

Evaluating Active Labor Market Policies:
Lessons from North America

Prof. Jeffrey Smith
Department of Economics
University of Western Ontario
and NBER
jsmith@julian.uwo.ca

Version of August 15, 2000

The author thanks Dan Black, Michael Lechner and Miana Plesca for valuable comments.

1. Introduction

It is an exciting time to be involved in evaluation research. 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. There is also tremendous excitement on the policy side, with many governments that hitherto largely avoided serious evaluation of active labor market policies now becoming interested in this area.

This paper presents remarks on two important areas related to the evaluation of active labor market policies. The first, the industrial organization of the evaluation industry, is rarely discussed but nonetheless quite important to the long-term success of evaluation as a guide to policy. Here I describe the evaluation industry in North America and argue that other countries can learn from certain aspects of it.

The second, more traditional topic, concerns recent developments in the econometric methods employed in evaluation. This section builds on work by James Heckman, myself and others summarized and synthesized in Heckman, LaLonde and Smith (1999).

In regard to econometric methods, I make four main points. First, I argue for the importance, both in substantive policy terms and in econometric terms, of taking into account differences in the effects of programs across persons. Second, I discuss both the strengths and limitations of social experimentation, concluding that experimentation represents an important evaluation tool that should neither be summarily dismissed nor uncritically accepted.

Third, I discuss the most recent developments in non-experimental matching methods, which center on the technique of propensity score matching. I emphasize that matching, while highly useful when the data exist to support its use, requires strong assumptions to justify it. It also requires difficult judgements about matching variables on the part of the evaluator that can, in the end, only be resolved by recourse to prior knowledge or to theory. As such, despite the recent stampede toward matching in the applied literature, matching does not represent a “magic bullet” that will solve the evaluation problem in every context.

Finally, 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 means that they will remain controversial in both the academic literature and the policy world. Despite this controversy, evaluators should pay attention to these 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 them.

2. Remarks on the Evaluation Industry and Evaluation Policy

In this section, I describe the organization of the evaluation industry in North America,¹ and offer some remarks as to aspects of it that serve as a model for other regions.

Looking to the North American model makes empirical sense as it is in North America that evaluation research and practice are the most advanced.² At the same time, I want to be clear that evaluation area in North America is not perfect. For example, evaluation research in sociology and psychology lags years (if not decades) behind economics in its statistical methods and related conceptual issues (see, e.g., Rossi and Freeman, 1993). At the same time, economists often pay little attention to important issues of program implementation and outcome measurement emphasized in other disciplines. What I do in this section is emphasize positive aspects of the North American experience that could be carried over to other contexts.

2.1 Who Are the Evaluators?

Four main groups perform most evaluations in North America. Each group has advantages and disadvantages in terms of their typical level of familiarity with the methodological literature, their capacity to get things done in a timely way, and their incentives to promote the interests of the program at the expense of the evaluation.

The first group consists of government employees. Many agencies have staff on-site, often with doctoral training in economics or related disciplines; these researchers perform “in house” evaluations. The advantages here are clear – the staff will be very

¹ I am using “North America” here in the sense in which it is often used in Canada, where it means Canada and the United States but not Mexico.

² See Riddell (1991) for an earlier, comprehensive assessment of the North American evaluation industry, which focuses both on issues of organization and of econometric methodology.

familiar with the programs their agencies operate and with the data the agency collects for evaluation purposes. They can also be closely monitored. The disadvantages are that government researchers, who often have many other responsibilities, may fall behind in terms of evaluation methods. They may also face strong incentives within the agency to produce positive findings.

The second group consists of for-profit consulting firms. These firms range from those devoted almost entirely to evaluation work or related empirical studies, such as Mathematica or Westat, to large consulting firms with small units that do evaluation work, such as Goss Gilroy. To borrow a phrase, these firms do the “20 percent that gets you 80 percent of the way.” Concerns about making a profit tend to rule out long excursions down methodological side streets. At the same time, the bottom line also means that the work generally gets done on time, in a professional way, and is packaged in a way that policymakers can easily use and understand. To the extent that the same firm does repeat business with a given agency, and that agency rewards positive evaluation findings, there can be incentive problems here, particularly if the agency constitutes a large part of the firm’s business.

Non-profit consulting firms, such as the Manpower Demonstration Research Corporation (MDRC) in the U.S. and the Social Research and Demonstration Corporation (SRDC), its Canadian offspring, generally resemble for-profit firms, but with some differences in emphasis. Non-profit firms will often take the lead in developing new policy ideas and then looking for funding sources willing to pay for a demonstration project and related evaluation. Non-profit firms, at least those at the high end such as MDRC, tend to have more of an academic culture. As these firms often rely relatively

more on foundations for funding, their incentive problems tend to be less than those of for-profit firms. On the other hand, at the low end of this group, the low salaries will sometimes mean that positions get filled with advocates more concerned with particular policy outcomes than with whether a program actually works.

The final group performing evaluations consists of academic researchers, predominately scholars in economics, sociology and psychology. Relative to the other groups, academics tend to be closer to the methodological cutting edge. At the same time, they are notorious for not getting things done on time and for sometimes producing reports that make great scholarly reading but are too technical for their policymaker clients. Academics also have the smallest incentive problem, as most earn only a small fraction of their income from evaluation consulting.

In many large-scale evaluations, the different groups will collaborate. For example, in the US National JTPA Study, one for-profit firm, Abt Associates, combined with two non-profit firms, MDRC and the National Opinion Research Center (NORC) and top academic researchers to perform the evaluation. Common practice in Canada is for both for-profit and non-profit firms to do evaluations on contract to the government and then to subcontract with academic researchers to provide advice on the evaluation design and the interpretation of the results. My impression is that these collaborations often combine the virtues of the various actors while partly canceling out their various vices, and so should be encouraged in practice.

2.2 Who Pays for the Evaluations?

In North America, in addition to multiple groups performing the evaluations, there are also multiple groups funding them. The largest source of funding is, of course, the government agencies that run the programs. Even here, however, the federal structures of both the U.S. and Canada mean the same program can be, and sometimes is, evaluated at both the federal and state or provincial level. This is particularly common in the case of labor market programs, as these tend to be run jointly by the federal government and the states or provinces in both countries.

The second source of evaluation funding consists of government research organizations, operating separately from the agencies that run the programs. Examples include the U.S. National Science Foundation (NSF) and the Social Science and Humanities Research Council of Canada. These agencies tend to fund methodological research, but in an evaluation context, methodological research often has immediate and important practical implications. For example, the methodological research by James Heckman and various co-authors using the data from the U.S. National Job Training Partnership Act (JTPA) Study was funded in part by the NSF. In addition to its methodological aspects, this work has affected how policymakers view the results from the JTPA experiment by showing the sensitivity of the experimental estimates (Heckman and Smith, 2000) and highlighting the important role of control group substitution into other training programs (Heckman, Hohmann, Smith and Khoo, 2000). I will have occasion to cite this work many times in Section 3.

The final source of funding is private foundations. Some foundations take an interest in the development of specific types of programs – for example, early education

programs for children in low-income families. These foundations then fund evaluation work designed to build up an evaluation record that in turn provides a base of evidence that supports expansion of the programs.

The existence of multiple sources of funding, some of which are not tied to the agencies that operate active labor market programs, is crucial for the emergence of dissenting views about program efficacy. The existence of funding sources with no direct interest in the survival of particular programs means that researchers need not fear that negative findings will result in a loss of access to future evaluation work.

2.3. Who Evaluates the Evaluators?

An important practical aspect of evaluation policy is quality control. Who or what makes sure that evaluators ask the correct questions and use the best methods to answer them?

This quality control has two aspects. One is simply knowledge. How does whoever funds the evaluation know whether they have received a good evaluation or a poor one?

The second aspect concerns the incentive problems alluded to above. Which institutions or individuals work against the incentive that is often present to slant an evaluation in favor of the program being evaluated? These are important questions that have received too little attention from researchers. I offer only a couple of practical observations here.

The first line of quality control is of course the staff of the government agencies that commission most evaluations. In my experience, the presence of staff with sufficient expertise to both help shape the evaluation agenda of the agency and to evaluate the evaluations it receives from its contractors vastly increases the quality and policy value of the evaluations that get done. Thus, while I think it is a bad idea for agencies to do most

of their evaluation work in house, I think it is very important for them to have persons on staff with the technical and management skills to direct evaluation work done on contract.

The academic world also plays several important roles in the quality control process. First, academics may be asked to review work done by profit or non-profit firms on behalf of a government agency. Second, most of those who do evaluation work in government or at firms are trained in academia. The values they inherit there – ideally including a strong devotion to empirical truth – act as an internal quality control mechanism that they carry with them in their work.

Third, in North America there is a fair amount of movement between academia, government and consulting. This includes, for example, the high level position at the U.S. Department of Labor recently occupied by, among others, Lawrence Katz of Harvard, Alan Krueger of Princeton and Harry Holzer of Michigan State. At a more mundane level, economists often hold positions in both worlds, such as those who have joint appointments at the Rand Corporation and the UCLA economics department. Individuals in government or in the consulting world who hope to return to academia, or who already spend part of their time there, have a strong incentive to maintain the academic quality of their work. In addition, movement of individuals between sectors quickens the pace at which methodological developments get put into evaluation practice. The North American experience suggests the value of having researchers move between academia, government and the consulting world, and in particular the value of bringing leading academics (temporarily) into research management and policy roles in government.

Finally, because many evaluation researchers in government and consulting either desire to keep in contact with the academic world or eventually return to it, their efforts to publish their work in scholarly journals also aids in quality control. Economists at the better consulting firms routinely publish the results of their evaluations in academic journals, in addition to whatever report gets written for the agency funding the evaluation. This desire to remain a part of the academic community, and to seek approval of their work through publication, helps to counterbalance the incentives otherwise present to simply give the agency the answers that it wants to hear. Governments can enhance the incentives provided by scholarly publication by including funds in their evaluation contracts for researchers to write up their results for publication in scholarly journals.

2.4 The Importance of Data

A recent theme in the scholarly literature on evaluation has been the importance of good data (see, e.g., Heckman, Ichimura, Smith and Todd, 1998). Indeed, a strong case could be made that relatively too much time has been spent worrying about estimator choice and relatively too little time worrying about the quality of the data to which the estimator is applied. The best econometric methods will not make up for poor data.

There are many things that government agencies can do to improve the quality and amount of data available for evaluation research. First, the data from evaluations should routinely be made available to the research community for re-analysis and for additional analyses addressing other evaluation questions. Funding for the preparation of public use files should therefore be incorporated in evaluation contracts. Data from past

evaluations in the U.S. that have been made available to the research community, such as the National Supported Work (NSW) Demonstration and the various MDRC work-welfare experiments, have provided the foundation for important new substantive and methodological findings. Of course, adequate precautions related to respondent consent and confidentiality must always be taken.

Second, governments can make administrative data collected as part of program operations available, again with appropriate privacy precautions, for evaluation research. To make the data most useful, data collection procedures for the administrative data should be designed with evaluation uses in mind. This may mean collecting a few additional variables (e.g., education is often missing from administrative data sets) or designing data entry procedures to reduce the number of missing and invalid values of key variables. In North America, administrative data have been used for evaluation purposes without additional data collection, and they have also been merged with survey data to provide longitudinal data on earnings and employment that would be difficult and expensive to obtain through surveys. Indeed, the only really solid estimates that we have regarding the long-term impacts (four or more years after treatment) of employment and training programs rely on matched administrative data – see Couch (1992) for the NSW demonstration and U.S. General Accounting Office (1996) for the JTPA experiment.

In addition to administrative data and survey data collected specifically for particular evaluations, general panel data sets have played an important role in evaluation research in the United States. The archetypes of the panel data sets I have in mind here are the U.S. National Longitudinal Surveys. For example, the National Longitudinal Survey of Youth, which also collects information on the children of its respondents,

provides the data for what is probably the best existing analysis of the U.S. Head Start program for disadvantaged children of pre-school age (see Currie and Thomas, 1995). Based on the volume of research that uses them, my impression is that panel data sets of this type are among the best bargains available in social science research.

3. Remarks on Recent Advances in Evaluation Methods

In this section, I remark briefly on several important recent themes in the literature on evaluation methods. In each case, I introduce the main issues and provide pointers to the recent literature for readers interested in a more detailed discussion.

3.1 Heterogeneity

A large amount of conceptual progress in the evaluation literature has resulted from thinking carefully and formally about models in which the impacts of programs differ across persons. In particular, thinking about the evaluation problem in the context of heterogeneous impacts makes it clear that there is not 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.

To see this more clearly, consider some very simple notation. Let Y_1 denote the outcome a person receives following participation in a program, should he or she participate in it. This outcome could represent 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. Of course, 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 .

We can now think about different possibilities for how the impact of a program varies among persons. In the simplest world, there is no variation, and the program has the same effect on everyone. In the above notation, $\Delta_i = \Delta$ for all persons i . While unlikely to hold in a literal sense, this “common effect” assumption may be good enough in some contexts (and may be a very poor approximation in others). It is this assumption that has largely guided the econometric and applied literatures on program evaluation in the past.

In a slightly more general world, the impact of treatment 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, impacts vary across persons and either the person or program staff or both have some information about its value 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 impact of treatment on the treated lies below the 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 treatment effects, 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 it might seem. In a world of heterogeneous 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.

3.2 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 or 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. Note that the creation of an experimental control group differs from the creation of the typical non-experimental comparison 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 what would have

happened to the participants, had they not participated, 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 consistent 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.³

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

³ 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.

skepticism about non-experimental methods. In these evaluations, different researchers using the same data set came to dramatically different conclusions about program effectiveness.⁴ 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 both non-experimental estimators themselves 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

⁴ 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).

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.⁵ 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 3.3. Heckman, Ichimura, Smith and Todd (1996,1998) and Heckman, Ichimura and Todd (1997) use the 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.⁶

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. The issues considered here have become better understood in the North American literature over the past decade, but it is important that they influence choices about experimental designs in Europe as well.

To begin with, social experiments cannot estimate all parameters of interest. This limitation has several dimensions. First, some “treatments” (broadly defined), such as

⁵ See the related analyses in Fraker and Maynard (1987) and LaLonde and Maynard (1987).

⁶ 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.

sex or family income while young, defy random assignment. Second, with some exceptions, social experiments are well suited to estimate the impact of treatment on the treated but poorly suited to estimate general equilibrium effects on persons not randomly assigned. I discuss these general equilibrium effects in more detail in Section 3.4. 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. 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 different 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, ongoing 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 (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. Heckman, Ichimura, Smith and Todd (1998) present evidence that these factors make a big difference to the amount of bias in non-experimental evaluations.

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 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 complicated. 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, 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 correct 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.”

3.3 Matching

Non-experimental research evaluating social programs over the past few decades has witnessed the rise and fall of various preferred estimation strategies. Two decades ago, the Heckman (1979) bivariate normal selection model dominated the literature. This model attempts to account for selection on “unobservables.” That is, it takes account of the fact that variables unobserved by the econometrician may affect participation in the program as well as the outcome of interest. Ten years ago, selection on unobservables remained a concern, but the method of choice was difference-in-differences, which assumed selection into programs based on a fixed, unobserved component of outcomes. In the past few years, perhaps because of the availability of richer data, selection on unobservables has diminished as a concern with the rise of interest in matching methods for program evaluation.

Matching methods are not new, even to this literature. 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 3.1). 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 running 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 this non-linearity 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.⁷ In this case, 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 for 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 for some participants, an impact estimate cannot be obtained for these participants. In this case, if impacts vary across persons as described in Section 3.1, 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)

⁷ 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

make the problem clear.⁸ 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 algorithm, other than comparing the resulting estimates to those from an experiment, exists to guide the researcher in making this decision.⁹ 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 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

the support issue and ways of dealing with it.

⁸ See Figure 2 in the first case and Figures 1 and 2 in the second case.

⁹ 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).

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 consistent¹⁰ 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

¹⁰ 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.

as the experimental impact estimate. In contrast, Dehejia and Wahba (1998,1999) use the 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.

3.4 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* effects¹¹ 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

¹¹ These substitution effects differ conceptually from those discussed in the context of social experiments, despite the similar terminology.

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 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

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.

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, 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 impact – 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 3.1, the researcher will also be

interested in the effect of the program on non-participants. This impact on non-participants may be decomposed, 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 3.1, 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.

4. Conclusions

Many countries that have historically not placed much emphasis on evaluation have now begun to do so. In the first part of this discussion, I describe the evaluation industry in North America, the region where evaluation practice and data collection are both most advanced. I highlight those features of the industry that play key roles in maintaining the quality of evaluation practice and which therefore are good candidates for adoption elsewhere.

Even in North America, the rapid methodological developments in evaluation research in the past two decades have sometimes outpaced both government data collection efforts and evaluation practice. In the second part of this discussion, I provide an informal overview of recent methodological developments and their implications for evaluation practice and policy. I also provide copious citations to the technical literature related to these developments. The main conclusion here is that there remains much room for improvement in evaluation practice.

REFERENCES

- Angrist, Joshua and Alan Krueger. 1999. "Empirical Strategies in Labor Economics." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics Volume 3A*. Amsterdam: North-Holland. 1277-1366.
- Barnow, Burt. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources*. 22. 157-193.
- Björklund, Anders and Håkan Regnér. 1996. "Experimental Evaluation of European Labour Market Policy." In Günther Schmid, Jacqueline O'Reilly and Klaus Schömann, eds., *International Handbook of Labour Market Policy and Evaluation*. Brookfield, VT: Edward Elgar. 89-114.
- Black, Dan, Jeffrey Smith, Mark Berger and Brett 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, Howard. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*. 82(2). 225-246.
- Bloom, Howard, Larry Orr, Stephen Bell, George Cave, Fred Doolittle, Winston Lin and Johannes 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, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives*. 9(2). 63-84.
- Burtless, Gary and Larry Orr. 1996. "Are Classical Experiments Needed for Manpower Policy." *Journal of Human Resources*. 21. 606-639.
- Calmfors, Lars. 1994. "Active Labor Market Policy and Unemployment – A Framework for the Analysis of Crucial Design Features." *OECD Economic Studies*. 22(1). 7-47.
- Couch, Kenneth. 1992. "New Evidence on the Long-term Effects of Employment and Training Programs." *Journal of Labor Economics*. 10(4). 380-388.
- Currie, Janet and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review*. 85(3). 341-364.
- Davidson, Carl and Stephen Woodbury. 1993. "The Displacement Effects of Reemployment Bonus Programs." *Journal of Labor Economics*. 11(4). 575-605.

Dehejia, Rajeev and Sadek Wahba. 1998. "Propensity Score Matching Methods for Non-Experimental Causal Studies." NBER Working Paper #6829.

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

Dickinson, Kathryn, Terry Johnson and Richard West. 1987. "An Analysis of the Sensitivity of Quasi-Experimental Net Estimates of CETA Programs." *Evaluation Review*. 11. 452-472.

Dolton, Peter and Donal O'Neill. 1996. "Unemployment Duration and the Restart Effect: Some Experimental Evidence." *Economic Journal*. 106(435). 387-400.

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

Forslund, Anders and Alan Krueger. 1997. "An Evaluation of the Swedish Active Labor Market Policy." In Richard Freeman, Birgitta Swedenborg and Robert Topel, eds., *The Welfare State in Transition*. Chicago: University of Chicago Press. 267-298.

Fraker, Thomas and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*. 22(2). 194-227.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*. 47(1). 153-161.

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

Heckman, James, Neil Hohmann, and Jeffrey Smith, with Michael 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, James and V. Joseph 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, James, Hidehiko Ichimura, Jeffrey Smith, and Petra 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, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*. 66(5). 1017-1098.
- Heckman, James, Hidehiko Ichimura and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*. 64(4). 605-654.
- Heckman, James, Robert LaLonde and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics Volume 3A*. Amsterdam: North-Holland. 1865-2097.
- Heckman, James, Lance Lochner and Christopher 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, James and Jeffrey Smith. 1993. "Assessing the Case for Randomized Evaluation of Social Programs." In Karsten Jensen and Per Kongshoj Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives*. Copenhagen: Danish Ministry of Labour. 35-96.
- Heckman, James and Jeffrey Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*. 9(2). 85-110.
- Heckman, James, and Jeffrey Smith. 1996a. "Experimental and Nonexperimental Evaluation." In Günther Schmid, Jacqueline O'Reilly and Klaus Schömann, eds., *International Handbook of Labour Market Policy and Evaluation*. Brookfield, VT: Edward Elgar. 37-88.
- Heckman, James, and Jeffrey 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: Campus Verlag. 186-213.
- Heckman, James and Jeffrey Smith. 2000. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study." In David Blanchflower and Richard Freeman, eds., *Youth Employment and Joblessness in Advanced Countries*. Chicago: University of Chicago Press for NBER. 331-356.
- Heckman, James and Jeffrey Smith, with Nancy 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, James, Jeffrey Smith and Christopher Taber. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics*. 80(1). 1-14.

- Imbens, Guido and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*. 62(4). 467-476.
- LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*. 76(4). 604-620.
- LaLonde, Robert and Rebecca Maynard. 1987. "How Precise Are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Review*. 11. 428-451.
- Lechner, Michael. 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, Bruce. 1995. "Lessons from the U.S. Unemployment Insurance Experiments." *Journal of Economic Literature*. 33(1). 91-131.
- Michalopoulos, Charles, David Card, Lisa Gennetian, Kristen Harknett and Philip Robins. 2000. *The Self-Sufficiency Project at 36 Months: Effects of a Financial Work Incentive on Employment and Income*. Ottawa: Social Research and Demonstration Corporation.
- Raaum, Oddbjørn and Hege Torp. 2001. "Labour Market Training in Norway – Effect on Earnings." *Labour Economics*. Forthcoming.
- Regné, Håkan. 2001. "A Nonexperimental Evaluation of Training Programs for the Unemployed in Sweden." *Labour Economics*. Forthcoming.
- Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*. 70(1). 41-55.
- Rossi, Peter and Howard Freeman. 1993. *Evaluation: A Systematic Approach*. 5th Edition. Newbury Park, CA: Sage.
- Torp, Hege, Oddbjørn Raaum, Erik Hernæs and Harald Goldstein. 1993. "The First Norwegian Experiment." In Karsten Jensen and Per Kongshoj Madsen, eds., *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives*. Copenhagen: Danish Ministry of Labour. 97-140.
- Riddell, Craig. 1991. "Evaluation of Manpower and Training Programmes: The North American Experience." In OECD, ed., *Evaluating Labor Market and Social Programs*. Paris: OECD. 43-72.
- Smith, Jeffrey and Petra Todd. 2000. "Is Propensity Score Matching the Answer to LaLonde's Critique of Nonexperimental Estimators?" Unpublished manuscript, University of Western Ontario.

U.S. General Accounting Office. 1996. *Job Training Partnership Act: Long-Term Earnings and Employment Outcomes*. Report No. GAO/HEHE-96-40.

White, Michael and Jane Lakey. 1992. *The Restart Effect: Evaluation of a Labour Market Programme for Unemployed People*. London, UK: Policy Studies Institute.