form prior to random assignment. A form requesting additional baseline and contact information (to aid follow-up of the participant) can be administered at the same time.

Data on the receipt of services is needed to understand differences between the receipt of services by the treatment and control/comparison groups and hence the interventions and

counterfactuals being tested. Program participants will likely vary in the intensity of the services received. And, as discussed above, study participants—in both the treatment and control/comparison groups—may also receive services from other programs.

The program is likely to be able to provide detailed and accurate data on service receipt among program participants. (Program administrators may need assistance in collecting these data.) However, these data are typically not available for the control/comparison group. Data on the service receipt of the control/comparison group are often unavailable from administrative records and hence need to be collected using a survey.

Correctly interpreting estimates of program effectiveness requires an understanding of the program as it is actually implemented, rather than how it is designed. This understanding requires an "implementation" or "process" analysis, which requires collecting detailed information on the program from program manuals, training materials, and budgets; interviewing both managers and front-line program staff; observing service provision; and talking with participants. If an impact is found, this information is important for replication. If no impact is found, or the impact is smaller than expected, this information will allow the evaluator to determine whether this was because the intervention was not implemented, because it was not implemented as designed, or because it was ineffective.

Finally, information on the cost of the program can be used to interpret the magnitude of a program impact and to inform others who may be considering replications of the program. A program may have positive impacts on earnings, but may not be cost-effective if its costs are high. Conversely, a low-cost intervention may be cost-effective even if it has modest impacts. With cost data, a benefit-cost analysis can be conducted that compares the cost of the intervention with the monetary value of the benefits of the employment services. The largest benefit of employment services is typically the increase in participant's earnings after they leave

the program, which is already measured in dollars. Other potential benefits from participation in employment services, such as any reduction in public assistance use or crime, can be valued in dollars (see, for example, McConnell and Glazerman 2001). In evaluations where it is difficult to place dollar values on program benefits (so that benefit-cost analyses are not possible), some researchers instead conduct cost-effectiveness studies where they compare the key impact estimates with the per-participant program costs. Benefits and costs are examined from different perspectives—usually society as a whole, taxpayers, the program's funder, and participants. Benefit-cost analysis is useful for comparing interventions to each other, and for identifying those interventions that improve participants' outcomes most efficiently.

## V. Recommendations

First, we urge the EU to invest in data collection for evaluating program impacts. As well as collecting baseline and outcomes data, data should also be collected on costs, the implementation of the program, and the receipt of services by members of both treatment and control/comparison groups.

Second, we recommend that the EU consider conducting experiments. While not always possible, there are many situations in which they can be done and can yield rigorous findings. They need not be large or expensive.

Third, if experiments are not feasible, we recommend that rigorous nonexperimental methods be used, such as regression discontinuity or propensity score matching methods. However, it is critical that these methods be carefully selected and applied to ensure that potential sample selection biases can be overcome to yield credible impact estimates.

Finally, we recommend that the EU invest in conducting rigorous impact evaluation, whether experimental or not. The stakes for the taxpayers, the participants, and the health of the economy are too high for labor market policies not to be based on strong evidence.

**References**


Benus, Jacob, Terry Johnson, Michelle Wood, Neelima Grover, and Ted Shen. "Self-Employment Programs: A New Reemployment Strategy: Final Impact Analysis of the Washington and Massachusetts Self-Employment Demonstrations." Unemployment Insurance Occasional Paper no 95-4. Washington, DC: U.S. Department of Labor, 1995.

Benus, Jacob, Sheena McConnell, Jeanne Bellotti and others. "Growing America Through Entrepreneurship: Findings from the Evaluation of Project GATE." Report prepared for the U.S. Department of Labor, Employment and Training Administration (May 2008).

Black, Dan, Jose Galdo and Jeffrey Smith. 2007. "Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Design." *American Economic Review Papers and Proceedings* 97(2): 104-107.

Bloom, H.S., L.L. Orr, S.H. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos. (1997). "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources,* vol. 32, no. 3.

Bloom, Howard S., Charles Michaelopoulos and Carolyn J. Hill, "Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects," in Howard S. Bloom (Ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*, New York: Russell Sage (2005), 173-235.

Cook, T. (2008). "Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics." *Journal of Econometrics*, 142(2), 636-654;

Cook, T.D., Shadish, W.R., and V.C. Wong (2008). "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 727(4): 724-750.

Corson, Walter, Paul Decker, Shari Dunstan, and Anne Gordon. "The New Jersey Unemployment Insurance Reemployment Demonstration Project." Report prepared for the U.S. Department of Labor, Employment and Training Administration (April 1989).

Corson, Walter, Paul Decker, Shari Dunstan, and Stuart Kerachsky. "Pennsylvania Reemployment Bonus Demonstration Project." Unemployment Insurance Occasional Paper 92-1. Washington, DC: U.S. Department of Labor, 1992

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94 (December 1999), 1053-1062.

Fraker, Tom and Rebecca Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* (1987), vol. 22, no. 2.

Glazerman, Steven, Dan M. Levy and David Myers, "Nonexperimental versus Experimental Estimates of Earnings Impacts," *Annals of the American Academy of Political and Social Science* 589 (September 2003), 63-93.

Glazerman, Steven, Allison McKie, Nancy Carey, and Dominic Harris. "Evaluation of the Teacher Advancement Program (TAP) in the Chicago Public Schools: Study Design Report." Report prepared for The Joyce Foundation, November 2007.

Greenberg, David H. and Mark Shroder. *The Digest of Social Experiments.* Urban Institute Press. Washington, DC: 2004.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66 (September 1998), 1017-1098.

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

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (April 1998), 261-294.

Heinrich, C., P. Mueser, and K. Troske. Workforce Investment Act Non-Experimental Net Impact Evaluation. Report prepared for the U.S. Department of Labor Employment and Training Administration (December 2008).

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design." *Biostatistics*, vol. 1, 2000.

Hotz, V. Joseph and John Scholz. "Measuring Employment and Income for Low-Income Populations with Administrative and Survey Data." In *Studies of Welfare Populations: Data and Research Issues*, edited by Michele Ver Ploeg, Robert A. Moffitt, and Constance F. Citro. Washington, DC: National Academies Press, 2001.

Imbens, Guido W., and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," Institute for Research on Poverty Discussion Paper no. 1340-08, University of Wisconsin, 2008.

Imbens, G. and T. Lemieux (2008). Waiting for Life to Arrive: Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*, 142(2), 615-635.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Employment and Training Programs with Experimental Data." *American Economic Review* 76(4, September): 604-20.

Levitan, Sar A. 1992. *Evaluation of Federal Social Programs: An Uncertain Impact.* Washington, DC: George Washington University, Center for Social Policy Studies.

Mallar, Charles, Stuart Kerachsky, Craig Thornton et al. "Evaluation of the Economic Impact of the Job Corps Program." Report prepared for the U.S. Department of Labor, Employment and Training Administration (September 1982).

Maynard, Rebecca and others. "Supported Work Demonstration: Effects During the First 18 Months After Enrollment." Report prepared for the U.S. Department of Labor, Employment and Training Administration and the Ford Foundation (April 1979).

McConnell, Sheena and Steven Glazerman. "National Job Corps Study: The Benefits and Costs of Job Corps." Report prepared for the U.S. Department of Labor, Employment and Training Administration, June 2001.

McConnell, Sheena, Elizabeth Stuart, Kenneth Fortson and others. "Managing Customers' Training Choices: Findings from the Individual Training Account Experiment." Report prepared for the U.S. Department of Labor, Employment and Training Administration (December 2006).

Mueser, Peter R., Kenneth R Troske, and Alexey Gorislavsky, "Using State Administrative Data to Measure Program Performance," *Review of Economic and Statistics* (vol. 89, no. 4, November 2007), pp. 761-783.

Peikes, Deborah N., Lorenzo Moreno and Sean Michael Orzol, "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs," *The American Statistician* 62 (August 2008), 222-231.

Reich, Robert B. *Locked in the Cabinet*. New York, NY: Vintage Books, 1997.

Rosenbaum, Paul R., and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983), 41-55.

Schirm, Allen, Elizabeth Stuart, and Allison McKie. "The Quantum Opportunity Demonstration: Final Impacts." Report prepared for the U.S. Department of Labor, Employment and Training Administration (July 2006).

Schochet, Peter Z., John Burghardt, and Sheena McConnell. 2008. "Does Job Corps Work? Impact Findings from the National Job Corps Study." *American Economic Review*, 98(5): 1864–86.

Schochet, Peter Z. "An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations" *Evaluation Review*, vol. 33, no. 6, December 2009a.

Schochet, Peter Z. "Statistical Power for Regression Discontinuity Designs in Education Evaluations." *Journal of Educational and Behavioral Statistics*, vol. 34, no. 2, 2009b.

Schochet, Peter Z. and John Burghardt. "Using Propensity Scoring to Estimate Program-Related Subgroup Impacts in Experimental Program Evaluations." *Evaluation Review*, vol. 31, no. 2, April 2007.

Seftor, Neil, Arif Mamun, and Allen Schirm. "The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation." Reported prepared for the U.S. Department of Education, Policy and Program Studies Service. January 2009.

Spiegelman, Stephen A. Christopher O'Leary, and Kenneth Kline. The Washington Reemployment Bonus Experiment." Unemployment Insurance Occasional Paper 92-6. Washington, DC: U.S. Department of Labor, 1992.

Smith, Jeffrey A. and Petra E. Todd, "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (March-April, 2005a), 305-53.

Smith, Jeffrey A. and Petra E. Todd, "Rejoinder," *Journal of Econometrics* 125 (March-April, 2005b), 365-75.

Social Policy Research Associates and Mathematica Policy Research. "Evaluation of the Trade Adjustment Assistance Program: Design Report" Report prepared for the U.S. Department of Labor Employment and Training Administration (August 2004).

Woodbury, Stephen A. and Robert G. Spiegelman. "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois." *American Economic Review*, vol. 77, no.4, 1987.