A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies

JEFFREY SMITH*

1. INTRODUCTION

Active labor market policies aim to affect the outcomes – earnings, employment, health, etc. – of those who participate in them. The task of evaluation research lies in devising methods to reliably estimate their effects on those outcomes, so that informed decisions about program expansion and termination can be made. The past thirty years have witnessed real progress in our understanding of how to undertake evaluations of active labor market policies. The chapters by HECKMAN, LALONDE and SMITH (1999) and ANGRIST and KRUEGER (1999) in the most recent *Handbook of Labor Economics* capture the rapid pace of ideas in this area and the lively intellectual debate it engenders.

This paper lays out the basic form of the evaluation problem and then considers different methods for solving it. In describing the evaluation problem in Section 2, I highlight the role of individual heterogeneity in program impacts. Such heterogeneity has important implications both for the choice of impact estimator and for the interpretation of the resulting estimates. The remainder of the paper considers alternative methods recently advanced in the literature and employed in practice for evaluating active labor market policies. All of the methods I consider have been employed not just in the evaluation of active labor market policies, but also more broadly in the applied economics literature.

I begin in Section 3 by considering social experiments, sometimes held up as the "gold standard" of evaluation. I clarify both the strengths and the weaknesses of experimental methods. I argue that experimentation represents an important evaluation tool that should neither be summarily dismissed nor uncritically accepted.

In Sections 4 and 5, I consider the two non-experimental methods most popular in the recent literature: difference-in-differences and propensity score matching. To keep the paper short, I leave a more general treatment of non-experimental evaluation methods to standard references such as HECKMAN and ROBB (1985a,b), HECKMAN and SMITH (1996), HECKMAN, LALONDE and SMITH (1999) and ANGRIST and KRUEGER (1999). I emphasize that both the difference-in-differences and propensity score matching methods depend critically on maintained assumptions about the nature of the process by

* Professor, Department of Economics, University of Western Ontario and NBER, jsmith@julian.uwo.ca. The author thanks Dan Black, Michael Lechner and Miana Plesca for valuable comments.

which participants select into a program. These assumptions may or may not hold empirically in any particular context; indeed, the findings in HECKMAN and SMITH (1999) suggest that the assumptions underlying the difference-in-differences estimator represent a very poor approximation to reality in the case of job training programs.

While how to choose among alternative non-experimental estimators remains an important issue, I conclude my discussion of partial equilibrium evaluation methods in Section 6 by arguing that the literature has spent relatively too much time worrying about estimator choice and relatively too little time worrying about data quality. This pattern dates back at least to LALONDE's (1986) paper. He evaluates a set of standard non-experimental estimators using comparison groups drawn from different labor markets than the program participants, whose earnings are measured in different ways than the earnings of the participants, and not all of whom are known to be eligible for the program. Understanding the importance of different aspects of data quality to solving the evaluation problem remains a research area with a very high marginal product.

Finally, in Section 7 I address the issue of general equilibrium effects. Such effects come about when programs affect the outcomes and behavior of non-participants as well as participants. As shown in recent work by HECKMAN, LOCHNER and TABER (1998) and others, taking account of general equilibrium effects can strongly alter the conclusions that would be drawn from a partial equilibrium analysis. At the same time, the difficult methodological issues surrounding the analysis of general equilibrium effects mean that they will remain controversial in both the academic literature and the policy world. Despite this controversy, evaluators should pay attention to general equilibrium effects, if only indirectly through examining the sensitivity of cost-benefit analyses to alternative assumptions about them. Such sensitivity analyses would represent an improvement on much current partial equilibrium research that simply ignores general equilibrium effects.

2. THE EVALUATION PROBLEM

The evaluation problem exists because we only observe persons either in the state of the world where they participate in a program or in the state of the world where they do not, but never both. Solving the evaluation problem requires obtaining credible estimates of the counterfactual outcomes that would have been realized had persons made different program participation choices.

To see this more clearly, consider some very simple notation. Let Y_1 denote the outcome a person receives in the state of the world where he or she participates in the program being evaluated. This outcome could consist of earnings, employment, health or any other outcome that a program intends to affect. Let Y_0 denote the same outcome, measured in the same way over the same time period, in the state of the world where the person does not participate in the program. As already noted, a person can only participate or not participate, so exactly one of the two potential outcomes is observed for

each person. Nonetheless, it makes sense conceptually to associate both possible outcomes with each person, and to think of the difference between the two outcomes for a given person as the impact of the program on that person. Put differently, the impact of participation in a program for a given person consists of the difference it makes to their outcomes. In formal terms, the impact for person i is given by

$$\Delta_i = Y_{1i} - Y_{0i},$$

where Δ_i is the notation for the impact for person i.

The older literature on evaluation typically assumes that the impact of a program is the same for everyone – that is, that the impacts are homogeneous. Under this assumption, $\Delta_i = \Delta$ for all i. While unlikely to hold in a literal sense, this "common effect" assumption may be a reasonable approximation in some contexts (and a very poor one in others). It is this assumption that has largely guided the econometric and applied literatures on program evaluation in the past.

In recent years, substantial conceptual progress has resulted from thinking carefully and formally about models in which the impact of a program differs across persons. In particular, thinking about the evaluation problem in the context of heterogeneous impacts makes it clear that there is not just one parameter of interest but many. It also makes it clear that estimators that produce consistent estimates of one parameter of interest may not produce consistent estimates of others.

Now consider some possibilities for how the impact of a program might vary among persons. The simplest world, with no variation, is the "common effect" world already mentioned. In a slightly more general world, the impact of a program varies across persons, but prior to the program neither the potential participant nor program staff have any information about the person-specific component of the impact. Put differently, programs have different effects on different persons, but no one can predict in advance who will gain more or who will gain less, so that the variation in impacts has no effect on who participates in the program. In this slightly more general world, the variation in impacts has few policy implications.

In the most general world, the impact varies across persons and either the person or program staff or both have some information about it prior to participation. In this most general world, the person-specific component of the impact does affect participation in the program. As a result, it has important policy implications, as it means that different policy changes, which include or exclude different sets of persons from the program, will have different mean impacts.

To see why the variation in impacts can have implications for policy, consider three parameters that might be of interest to a policymaker. Consider these parameters in the context of a voluntary program that serves part but not all of some population of interest, for example, a voluntary job-training program for persons receiving social assistance. One parameter of obvious interest is the effect that the program has on its current participants. The literature calls this parameter the impact of "treatment on the treated"

(TT) or, in the case of our example, of training on the trained. When combined with information on program costs, and putting aside for the moment the issue of general equilibrium effects other than tax effects, this parameter answers the policy question of whether or not the program should be eliminated. In a strict cost-benefit world, a program for which the mean impact of treatment on the treated lies below the per-participant cost of the program (including the deadweight costs associated with the taxes that finance the program) should be eliminated.

Program elimination is often not the only, or even the primary, policy proposal of interest. Suppose instead that the policy of interest is a 10 percent reduction in the number of persons served under the program, to be accomplished in some specified way, such as by instituting a small fee for the training materials, or by rationing the available spaces on a first-come, first-served basis. In this case, the parameter of interest is not the impact of the program on all those it currently serves, but rather its impact on the 10 percent of persons whom it would cease to serve were the policy change put in place.

In a world of heterogeneous impacts, it could well be that the mean impact for this marginal group does not exceed the costs of providing services to them, while the mean impact for the other 90 percent of participants would suffice to pass a cost-benefit test. Indeed, if those who benefit the most from the program are those who are most eager to participate (and therefore most willing to pay the training fee or get to the program office first), then this is exactly what one might expect. A very simple economic model of program participation indicates that if potential participants have some idea of their person-specific gain from the program, then those with the largest gains should be the most likely to participate, all else equal.

This marginal impact parameter is an example of what IMBENS and ANGRIST (1994) call a "local average treatment effect" or LATE. It is a treatment effect at the margin of participation defined relative to some instrument, where in this case the instrument would be the mechanism used to reduce participation, such as the small fee for training materials. This LATE measures the mean impact of the program on those persons whose participation status changes due to the change in the policy instrument.

Rather than seeking to eliminate or cut the program, the policy proposal under consideration may seek to expand the program to all eligible persons. In the context of our example, this would mean making the job training program mandatory for all social assistance recipients. The policy question of interest now becomes whether or not the mandatory program would pass a cost-benefit test. The impact parameter of interest becomes what the literature calls the "average treatment effect" (ATE). This parameter gives the mean impact of treatment on all persons eligible for it, rather than just on those who choose to voluntarily participate. Thinking again about a simple model of program participation in which those with the largest expected gains participate, we would expect the ATE to be less than the impact of treatment on the treated.

Of course, in a common effect world, all three impact parameters – TT, LATE and ATE – are the same. This simplicity is part of the attraction of the common effect world, however unrealistic the common effect assumption might seem. In a world of heteroge-

neous program impacts, when agents or program staff have some information about the impacts, these three impact parameters will likely differ, and the differences can matter for policy purposes.

Heterogeneity in the effects of programs also has implications for some commonly used non-experimental evaluation strategies, such as the method of instrumental variables. HECKMAN, LALONDE and SMITH (1999) and HECKMAN (1997) discuss these issues in more detail. Finally, in addition to the TT, LATE and ATE parameters, we can also define a number of other parameters of interest, such as the variance of impacts among participants. HECKMAN, SMITH and CLEMENTS (1997) discuss the estimation of such parameters.

3. SOCIAL EXPERIMENTS

Social experiments have become the method of choice in the evaluation of social programs in North America. High profile evaluations such as the National JTPA Study in the U.S. (see BLOOM *et al.*, 1997) and the Self-Sufficiency Project in Canada (see, e.g., MICHALOPOULOS *et al.*, 2000) have brought about real changes in the views and, in the first case, the actions of policymakers. With a few exceptions such as the Restart experiments in Britain (see, e.g., WHITE and LAKEY, 1992, and DOLTON and O'NEILL, 1996), some random assignment evaluations of training programs in Norway (see TORP *et al.*, 1993), and a small experiment in Sweden described in BJÖRKLUND and REGNÉR (1996), these methods have only recently emerged as an evaluation alternative in most European countries. In this section, I consider the costs and benefits of social experiments, concluding that they represent an important tool for evaluation, but one that requires careful implementation and interpretation. For additional (and sometimes more technical) discussion of social experiments, see BJÖRKLUND and REGNÉR (1996) BURTLESS and ORR (1986), BURTLESS (1995), HECKMAN and SMITH (1993,1995,1996a,b), and HECKMAN, LALONDE and SMITH (1999).

Ideally, social experiments take persons who would otherwise participate in a program and randomly assign them to one of two groups. The first group, called the treatment group, receives the program as usual, and the second group, called the control group, is excluded from it. Experimental control groups differ from traditional non-experimental comparison groups composed of naturally occurring non-participants because, up to sampling variation, they have the same distribution of observed and unobserved characteristics as the participants in the experimental treatment group. In a non-experimental evaluation, statistical techniques are used to adjust the outcomes of persons who choose not to participate to "look like" what the participants would have experienced, had they not participated. In contrast, an experiment directly produces the counterfactual of interest by forcing some potential participants not to participate.

As a result of random assignment, under certain assumptions a simple comparison of the mean outcomes in the experimental treatment and control groups produces a consis-

tent estimate of the impact of the program on its participants. In terms of the parameters of the preceding section, a social experiment produces a consistent estimate of the impact of treatment on the treated. With clever designs, social experiments can also be used to obtain estimates of the average treatment effect, as in the British Restart experiment where persons were randomly denied an otherwise mandatory treatment. Similarly, random assignment at the policy margin, as in the evaluation of "profiling" (assigning treatment based on the predicted duration of unemployment) unemployment insurance claimants by BLACK, SMITH, BERGER and NOEL (2000), yields experimental estimates of a LATE.

Beyond the simple fact that, in the absence of the problems discussed later in this section, social experiments produce consistent estimates of the impact of treatment on the treated, social experiments have several advantages relative to standard non-experimental methods. First, social experiments are simple to explain to policymakers. Most educated persons understand the idea behind random assignment.[1]

Second, experiments are less controversial than non-experimental methods. In North America, the widely varying estimates of the impact of the Comprehensive Employment and Training Act programs described in BARNOW (1987) led to serious skepticism about non-experimental methods. In these evaluations, different researchers using the same data set came to dramatically different conclusions about program effectiveness.[2] In contrast, experiments are held to deliver "one number" rather than the panoply of different estimates often produced in non-experimental evaluations. This point is sometimes overstated by advocates of experiments in light of the observed sensitivity of experimental impact results to various empirical judgement calls (see HECKMAN and SMITH, 2000). Despite this sensitivity, however, experimental impact estimates, because of the simple and straightforward methodology that underlies them, remain compelling relative to non-experimental estimates.

Third, it is hard to cheat on an experiment. That is, if the person, firm or organization conducting the evaluation prefers to find that a program works well or does not work well, relying on an experimental evaluation makes it more difficult for them to generate the impact estimate they want. In contrast, a smart non-experimental evaluator could use the information in the literature about the biases commonly associated with specific non-experimental estimators to strategically choose an estimation strategy that would produce the desired findings. Forcing an experiment on the evaluator makes such manipulation much more difficult as it removes the choice of estimator from the evaluator's strategic toolkit.

Fourth, experiments provide a valuable opportunity to calibrate individual non-ex-

1. Of course, all of the complex issues associated with any impact estimate, whether experimental or non-experimental, remain. This includes issues such as the extent to which impact estimates for one program and one population can be generalized to other, similar, programs or to other populations.
2. Note that some of these differences were due to choices about how to handle the data, rather than what non-experimental estimator to use. See DICKINSON, JOHNSON and WEST (1987).

perimental estimators and, more broadly, to examine the efficacy of strategies for systematically choosing among alternative non-experimental estimators. LALONDE's (1986) paper uses data from the U.S. National Supported Work Demonstration (NSW) experiment to examine the biases associated with the common evaluation strategy of drawing a comparison group from an existing national data set and then applying standard non-experimental techniques.[3] His finding that the estimates produced by standard estimators rarely came close to the experimental estimates played a major role in the shift to social experiments in North America.

In more recent work, DEHEJIA and WAHBA (1999a, b) and SMITH and TODD (2000) use the same NSW data to examine the performance of propensity score matching, which I discuss in detail in Section 5. HECKMAN, ICHIMURA, SMITH and TODD (1996, 1998) and HECKMAN, ICHIMURA and TODD (1997) use the experimental data from the National JTPA Study to examine matching methods and to characterize the nature of selection bias more generally. Finally, HECKMAN and HOTZ (1989) find, using the NSW data, that choosing among alternative non-experimental estimators using specification tests reduces the bias associated with non-experimental methods.[4]

While social experiments have a number of advantages over standard non-experimental methods, they do not represent a simple solution to every possible evaluation problem. The remainder of this section considers limitations and potential problems with social experiments.

To begin with, social experiments cannot estimate all parameters of interest. This limitation has several dimensions. First, some "treatments" (broadly defined), such as sex or family income while young, defy random assignment. Second, while social experiments are generally well suited to estimate the impact of treatment on the treated, they are poorly suited to estimate general equilibrium effects on persons not randomly assigned. I discuss these general equilibrium effects in more detail in Section 7. Finally, even within the standard, partial equilibrium evaluation context, parameters that depend on the link between outcomes in the participation and non-participation states, such as the variance in impacts among participants, require additional, non-experimental assumptions to estimate, even with experimental data. HECKMAN, SMITH and CLEMENTS (1997) discuss this latter issue in detail.

Second, the presence of random assignment may disrupt the operation in the program, resulting in an impact estimate that corresponds to something other than the program as it normally operates. The literature refers to this as "randomization bias." Consider three examples. First, if the number of persons in the program is the same during the experiment as at other times, program operators will have to recruit additional potential participants during the experiment in order to fill the control group. These additional recruits, who will be randomly divided between the treatment and control groups, may have a dif-

3. See the related analyses in FRAKER and MAYNARD (1987) and LALONDE and MAYNARD (1987).
4. Section 8.4 of HECKMAN, LALONDE and SMITH (1999) discusses the limitations of the specification testing strategy they examine. See REGNÉR (2001) and RAAUM and TORP (2001) for recent applications to evaluating European active labor market policies.

ferent impact from the program than those who would normally participate. Second, randomization might affect survey response rates in the treatment and control groups in ways that would not occur in a non-experimental evaluation. Experimental controls, denied the opportunity to participate in the program, might refuse to participate in the data collection as well. Finally, if participants normally undertake activities affecting their impact from the program prior to starting it, the threat of random assignment may cause them to cut back on these activities, as they may turn out to be wasted.

Third, experiments are sometimes more expensive than non-experimental methods. Random assignment does have costs, as it typically requires substantial staff training, on-going staff monitoring and information provision to the potential participants, who typically must sign a contract agreeing to random assignment. At the same time, as pointed out by HECKMAN, LALONDE and SMITH (1999) (see Section 8.1), this case can be overstated. Non-experimental evaluations are inexpensive when they rely on existing national data sets for their comparison groups. However, using national data sets almost always means not drawing the comparison group from the same local labor markets as the participants, and often means not measuring the outcome variables in the same way for participants and non-participants. If these factors are important to reducing bias, then the savings associated with using an existing national data set comes at the cost of biased estimates.

Fourth, random assignment sometimes engenders political controversy or bad publicity. For example, in the U.S. National JTPA Study, evaluators had to contact around 200 of the approximately 600 JTPA training centers in the U.S., and had to pay US$ 1 million in budgetary side payments, in order to find 16 training centers that would voluntarily participate in the experiment. According to DOOLITTLE and TRAEGER (1990), a primary concern of the training centers that chose not to participate was the potential for negative publicity associated with using random assignment.

Finally, interpretation of experimental estimates is complicated in situations where members of the experimental treatment group drop out of the experiment prior to receiving any (or receiving full) treatment, and where experimental control group members can participate in alternative programs offering the same or similar services. If only treatment group dropouts pose a problem, then standard methods exist for retrieving an estimate of the impact of treatment on the treated (see, e.g., BLOOM, 1984, and HECKMAN, SMITH and TABER, 1998).

In the presence of control group substitution into alternative programs similar to the one being evaluated, things become much more difficult. The experimental estimate now compares the program being evaluated to the other programs in the environment, rather than to no program at all. If the other programs work as well or as poorly as the one being evaluated by the experiment, and if roughly equal numbers participate in some program in the experimental treatment and control groups, then the experimental impact estimate will be zero regardless of the impact of the program relative to no program at all. HECKMAN, HOHMANN, SMITH and KHOO (2000) show that careful interpretation is crucial in such circumstances, and that obtaining estimates of the impact of the

program relative to no program requires application of non-experimental methods to the experimental data.

To conclude, experimental methods have proven very successful in North America at providing convincing estimates of the impact of both demonstration programs and existing programs. At the same time, as experience with experiments has grown, it has been recognized that in practice, their design and interpretation is often more difficult than it might first appear. Issues of randomization bias, dropout from the program among treatment group members, and substitution into alternative programs among experimental controls, complicate the development and interpretation of experimental evaluations. These limitations certainly do not indicate that experiments should be avoided. Instead, they indicate that, in the words of Burt Barnow, "experiments are not a substitute for thinking."

4. DIFFERENCE-IN-DIFFERENCES

In situations where experimental data are unavailable or subject to the types of problems outlined in the preceding section, evaluators must rely instead on non-experimental evaluation methods. These methods rely on naturally occurring variation in program participation, combined with statistical adjustment of the observed outcomes of non-participants, to produce impact estimates. The statistical adjustments that define each non-experimental estimator derive their motivation from models of program participation and its relationship to outcomes in the participation and non-participation states.

At a crude level, the various non-experimental estimators can be divided into those primarily concerned with selection on observables and those primarily concerned with selection on unobservables. Models in which participation is random conditional on some set of observed covariates motivate estimators that deal only with selection on observables. This class of estimators includes the propensity score matching estimator discussed in Section 5. The estimators in this class differ mainly in how the conditioning gets done.

Models in which factors other than observed covariates jointly affect participation and outcomes motivate estimators that deal with selection on unobservables. This class of estimators includes, among others, the difference-in-differences estimator considered here, the classical HECKMAN (1979) bivariate normal estimator and the second-differences estimator considered in HECKMAN and HOTZ (1989). These estimators differ because the assumptions that the underlying models make about the inter-relationship between the participation and outcome processes differ.

The difference-in-differences estimator builds on a model that decomposes the outcome equation unobservable into two components.[5] One component is time-invariant –

5. MOFFITT (1991) provides a very clear introduction to the basics of longitudinal estimators, of which the difference-in-differences estimator is just one example. HECKMAN (1996), commenting on EISSA (1996), provides a more detailed critique of the difference-in-differences estimator.

a so-called fixed effect – and the other is transitory. In notation, the outcome equation is

$$Y_{it} = \beta_0 + \beta_1 X_{i1t} + ... + \beta_k X_{ikt} + \Delta_i D_i + \mu_i + \epsilon_{it},$$

where Y_{it} is again the outcome variable of interest for person i in period t, the X_i denote observed determinants of outcomes with associated coefficients β_1, ..., β_k, D_i and Δ_i denote the participation indicator and person-specific impact, respectively, μ_i is the unobserved, time-invariant fixed effect and ϵ_{it} is the transitory component of the unobservable. The model that motivates the difference-in-differences method assumes that participation depends on the fixed effect μ_i but not on the transitory component ϵ_{it}. The usual story is that participants are more able or more motivated (or perhaps less able or less motivated) than non-participants, and that these differences in ability or motivation affect their outcomes in every period.

In the context of this model, simply running a regression of outcomes on X_i and D_i will result in inconsistent estimates, because the unobserved fixed effect is correlated with the participation indicator D_i. However, because the fixed effect is time-invariant, it can be differenced out. Assuming that one period of pre-program data is available – that is, one period of data before the participants have participated, the following equation can be estimated:

$$Y_{it} - Y_{is} = \beta_0 + \beta_1 (X_{i1t} - X_{i1s}) + ... + \beta_k (X_{ikt} - X_{iks}) + \Delta_i D_i + (\epsilon_{it} - \epsilon_{is}),$$

where s denotes a period prior to participation in the program (for the participants). This estimator consists of the difference between the before-after earnings difference for participants and the before-after earnings difference for the non-participants – hence the name "difference-in-differences." Under the assumption that selection into the program depends only on the fixed effect and not on the transitory component, it provides consistent estimates of the impact of treatment on the treated.

How well does this assumption of selection on a time-invariant fixed effect correspond to the facts? This assumption implies that there should be a fixed difference in outcome levels between participants and non-participants prior to participation. There should also be a fixed difference after participation, which will equal the fixed difference before participation plus the impact of treatment on the treated.

In contrast to these assumed fixed differences over time, the data from a wide variety of evaluations exhibit a phenomenon known as Ashenfelter's dip. As discussed in Section 4.1 of HECKMAN, LALONDE and SMITH (1999), Ashenfelter's dip consists of the recurring pattern whereby the mean earnings of participants decline in the period leading up to program participation. This dip is consistent with selection on the transitory component of earnings rather than, or in addition to, selection on a fixed effect.

The data from the experimental control group from the U.S. National JTPA Study presented in HECKMAN and SMITH (1999) display another pattern inconsistent with the

difference-in-differences model. For most demographic groups, the control group data reveal that the earnings of persons who would have participated in the program but for random assignment increase relative to the earnings of ordinary non-participants in the post-program period. Thus, the assumption of a fixed difference in the post-program period also fails to hold in the JTPA data. The difference-in-differences impact estimates presented in HECKMAN and SMITH (1999) show that the failure of the assumptions justifying the estimator to hold empirically results in estimates that differ strongly from the corresponding experimental impact estimates. They also find, not surprisingly given the control group earnings patterns, that the difference-in-differences estimates are quite sensitive to the precise choice of the "before" and "after" periods used in constructing them. Overall, the available evidence suggests that the difference-in-differences estimator, though motivated by plausible stories about differences in motivation or ability, may be a poor choice in many evaluation contexts.

5. MATCHING

Unlike the method of difference-in-differences just considered, matching methods concern themselves solely with selection on observable variables. As such, they require very rich data in order to make the estimates they generate credible. Matching methods are not new, even to the literature on program evaluation. Some of the evaluations of the U.S. Comprehensive Employment and Training Act (CETA) reviewed in BARNOW (1987) use modified forms of matching. What is new is the use of "propensity score" matching, developed in ROSENBAUM and RUBIN (1983). Propensity score matching, rather than using a vector of observed characteristics X, matches participants and non-participants based on their estimated probability of participation $P(X)$. ROSENBAUM and RUBIN (1983) show that when matching on X produces consistent estimates, so does matching on $P(X)$.

The advantage of matching on $P(X)$ rather than X is that $P(X)$ is a scalar, while X may have many dimensions. When X is of high dimension, matching becomes difficult because for some values of X among participants no close matches will be found among comparison group members. This problem becomes less important (though it does not disappear as I note below) when matching on the scalar $P(X)$.

Matching, whether on X or on $P(X)$, relies on a conditional independence assumption. This assumption states that, once you condition on X or on $P(X)$, participation in the program is independent of the outcome in the non-participation state (Y_0 in the notation defined in Section 2). This is not a trivial assumption. It requires that all variables that affect both participation and outcomes in the absence of participation be included in the matching. Clearly, making this conditional independence assumption plausible in practice requires access to very rich data. It also requires careful thought, guided by economic theory, about what variables do and do not affect participation and outcomes.

At this point, the reader may wonder how matching methods differ from simply run-

ning regressions. After all, running a regression of outcomes on a participation indicator and X produces an impact estimate that conditions on X. I consider two important differences here. First, matching is non-parametric. As such, it avoids the functional form restrictions implicit in running a linear regression. The evidence presented in DEHEJIA and WAHBA (1998) and in SMITH and TODD (2000), who directly compare matching and regression estimates constructed using the same X, suggests that avoiding these functional form restrictions can be important to reducing bias. Of course, with a sufficient number of higher-order and interaction terms included in the regression, this difference fades. However, the inclusion of such terms (other than age or education squared) is uncommon in practice.

Second, matching vividly highlights the so-called "support" problem. The support of a distribution is the set of values for which it has positive density – that is, the set of values with a non-zero probability. It is relevant to matching because it will sometimes be the case empirically that for certain values of X or of $P(X)$ present in the participant sample there will not be any observations present in the non-participant sample.[6] In such cases, the support of the two samples differs. Moreover, the common support – the set of values where there are observations in both samples – may not include all of the participant observations. Note that for the estimation of the impact of the treatment on the treated, it does not matter if there are non-participant observations with no analogues in the participant sample. All that is required to estimate the treatment on the treated parameter is that there be analogues for each of the participants in the non-participant sample. Note also that if there are values of X such that $P(X) = 1$, then participants with such values necessarily lie outside the common support because their probability of not participating is zero.

When the support condition fails and there are no non-participants to match with for some participants, an impact estimate cannot be obtained for these participants. In this case, if impacts vary across persons as described in Section 2, matching will produce an impact estimate whose population analogue differs from that estimated by other estimators that do not drop observations lacking a common support. Matching highlights the common support problem in the sense that it makes it easy to see when the support condition fails. In the propensity score matching case, simple histograms such as those presented in HECKMAN, ICHIMURA, SMITH and TODD (1998) and DEHEJIA and WAHBA (1999) make the problem clear.[7] In contrast, in analyses that estimate impacts simply by running regressions on X, the issue is rarely even investigated.

Some caveats also apply to the use of matching methods. I mention three of the most important here. First, while matching removes from the researcher the need to make decisions about functional form, it does not remove the problem of variable selection. That is, the researcher must decide what variables to include in X. No deterministic algo-

6. The extent of the support problem implicitly depends on the tolerance of the researcher for poor (i.e., not very comparable) matches. See Heckman, Ichimura, Smith and Todd (1998) for an extended discussion of the support issue and ways of dealing with it.
7. See Figure 2 in the first case and Figures 1 and 2 in the second case.

rithm, other than comparing the resulting estimates to those from an experiment, exists to guide the researcher in making this decision.[8] HECKMAN, ICHIMURA, SMITH and TODD (1998) show that the estimates produced by matching can be quite sensitive to the choice of variables used to construct $P(X)$.

Second, the choice of matching method can make a difference in small samples. A number of different matching methods coexist in the literature (see HECKMAN, ICHIMURA and TODD, 1997, for an extended discussion). The most common is nearest neighbor matching, in which the non-participant closest (in terms of X or $P(X)$) to each participant is chosen as the participant's match. The outcome of the nearest neighbor approximates the participant's counterfactual non-participation outcome – that is, it approximates what would have happened to the participant, had he or she not participated. Nearest neighbor matching can be operationalized with more than one nearest neighbor and with and without replacement, where "with replacement" means that a given non-participant observation can form the counterfactual for more than one participant. Alternatives to nearest neighbor matching include kernel matching, in which a weighted average of the outcomes of observations close to each participant provides the counterfactual, or local linear matching, in which a local linear regression is run for each participant to obtain the counterfactual. These methods are all consistent[9] as they all become closer and closer to comparing only exact matches as the sample size grows. However, in small samples they can provide somewhat different answers, and certain methods have properties that make them a better choice in particular contexts.

Third, it is important to get the correct standard errors. The estimation of the propensity scores (if propensity score matching is used) and the matching itself both add variation beyond the normal sampling variation (see the discussion in HECKMAN, ICHIMURA and TODD, 1998). In the case of nearest neighbor matching with one nearest neighbor, treating the matched comparison sample as given will understate the standard errors. In practice, most researchers report bootstrapped standard errors.

A small literature has accumulated over the past few years that uses experimental data to evaluate the performance of matching. Two sets of papers give somewhat different results. The first set of papers – HECKMAN, ICHIMURA, SMITH and TODD (1996, 1998) and HECKMAN, ICHIMURA and TODD (1997) – uses the data from the U.S. National JTPA Study. These papers find that matching substantially reduces the raw bias in earnings between participants and eligible non-participants drawn from the same local labor markets and with earnings information collected in the same way. At the same time, the bias that remains in the preferred specification is of the same order of magnitude as the experimental impact estimate. In contrast, DEHEJIA and WAHBA (1998, 1999) use the

8. The "balancing test" proposed in ROSENBAUM and RUBIN (1983) and applied by DEHEJIA and WAHBA (1998, 1999) and by LECHNER (1999) aids the researcher in determining whether or not to include higher-order and interaction terms for a given X. It does not aid the researcher in selecting the variables to include in X. See the discussion in SMITH and TODD (2000).

9. Statistically, an estimator is consistent if the probability that it deviates from the population parameter value by any given amount goes to zero as the sample size increases.

data from the U.S. National Supported Work Demonstration and reach more optimistic conclusions. They apply propensity score matching methods to a subset of the data from LaLonde (1986) that allows matching on pre-program earnings variables. In their preferred specification, matching eliminates the vast majority of the bias. Smith and Todd (2000) argue that the Dehejia and Wahba results depend crucially on their choice of subsample and of X variables. Changing either choice leads to results that look more like those found using the data from the JTPA experiment.[10]

6. BETTER DATA HELP A LOT

A common theme of much of the evaluation literature in the 1970s and 1980s, such as LaLonde (1986), Barnow (1987) and Heckman and Hotz (1989), is that of estimator choice. In this strand of the literature, the evaluation problem is posed as follows: given the available data, what estimator will produce consistent estimates of program impacts. Left out of the discussion are the data themselves, and the role that they play in allowing consistent estimates.

More recent work by Heckman, Ichimura, Smith and Todd (1998) highlights the importance of particular data issues and shows that they often contribute as much or more to the total bias as the choice of non-experimental estimator. Their work focuses on two key factors, already mentioned briefly above. The first factor consists of drawing comparison group members from the same local labor markets as participants. They estimate the importance of this factor in two ways. First, they mismatch the four experimental sites at which special comparison group data were collected as part of the U.S. National JTPA Study. These data rely on the same survey instruments that were administered to the experimental sample, and so allow the isolation of bias due to geographic mismatch. They find that putting non-participants in the same local labor markets as participants strongly reduces the bias in non-experimental estimates.

The second factor they consider is differences in the way in which the dependent variable, typically earnings, is measured. As already noted, the influential LaLonde (1986) study uses comparison groups with earnings measured in different ways than the participants. Evidence from comparisons of multiple earnings measures in the data from the National JTPA Study presented in Smith (1997) illustrates the potential for bias due to measurement error of this sort. In that paper, I show that different earnings measures for the same persons and the same time-period – including two survey-based measures and two administrative measures – can yield substantially different estimates of mean earnings. Heckman, Ichimura, Smith and Todd (1998) examine this issue by constructing a comparison group from a national U.S. data set, the Survey of Income and Program Participation (SIPP). Their SIPP analysis combines the effect of local labor

10. For a somewhat more optimistic view of matching, see the long series of papers by Rubin and various co-authors. Rubin and Thomas (2000) is among the most recent and cites many of the earlier papers.

market mismatch with those of different earnings measures and shows again that the data, rather than the estimator, can have a substantial effect on the bias associated with non-experimental methods.

Indeed, part of the reason why social experiments often look good in comparison with non-experimental estimators is precisely that social experiments always collect data that satisfy these conditions. The control group is always drawn from the same local labor market and outcome variables for the two groups are always measured in the same way. This particular feature of social experiments can, and should, be carried over to non-experimental evaluations. More generally, as HECKMAN, LALONDE and SMITH (1999) argue, the literature has probably put relatively too much effort into worrying about the problem of estimator selection, important as that problem is, and relatively too little effort into studying the role of data quality in reducing bias in non-experimental evaluations.

7. GENERAL EQUILIBRIUM EFFECTS

General equilibrium effects occur when a program affects persons other than its participants. For example, an active labor market program that provides job search assistance to the long-term unemployed may increase the speed with which its participants obtain work, but may also slow down the return to work of the short-term unemployed. This effect is called *displacement* (see, e.g., CALMFORS, 1994). In this example, long-term unemployed persons with improved job search skills due to the program take jobs that would otherwise have been taken by short-term unemployed persons. Related to this are *substitution* effects[11] where, e.g., subsidies to one group of workers cause employers to substitute them for other workers and *deadweight* effects, where, e.g., activity that would have occurred anyway is subsidized. CALMFORS (1994) also notes the importance of *tax* effects, whereby the taxes collected to finance a program distort the choices of both participants and non-participants. A complete accounting of either the cost-benefit performance of a program or of its distributional effects must include these general equilibrium effects.[12]

General equilibrium effects will only be important in certain contexts. At the simplest level, such effects will play a more important role in the evaluation of large (relative to the relevant population) programs than in the evaluation of small ones. Thus, a small demonstration program that treats 100 individuals in a large, urban labor market will not generate noticeable general equilibrium effects. On the other hand, a universal program

11. These substitution effects differ conceptually from those discussed in the context of social experiments, despite the similar terminology.
12. Note that general equilibrium effects differ from what are sometimes called "macro" effects, whereby the state of the economy affects program effectiveness. For example, a given program may have a larger impact when the unemployment rate is four percent than when it is ten percent. Such effects may be important in some cases, but they are not general equilibrium effects as defined in this section.

that provides a generous subsidy to attending university almost certainly will have important general equilibrium effects. Of course, a program with no partial equilibrium effects will likely not have general equilibrium ones either. A training program that does not improve the human capital of its participants will not lead them to displace non-participants in the labor market (although the taxes required to pay for it may alter the labor supply choices of both trainees and non-trainees).

General equilibrium effects cause problems for evaluation researchers because the partial equilibrium methods they most commonly use either miss these effects entirely or, perhaps worse, are biased by them. To see how problems can arise, consider a simple evaluation of a training program that compares the earnings of a sample of participants with those of a comparison group of similar non-participants. If the program has important displacement effects, then these effects will show up in lower average earnings among the comparison group members, some of which will have been displaced. This leads to an upward bias in the estimated impact of the program on its participants. Of course, due to the displacement effects, the impact on participants alone is an upward biased estimate of the overall social impact of the program. Note that this problem of partial equilibrium evaluation methods being unable to pick up general equilibrium effects extends to social experiments.

To help illustrate the potential importance of general equilibrium effects to policy evaluation, and to give a sense of some of the magnitudes that have been estimated in the literature, consider the following three examples. The first two examples both concern the U.S. unemployment insurance (UI) bonus experiments, which receive a careful survey in MEYER (1995). In the bonus experiments, UI recipients who found a job within a certain period – relatively short by U.S. standards and extremely short by European ones – after the start of their UI spell and held it for at least a certain period (usually four months) received a cash bonus.

The first example is due to MEYER (1995). He notes that in a permanent UI bonus program, rather than in a demonstration, the presence of the bonus and the rules for its receipt would become widely known. As a result, both worker and firm behavior would change in several dimensions. For example, in the U.S., many persons who have short spells of unemployment between jobs, and who are eligible for UI, presently do not collect any UI, presumably due to the fixed costs in terms of time and trouble necessary to obtain UI, and perhaps due to stigma as well. The bonus would lead some of these persons to apply for and receive some UI, in order to collect the bonus. This is a classic example of a deadweight effect, in which persons receive a bonus for behavior they would have engaged in anyway. This general equilibrium effect would reduce the net effect of the program relative to that estimated by the experiments.

In the second example, DAVIDSON and WOODBURY (1993) estimate a Mortensen-Pissarides structural search model in order to estimate the displacement effects of the bonus. They find substantial displacement effects among unemployed workers who are not eligible for UI (and, therefore, not eligible for the bonus) due to working too little in the previous year. Overall, their results indicate that 30 to 60 percent of the gross im-

pact – that is, of the partial equilibrium impact as estimated by the experiments used to evaluate the bonus program – is offset by displacement.

The third example comes from HECKMAN, LOCHNER and TABER (1998). For the U.S., they consider a policy of subsidies to attend college or university. They develop a rational expectations, perfect foresight, overlapping generations model of the U.S. economy that includes heterogeneous skills (levels of schooling in their case) with separate and endogenous prices. Using this framework, they simulate the effects of a revenue-neutral $500 increase in the present subsidy to attending college or university. Their partial equilibrium increase in attendance, calculated with skill prices fixed, is 5.3 percent in the steady state. In sharp contrast, the general equilibrium increase in attendance, calculated with changing skill prices, is only 0.46 percent. The strong difference arises because increasing the number of college and university graduates depresses their wage in the labor market, and correspondingly increases the wage of the now more scarce high school graduates. These changes in prices mute the effect of the subsidy – by their calculations by over 90 percent.

Two important issues arise in contexts likely to include general equilibrium effects. First, additional parameters of interest become relevant. In a general equilibrium context, in addition to the parameters discussed in Section 2, the researcher will also be interested in the effect of the program on non-participants. This impact on non-participants may be decomposed in various ways, e.g., into effects through the labor market and effects through the tax system. In certain contexts, such as that of HECKMAN, LOCHNER and TABER (1998), variants of the local average treatment effect (LATE), defined in Section 2, can be constructed. In their model, the subsidy policy moves some persons from high school to college and others from college to high school. They define a LATE for each group as well as an overall LATE consisting of a weighted average of the two.

The second issue, of course, is how to estimate the general equilibrium effects. One strand of the literature uses variation in program scale across jurisdictions, combined with data at the jurisdictional level, to estimate the effects. A recent example is FORSLUND and KRUEGER (1994). The other strand of the literature estimates structural, general equilibrium models. Both the DAVIDSON and WOODBURY (1993) and HECKMAN, LOCHNER and TABER (1998) papers use such models. They have the advantage that they make explicit assumptions about the mechanism generating the general equilibrium effects. They also provide a framework that allows for estimation of many evaluation parameters of interest. The key disadvantage of such models, other than their computational and conceptual complexity, is the strong assumptions they require about functional forms of economic relationships and about the values of key economic parameters.

As structural general equilibrium models have only recently begun to penetrate the evaluation literature in significant numbers, their conclusions remain controversial and their value relative to more traditional methods (and relative to their high cost of production) remains an open research question. What remains more certain is the likely importance, despite the literature's general avoidance of the topic, of the general equilibrium effects of active labor market policies.

8. CONCLUSIONS

This paper has reviewed and commented on recent developments in evaluation research. I have outlined recent methodological developments and their implications for evaluation practice and policy. I have also provided copious citations to the technical literature related to these developments. My main conclusion is that while much has been learned over the past three decades, there remains a lot of room for improvement in econometric evaluation methodology and, even more so, in evaluation practice.

REFERENCES

ANGRIST, J. and A. KRUEGER (1999), "Empirical Strategies in Labor Economics", in: O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics Volume 3A*, Amsterdam, 1277–1366.

BARNOW, B. (1987), "The Impact of CETA Programs on Earnings: A Review of the Literature", *Journal of Human Resources*, 22, 157–193.

BJÖRKLUND, A. and H. REGNÉR (1996), "Experimental Evaluation of European Labour Market Policy", in: G. Schmid, J. O'Reilly and K. Schömann, eds., *International Handbook of Labour Market Policy and Evaluation*, Brookfield, VT, 89–114.

BLACK, D., J. SMITH, M. BERGER and B. NOEL (2000), "Is the Threat of Reemployment Services More Effective Than the Services Themselves: Experimental Evidence from the UI System", Unpublished manuscript, University of Western Ontario.

BLOOM, H. (1984), "Accounting for No-Shows in Experimental Evaluation Designs", *Evaluation Review*, 82(2), 225–246.

BLOOM, H., L. ORR, S. BELL, G. CAVE, F. DOOLITTLE, W. LIN and J. BOS (1997), "The Benefits and Costs of JTPA Title II-A Programs: Findings from the National Job Training Partnership Act Study", *Journal of Human Resources*, 32(3), 549–576.

BURTLESS, G. (1995), "The Case for Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, 9(2), 63–84.

BURTLESS, G. and L. ORR (1996), "Are Classical Experiments Needed for Manpower Policy", *Journal of Human Resources*, 21, 606–639.

CALMFORS, L. (1994), "Active Labor Market Policy and Unemployment – A Framework for the Analysis of Crucial Design Features", *OECD Economic Studies*, 22(1), 7–47.

DAVIDSON, C. and S. WOODBURY (1993), "The Displacement Effects of Reemployment Bonus Programs", *Journal of Labor Economics*, 11(4), 575–605.

DEHEJIA, R. and S. WAHBA (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies", NBER Working Paper #6829.

DEHEJIA, R. and S. WAHBA (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94(448), 1053–1062.

DICKINSON, K., T. JOHNSON and R. WEST (1987), "An Analysis of the Sensitivity of

Quasi-Experimental Net Estimates of CETA Programs", *Evaluation Review*, 11, 452–472.

DOLTON, P. and D. O'NEILL (1996), "Unemployment Duration and the Restart Effect: Some Experimental Evidence", *Economic Journal*, 106(435), 387–400.

DOOLITTLE, F. and L. TRAEGER (1990), *Implementing the National JTPA Study*, New York.

EISSA, N. (1996), "Labor Supply and the Economic Recovery Tax Act of 1981", in: M. Feldstein and J. Poterba, eds., *Empirical Foundations of Household Taxation*, Chicago, 5–32.

FORSLUND, A. and A. KRUEGER (1997), "An Evaluation of the Swedish Active Labor Market Policy", in: R. Freeman, B. Swedenborg and R. Topel, eds., *The Welfare State in Transition*, Chicago, 267–298.

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

HECKMAN, J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47(1), 153–161.

HECKMAN, J. (1996), "Comment.", in: M. Feldstein and J. Poterba, eds., *Empirical Foundations of Household Taxation*, Chicago, 32–38.

HECKMAN, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator", *Journal of Human Resources*, 32(3), 441–461.

HECKMAN, J., N. HOHMANN, and J. SMITH, with M. KHOO (2000), "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment", *Quarterly Journal of Economics*, 115(2), 651–694.

HECKMAN, J. and V. J. HOTZ (1989), "Choosing Among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84(408), 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*, 93(23), 13416–13420.

HECKMAN, J., H. ICHIMURA, J. SMITH, and P. TODD (1998), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66(5), 1017–1098.

HECKMAN, J., H. ICHIMURA and P. TODD (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64(4), 605–654.

HECKMAN, J., R. LALONDE and J. SMITH (1999), "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics Volume 3A*, Amsterdam, 1865–2097.

HECKMAN, J., L. LOCHNER and CH. TABER (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents", *Review of Economic Dynamics*, 1(1), 1–58.

HECKMAN, J. and R. ROBB (1985a), "Alternative Methods for Evaluating the Impact of Interventions" in: J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York, 156–246.

HECKMAN, J. and R. ROBB (1985b), "Alternative Methods for Evaluating the Impact of Interventions: An Overview", *Journal of Econometrics*, 30(1–2), 239–267.

HECKMAN, J. and J. SMITH (1993), "Assessing the Case for Randomized Evaluation of Social Programs.", in: K. Jensen and P. K. Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives*, Copenhagen, 35–96.

HECKMAN, J. and J. SMITH (1995), "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, 9(2), 85–110.

HECKMAN, J. and J. SMITH (1996a), "Experimental and Nonexperimental Evaluation", in: G. Schmid, J. O'Reilly and K. Schömann, eds., *International Handbook of Labour Market Policy and Evaluation*, Brookfield, VT, 37–88.

HECKMAN, J., and J. SMITH (1996b), "Social Experiments: Theory and Evidence", in: *Ökonomie und Gesellschaft, Jahrbuch 13: Experiments in Economics – Experimente in der Ökonomie*, Frankfurt/Main and New York, 186–213.

HECKMAN, J. and J. SMITH (1999), "The Pre-Programme Dip and the Determinants of Participation in a Social Programme: Implications for Simple Program Evaluation Strategies", *Economic Journal*, 109(457), 313–348.

HECKMAN, J. and J. SMITH (2000), "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study", in: D. Blanchflower and R. Freeman, eds., *Youth Employment and Joblessness in Advanced Countries*, Chicago, 331–356.

HECKMAN, J. and J. SMITH, with N. CLEMENTS (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", *Review of Economic Studies*, 64(4),487–537.

HECKMAN, J., J. SMITH and CH. TABER (1998), "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics*, 80(1), 1–14.

IMBENS, G. and J. ANGRIST (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62(4), 467–476.

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

LALONDE, R. and R. MAYNARD (1987), "How Precise Are Evaluations of Employment and Training Programs: Evidence from a Field Experiment", *Evaluation Review*, 11, 428–451.

LECHNER, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification", *Journal of Business and Economic Statistics*, 17, 74–90.

MEYER, B. (1995), "Lessons from the U.S. Unemployment Insurance Experiments", *Journal of Economic Literature*, 33(1), 91–131.

MICHALOPOULOS, CH., D. CARD, L. GENNETIAN, K. HARKNETT and PH. ROBINS (2000), *The Self-Sufficiency Project at 36 Months: Effects of a Financial Work Incen-*

tive on Employment and Income, Ottawa: Social Research and Demonstration Corporation.

MOFFITT, R. (1991), "Program Evaluation with Nonexperimental Data", *Evaluation Review*, 15(3), 291–314.

RAAUM, O. and H. TORP (2001), "Labour Market Training in Norway – Effect on Earnings", *Labour Economics*, Forthcoming.

REGNÉR, H. (2001), "A Nonexperimental Evaluation of Training Programs for the Unemployed in Sweden", *Labour Economics*, Forthcoming.

ROSENBAUM, P. and D. RUBIN (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70(1), 41–55.

RUBIN, D. and N. THOMAS (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates", *Journal of the American Statistical Association*, 95(450), 573–585.

SMITH, J. (1987), "Measuring Earnings Levels Among the Poor: Evidence from Two Samples of JTPA Eligibles", Unpublished manuscript, University of Western Ontario.

SMITH, J. and P. TODD (2000), "Is Propensity Score Matching the Answer to LaLonde's Critique of Nonexperimental Estimators?", Unpublished manuscript, University of Western Ontario.

TORP, H., O. RAAUM, E. HERNÆS and H. GOLDSTEIN (1993), "The First Norwegian Experiment" in: K. Jensen and P. K. Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives*, Copenhagen, 97–140.

WHITE, M. and J. LAKEY (1992), *The Restart Effect: Evaluation of a Labour Market Programme for Unemployed People*, London, UK: Policy Studies Institute.

SUMMARY

This paper considers different methods for solving the evaluation problem. I highlight the role of heterogeneity in program impacts in defining evaluation parameters of interest and in interpreting estimated program impacts. I discuss the strengths and weaknesses of social experiments and conclude that they require careful implementation and interpretation. I review and critique two popular non-experimental evaluation methods: difference-in-differences and propensity score matching. I find that the former relies on assumptions at odds with the empirical data and that the latter is not a magical solution to all evaluation problems. Finally, I argue for the importance of paying attention to data quality and general equilibrium effects.

ZUSAMMENFASSUNG

Dieser Aufsatz betrachtet verschiedene Methoden, um das Evaluationsproblem zu lösen. Zunächst wird die Rolle der Heterogenität in den Programmauswirkungen hervor-

gehoben, um die interessierenden Evaluationsparameter zu definieren und um die geschätzten Programmauswirkungen zu interpretieren. In der Folge diskutiere ich die Stärken und Schwächen von sozialen Experimenten und komme zum Schluss, dass sie einer vorsichtigen Anwendung und Interpretation bedürfen. Zwei gängige nicht-experimentelle Evaluationsverfahren, *difference-in-differences* und *propensity score matching*, werden geprüft und kritisiert. Dabei stelle ich fest, dass ersteres auf Annahmen basiert, die nicht mit den empirischen Daten übereinstimmen und dass letzteres keine überzeugende Lösung für alle Evaluationsprobleme darstellt. Schliesslich wird erörtert, wie wichtig es ist, die Datenqualität und generelle Gleichgewichtseffekte zu beachten.

RESUME

Cet article examine différentes méthodes résolvant le problème d'évaluation. D'abord, le rôle de l'hétérogénéité dans les effets du programme est souligné; rôle important pour la définition des paramètres d'évaluation en question et l'interprétation des effets estimés du programme. Ensuite, je discute les avantages et désavantages d'expériences sociales et je conclus qu'elles doivent être appliquées et interprétées avec grand soin. Deux méthodes courantes d'évaluation non expérimentale, le *difference-in-differences* et le *propensity score matching*, sont examinées et critiquées. J'en conclus que la première est basée sur des hypothèses ne concordant pas avec les données empiriques et que la deuxième ne constitue pas une solution convaincante à tous les problèmes d'évaluation. Finalement, j'insiste sur l'importance de la qualité des données et de l'observation d'effets d'équilibre général.