

Improving the Evaluation of Employment and Training Programs in Canada

Jeffrey Smith
Department of Economics
University of Maryland
smith@econ.umd.edu

Arthur Sweetman
School of Policy Studies
Queen's University
sweetman@qsilver.queensu.ca

Version of November 8, 2001

Both authors thank Bill Warburton of the BC Ministry of Social Development and Economic Security for helpful discussions on the issues covered in this paper. Smith thanks the members of the HRDC Medium Term Indicators Advisory Panel, in particular Walter Nicholson, for useful discussions as well.

© Copyright Jeffrey Smith and Arthur Sweetman, 2001. Used with permission by the Human Resources Development Canada (HRDC) Conference on Evaluation Methodologies, in Ottawa, November 16 and 17, 2001.

1. Introduction

HRDC and the provincial governments spend a lot of money on employment, training and other related programs. It is always, especially in the current fiscal climate, the government's responsibility to ensure that this money is well spent since there are always high quality alternative uses of the funds. For some, determining that program funds are well spent involves what is sometimes called a "process evaluation" or an audit. Was the money allocated to a particular project actually spent on that project? Were the correct procedures followed? These, and similar, questions are posed and answered in such an exercise.

Process evaluations do not, however, usually answer another set of questions, on which we focus here. Does the program help, in some suitably defined sense, those who participate? That is, does it improve some outcome (e.g. earnings, or the risk of unemployment) relative to what it would have been without participation in the program? Does the program pay for itself? Or, the very hard question: Does the program help (or hurt) those not in the program? These questions are usually considered to be part of an "impact" evaluation. What is the "impact" of the program or treatment, including unintended consequences such as displacement of persons not on the program or "eligibility effects" wherein some persons take undesired actions in order to become eligible for services?

While the state of evaluation in Canada is strong on some dimensions, it is weak on others. Canada can rightly take pride in the high-quality experimental evaluation of the Self-Sufficiency Project (see Michalopoulos, et al., 2000), funded by HRDC and implemented by the Social Demonstration and Research Corporation. This evaluation has affected the policy debate not only in Canada but in the United States and Britain as well. It provides important evidence on the responsiveness of labor supply in low-skill populations that are sometimes

characterized as being unable to find and keep full-time work. Canada (and HRDC in particular) can take pride in its path-breaking efforts to develop statistical models to aid in service assignment in the form of the Service Outcomes and Measurement System (SOMS) and to incorporate a comparison group into its performance management system through the Medium Term Indicators (MTI) project. A large number of high quality evaluations can be found on the HRDC Evaluation and Data Development (EDD) web page.

At the same time, much work remains to be done. Canada lags behind the U.S. in the broad utilization of random assignment methods. Program implementation and data collection at both the federal and provincial levels often makes the conduct of high quality non-experimental evaluations difficult or impossible. Evaluation results are often buried in government reports rather than being publicized and subject to peer review at scholarly journals. Evaluations are rarely, if ever, subject to replication by outside researchers. We make proposals throughout the paper regarding ways in which Canada can improve the integrity of its evaluation process, which will lead to improvements in policymaking based on sound evidence.

The field of program evaluation is rapidly advancing in terms of technique (see, e.g., Heckman, LaLonde and Smith, 1999, and Angrist and Krueger, 1999), and modern computers and data sets are allowing the potential for much higher quality than was possible even a few years ago. In this paper, we provide a short tour of some of the key issues in impact evaluation, and then discuss the importance of making evaluation a key element in the planning and development of new programs and reforms of existing programs. Aside from allowing for good decision-making, incorporating evaluation into the design process also provides appropriate incentives for policymakers because they know that their work will

have to pass a test of sorts. (Of course, program design and summative evaluation should usually be conducted by different parties to ensure that there is no conflict of interest. This does not rule out ongoing internal monitoring and evaluation to inform program operators). An institutional culture of serious evaluation raises the political costs of inefficient allocation of public funds. More efficient allocation of public funds is good news for taxpayers, who either get some of their hard-earned money back or see it spent on better programs, and for program participants, who receive a larger benefit from the time and energy they spend receiving services.

It is possible to have evidence-based policymaking, at least some of the time. A number of solid empirical studies guided the last unemployment insurance reform in Canada. In the U.S., when the experimental evaluation of the Job Training Partnership Act program found that it had no impact on youth, the youth component of the program had its budget cut by over eighty percent. Evidence and efficiency will always conflict with ideology and rent seeking (on at least one side of an issue) in debates over policy. One important goal of this paper is to outline ways in which to strengthen the evidence and efficiency side of the conflict.

The remainder of the paper proceeds as follows. In Section 2, we introduce the idea that programs may have different impacts on different persons, and show that this has important implications for the meaning of the impact estimates obtained from evaluations. In Section 3, we discuss random assignment methods for program evaluation, while in Section 4 we review recent thinking on the non-experimental alternatives to random assignment. Section 5 discusses the often-ignored issue of general equilibrium effects—that is, program effects on persons other than program participants. Section 6 highlights the importance of

data quality, particularly for non-experimental evaluation, while Section 7 discusses methods for improving evaluation quality through institutional changes in the ways in which evaluations are conducted. In Section 8, we emphasize the importance of cost-benefit analysis and discuss recent advances in the way in which such analyses are conducted for employment and training programs. Finally, Section 9 summarizes our main points and lays our conclusions regarding how to improve evaluation in Canada.

2. Integrity and Interpretation

Integrity is a key element of any evaluation. Perhaps most importantly this includes the integrity of the evaluation process, including for example, the nature and scope of the questions asked, and the use to which findings are put. The integrity of the researcher is also important, since, as John Cragg noted in his 1994 Innis Lecture to the Canadian Economic Association: a skilled researcher can take the data into the basement and beat a variety of “truths” out of it.

The aspect of integrity on which we focus in this section is that of interpretation, which is foundational. In recent years there has been a large research effort to try to decrease the gap between the parameters that are estimated in evaluations and the interpretation of those parameters. This effort has two aspects. The first seeks to clarify exactly what policy question or questions one is trying answer in an evaluation. For example, do we want to know if we should eliminate the program, expand it a little bit in a certain way, or make it mandatory for everyone in the eligible population? The second aspect seeks to clarify the links between the estimates provided by particular evaluation strategies and specific policy

questions. In particular, this second aspect clarifies the set of policy questions that are *not* answered by particular empirical results – that is, the limits of interpretation.

This focus on the meaning of estimates arises from a realization that we live in a world in which the impacts of programs vary across individuals (see, e.g., the discussion and evidence in Heckman, Smith and Clements, 1997, and Heckman and Smith, 1998). In such a world, the impact of a program on those it presently serves may not be a good guide to its impact on the new participants it would acquire if it expanded or was made mandatory. In such a world, the fact that different econometric evaluation procedures (different ways of comparing participants and non-participants) estimate different parameters has crucial implications for the meaning of those estimates, and for their link to the policy questions of interest. Thinking about evaluation in a world where impacts vary across persons also means that more effort must be put into the design and implementation of both programs and their evaluations, in order to insure that the answers generated by an evaluation correspond to the questions policymakers care about.

2.1 Alternative Parameters of Interest

An (overly simple) example illustrates the issues. Consider a mammography-screening program. The incidence of breast cancer is a function of several risk factors; in this case we consider only one, age, as depicted (somewhat hypothetically) in Figure 1. The curve in this case represents the true underlying incidence, the x-axis is age, and the y-axis is the incidence of detection (for this simple example, we assume perfect screening so that we can ignoring problems of false negatives, false positives etc.).

Suppose that a health program offers screening to women over the age of 45, and that the true shape of the incidence curve, shown in figure 1, is unknown to those managing the program. The policy issue here is that it is good to find breast cancer early, both for the person who has it and the public that finances the health care system. At the same time, the tests are costly in terms of both time and resources, and may have negative side effects on those who receive them. As a result, it is not optimal to screen populations wherein the incidence rate is below a certain level.

All that analysts observe is the fraction of those tested who have a positive test result – which is the impact treatment on the treated (TT) in this context. Analysts calculate that, based on the TT, it is well worth expanding the program. The program is then expanded to include those aged 40 and over. However, it is quickly observed after the policy change that the average detection rate falls. An analyst looks at the new data and observes the previously unrealized relationship between age and detection rates (although the analyst sees data only for those beyond age 40); she calculates the incidence rate for the marginal group, those aged 40 to 45. The incidence rate for this marginal group is the local average treatment effect (LATE) for that group (see Imbens and Angrist, 1994). Other groups, such as women aged 35 to 39, would have their own (different) LATE. A comparison of the estimate of the LATE to the estimate of the TT reveals that the TT was not the best parameter on which to base a decision about expanding the program. Rather, an estimate of the LATE should have been used. In contrast, TT is the right parameter to look at if the policy of interest is whether to keep or scrap the current program that screens only women aged 45 and older. Along the same lines, if a proposal to make the program universal was entertained, then the average treatment effect (ATE) for the entire population is the relevant parameter for the cost-benefit

analysis. The general lesson is that different parameters answer different policy questions and require different evaluation designs.

Of course, the shape of the incidence curve in this case is well understood today (although it was not always so), and there are many issues (like the stress associated with false positive tests) that need to be considered if decisions were really to be made about such a program. Nevertheless, this illustration emphasizes the variety of impact parameters associated with a program.

Even this simple example becomes somewhat more complex if we consider that risk factors in addition to age might have been used as criteria for entrance into the program. Suppose that a medical officer had suggested that, rather than the minimum age being decreased for all women, that only younger women with particular family medical histories be screened. This would imply a different source of variation in the data, and a different LATE, one specific to the new policy change. In general, each source of variation in the data (each risk factor, or set of risk factors, used to determine eligibility for screening) generates a different LATE. Further, the example assumes that program planners observe all of the risk factors. This is rarely the case in practice. Rather, if selection into the program is voluntary, and not all risk factors are observed, then estimating program impacts becomes very difficult since people who will benefit the most (or are aware that they will benefit the most) will likely self-select into the treatment.

In the simple example $TT > LATE > ATE$. Although some simple economics (people are more likely to participate when they benefit more) suggests that this relationship should be a common one, it need not hold in general. Trivially, if the treatment were the distribution of lottery tickets to some fraction of the population, with each ticket having the

same chance of winning, then $TT = LATE = ATE$. A more policy relevant example where this ranking may be reversed relates to the economic return to high school completion. Card (2001) argues that high school dropouts have a higher economic return to high school completion than do those who complete high school. If a program were devised to keep those students in school (perhaps via an increase to the compulsory schooling age) then the LATE for this program would exceed the TT.

This education example also emphasizes the distinction between an outcome and an impact. Those who drop out of high school have, on average, worse labor market outcomes, for example wages and employment levels, than those who complete high school (and often proceed to post-secondary education). However, the impact — the value added of the education — is higher for the dropouts than for the group with superior outcomes. People who participate in a program may have desirable outcomes, but on its own that says little about the program because that group may have had good outcomes even without it. An impact is the increment, or value added, of a program, not just an outcome. A similar lesson about the difference between outcome levels and program impacts is provided by the small literature that compares program performance standards based on outcome levels to program impacts, typically estimated using randomized experiments. Heckman, Heinrich and Smith (2002) survey this literature and show that the relation between outcome levels and program impacts is often close to zero.

It is easy to understand the distinction between the three alternative parameters, and easy to identify the marginal group, in the mammography example. In many situations the analysis is not so straightforward. In the case of a program expansion, even identifying those who participate in the expanded program, but would not have participated in the pre-

expansion program, can be difficult. Since people generally self-select into government employment and training programs, the impact on those who are treated can be quite different from that on other groups. Since the selection process is not usually understood, the impact of the program is not usually obvious. What is required, then, is careful evaluation, starting from the design and implementation of the program and including precise specification of the evaluation questions of interest and the choice of evaluation strategy that will answer those questions. It is to the set of available evaluation strategies that we now turn our attention.

3. Dealing with Self-Selection into Programs: Random Assignment

3.1. Experiments with Random Assignment to Treatment and Control Groups

In the medical literature the gold standard for estimating the impact of a treatment is an experiment with participants randomly assigned into either a treatment or control group. When performed in the context of social policy, an experiment with random assignment to treatment and control groups is termed a social experiment. Random assignment has been used extensively in the U.S. to evaluate employment and training programs including Supported Work, the Job Training Partnership Act, the Job Corps and numerous state welfare-to-work programs. It is presently being used in the U.K. to estimate the impact of the New Deal for Disabled People.

Relatively few large-scale social experiments have been conducted in Canada; the Mincome experiment in Manitoba in the 1960s, and the Self-Sufficiency Project (SSP) that is now wrapping up, are the two primary examples. Mincome (see Hum and Simpson, 1993) tested the labor market impact of a guaranteed minimum income policy. The SSP (see

Michalopoulos, et al., 2000), conducted in British Columbia and New Brunswick, examined the impact of a generous wage subsidy conditional on full time employment. It produced convincing evidence that when subsidies are generous enough, some income assistance recipients can find and hold full time jobs.

In an experiment, the difference between the treatment and control groups is the impact of the treatment. (Following the literature, we will use “treatment” as a synonym for program participation). This approach (or any impact analysis), obviously, cannot estimate the counterfactual for an individual; rather it estimates the average counterfactual for an entire group. A counterfactual, in this environment, is the value of the outcome that is not observed: treatment for those not treated, and no treatment for those who are treated. The mean outcome in the control group provides an estimate of the mean outcome that treatment group members would have received, had they not received the treatment.

This type of randomized experiment has great power to identify the causal impact of a treatment – in contrast to a simple correlation between the treatment and an outcome, which might be caused by some other factor. Confidence in the result rests on the randomization process. In large samples, randomization ensures that both the observed and (importantly) the unobserved characteristics of the treatment and control groups have the same distributions. As a result, treatment status is uncorrelated with both the observed and unobserved characteristics of persons in the experiment, and a simple difference in means between the treatment and control groups provides an unbiased estimate of the impact of treatment on the treated.

As mentioned, the idea of a social experiment is to use random assignment to ensure that there is no correlation between a person’s characteristics and her or his assignment to

treatment. In contrast, when people are treated because they select into a program it is hard to know what portion of the observed outcome is the result of the impact of the treatment, and what part of the outcome results from the person's characteristics or situation; the value of the treatment is hard to isolate. People who select to undergo a treatment, for example a training program, are different than those who opt not to take the program. Those treated may have, for example, poor outcomes relative to those not treated, but they might have had poor outcomes even in the absence of the treatment. The correlation between average outcome and treatment tells us little about the causal impact of the program in situations with self-selection. If we want to estimate the impact of the program, we need to understand the counterfactual, which is not normally observed.

Social experiments are sometimes viewed with moral apprehension. In the medical context this is sometimes a very serious issue. But, if a society really does not know if a treatment “works” — if a new drug has been developed, looks promising in non-human trials, and has the potential to improve human welfare — some type of test is required to see whether its expected potential is “true”. There must be a weighing of ethical costs and benefits. In the social program context this balancing of costs and benefits is also necessary, although both the costs and benefits usually are of a very different type. If resources are put into programs that do not have an impact (or have a small impact), then those resources are not being put into other programs that might have, or could be demonstrated to have, an impact (or a bigger impact). Alternatively, the government could return the resources devoted to programs that fail to pass cost-benefit tests to the taxpayers.

There is no need, in general, to evaluate programs that we know work, or that are known not to work (although there may be political value to evaluating programs that do not

work but do provide rents to influential groups). Rather, evaluation (as distinct from monitoring) is required in situations where there is genuine uncertainty, which can derive from a lack of, or a lack of confidence in the, existing evidence. If we wrongly believe that a program “works” we may not even look for alternatives. The goal is to design systems that make the best use of the resources available so that society is well served.

While the gold standard, social experiments are not, as Burt Barnow likes to say, a substitute for thinking. For a wide variety of reasons, the results of randomized trials may not provide evidence that is relevant for the policy question of interest, or may require additional interpretation or adjustment beyond a simple comparison of means. First, when treatment group members drop out of the program or control group members find close substitutes to the experimental treatment from other sources, the difference in outcomes between the treatment and control groups no longer estimates the impact of treatment on the treated. Instead, it estimates a local average treatment effect – the mean impact on those persons who participate in the treatment group but not in the control group. Heckman, LaLonde and Smith (1999), Heckman, Smith and Taber (1998) and Heckman, Hohmann, Smith and Khoo (2000) discuss these issues in more detail and present evidence on the empirical importance of treatment group dropout and control group substitution. Second, experiments may require the program to dig deeper into its eligible population than it normally would in order to fill up the control group. In this case, the experimental impact estimate does not provide an estimate of the mean impact of the program on those it normally serves in the absence of the experiment, which is likely to be the parameter of interest. Finally, as we discuss in Section 5, experiments do not capture any general equilibrium

effects of a program and may, indeed, provide a biased estimate of the mean impact of treatment on the treated due to general equilibrium effects on the control group.

In general, traditional social experiments are clearly wrong for many policy questions where other evaluation methods are more appropriate. A randomized trial like the SSP is very expensive, and this makes it infeasible in situations that involve smaller and less costly policy issues. In some cases, ethical considerations forbid randomization. Moreover, unlike pharmaceutical trials in the medical context, randomization in social experiments cannot be made “blind” or “double blind”. That is, the treatment and control groups, and those delivering services to each, are normally aware of their status, and this may affect their actions and thereby make the program during the experiment operate differently from the program in the absence of an experiment.

3.2. Randomization at the Margin

The concept of random assignment is, however, not only applicable to large-scale experiments like the SSP. It can sometimes be incorporated into aspects of ongoing programs, or used in the introduction of new programs, in creative ways to generate good evidence on program impacts without the political and budgetary costs of large-scale social experiments. These costs can include potential bad press as well as legal requirements not to deny service to anyone. Understanding the concept of random assignment, and looking for occasions to implement it creatively, can produce large gains in useful information at low cost. This can be a very fruitful avenue for evaluation; in some cases it is superior to a traditional experiment.

Black, Smith, Berger and Noel (2001) provide an excellent example of the creative use of randomization from the state of Kentucky. Kentucky employs a statistical model to estimate the expected duration of each new unemployment insurance (UI) claim as a function of the claimant's characteristics and his or her local economic characteristics. The state then converts this into a score between one and twenty, with twenty indicating benefit exhaustion and one indicating a very short predicted duration. In each local UI office in each week, new UI claimants are assigned to receive mandatory employment and training services based on their score. Assignment starts with the high scores and proceeds until the number of slots available in that office in that week has been filled. For the marginal score (the one where the slots run out), the number of claimants typically exceeds the number of remaining slots; it is within this marginal office-week-score cell that random assignment takes place. Put differently, random assignment serves to allocate the remaining slots among claimants with the lowest score that receives services in each office in each week.

The benefits to this approach are several. First, it was cheap to implement and invisible to program case workers. Second, it allowed the state to treat everyone with the highest score, which it wanted to do. Third, it was perceived as fair by program staff. Fourth, and most important, it provides real experimental data, at least for the marginal group. The disadvantage, of course, is that the experimental impact estimate is an estimate of a local average treatment effect, not an estimate of the impact of treatment on the treated. Overall, Kentucky's clever use of random assignment provided a high quality evaluation and illustrates the important point that you can benefit from randomization even without running a traditional social experiment.

3.3. Other Variants of Random Assignment

Two other variants of random assignment also offer attractive alternatives that retain some of the benefits of traditional social experiments while overcoming some of their political, ethical, and technical problems. The first variant employs random assignment to allocate program participants among alternative treatment regimes, rather than to a treatment group and an untreated control group. Ideally, one treatment consists of low-intensity services, such as job search assistance, which can serve as a baseline treatment against which the impact of the more intensive treatments is measured. In this scheme, no participants go without some treatment, which sidelines most ethical objections to random assignment.

The second variant, which has been used in medical contexts but never in a social experiment, allows the program to operate as normal, but randomly assigns additional services designed to increase the participation rate. This variant will work best in programs with fairly high dropout rates prior to major service receipt and in cases where there is some prior knowledge of activities that will, in fact, induce additional participation in the program. The advantages and disadvantages here are similar to the first variant. There is little ethical objection to random assignment of outreach activities, but the resulting estimate corresponds to a local average treatment effect rather than to the impact of treatment on the treated.

Other ways of relatively innocuously introducing aspects of randomization are also feasible, and can facilitate valuable and credible evaluations. For example, if some aspect of a program is to be phased in across regions, then the start dates for each region can be assigned randomly. An exception to this occurs for small sample sizes, for example when there is a small number of regions. Randomization gets its statistical power from large sample sizes where, on average, the characteristics of the treatment and control groups are

approximately the same. For smaller groups it can be better not to use random assignment, but to select a few key characteristics and assign units to the treatment and control group, perhaps as case and matched control pairs, based on matching those characteristics. (This is developed in the medical literature using minimum chi-squared statistical techniques.) If a program is to be phased in anyway, it only makes sense to do it in a manner that maximizes the information that can be gained about the program's operation so that better informed management decisions can be made.

Overall, it is hard to overemphasize the potential to conduct persuasive yet inexpensive evaluations once one understands the concepts associated with randomized trials and thinks about how to implement those concepts creatively. High quality, evidence based, public policy can result.

4. Dealing with Self-Selection into Programs: Non-experimental Methods

4.1 When Everything that Matters is Observed - The Rare Case

Consider the case where self-selection occurs, but the analyst observes all the variables with important effects on both participation and on the outcome of interest in the absence of participation. The econometric literature calls this situation "selection on observables". Estimating the impact of a program on those treated is relatively straightforward in this situation. Matching methods, such as those described in Heckman, Ichimura and Todd (1997), and Smith and Todd (2002) provide consistent estimates of the impact of treatment on the treated. Under some additional assumptions, regression analysis can be used instead, although matching is preferred for statistical reasons when feasible.

Angrist (1998) has done one of the rare studies that falls into this category. It involved the U.S. military; he is interested in the impact of being accepted, or not, into the military on those who apply. In this case all applicants write a series of tests and all decisions are by the military based on the results of the tests. All relevant variables are observed, and the impact of the military's decision can be estimated because the selection process is well understood. This type of situation is *very* rare. Almost all of the time the analyst does not observe some factors that affect both participation and outcomes.

When unobservable factors (often simply called “unobservables”) influence both the decision-making process and outcomes in the absence of participation, statistical or econometric analysis is more difficult. In a regression context, the unobservable variables that matter, because they are not measured, are implicitly included in the error term. This implies that the error term is correlated with the treatment variable. This violates the assumptions of both matching and OLS regression. In order to produce estimates of causal effects, rather than just correlations, both of these methods require that participation be unrelated to the unobservables, conditional on the available observed variables. Experiments work because randomization ensures that both the observable and the unobservable factors that affect outcomes and participation are unrelated to treatment. When everything that matters is observed, it is similarly true that the error term, and the unobservables it embodies, is not correlated with the treatment. Note that this is not a “problem with OLS”, rather it reflects a fundamental logical requirement for identifying a treatment effect.

The selection on observables case can be made less rare through careful data collection and by building up a store of knowledge about the key factors that determine participation in employment and training programs as well as outcomes in the absence of

participation. It also helps when the administrative rules used to select persons into the program are well defined. Effective use of evaluation methods that assume selection on observables likely means employing both survey and administrative data sets. Care must also be taken in the sampling to insure that a sufficient number of non-participants with observable characteristics similar to those of participants appear in the sample. This will typically require stratified sampling of some sort based on prior knowledge of important determinants of participation conditional on program eligibility.

4.2. Instrumental Variables

Instrumental variables (IV) can sometimes provide consistent estimates in contexts where selection into a program occurs on unobservables—on factors not observed by the analyst such as motivation, ability, or a case-worker’s face-to-face evaluation of a client. An instrument is a variable that affects participation in a program, but does not affect the outcome of interest (e.g., earnings) other than through its effect on participation. Intuitively it can be thought of as a type of partial randomization that occurs “naturally”. If we can isolate this source of randomization, we can use it to estimate the impact of the treatment on the outcome of interest. Instruments can vary at the individual level, as when distance to the training provider is used as an instrument, or at an aggregate level, as when variations in rollout dates across jurisdictions serve as an instrument. Random assignment creates an ideal instrument because by construction the treatment variable determines participation but is uncorrelated with everything else. Two-stage least squares is one commonly used IV estimator. The so-called Heckman (1979) two-step estimator (which in fact should be

estimated in one step) is another frequently used IV estimator. It imposes additional assumptions about the precise statistical distributions of the unobserved factors.

Consider, for example, a policy or institutional change that increases (or decreases) the percentage of the potential population that is treated. For example, a one-year increase in the compulsory schooling age causes some people who would otherwise drop out of school to stay in longer. This increase in schooling is, *for the subgroup that is at risk of dropping out*, not subject to individual choice. Instrumental variable techniques can be used to estimate the impact of the extra education for the group affected by the policy change.

Note that the estimate resulting from a compulsory schooling increase needs careful interpretation. If the impact of schooling varies across people, then the estimate is not generalizable. It is the average impact of additional education on those whose level of education is affected by the policy change. If different instruments, that is different sources of variation each of which is not correlated with the unobservables, are employed, then different estimates may be obtained. Each estimate may be “correct” but reflect the value of education, or different aspects of education, to different subgroups of the population. Some of these estimates may answer questions that are relevant for a particular policy; others may not. Credible sources of variation that can be used in instrumental variable estimation may not be very common in many contexts. Like social experiments, advanced statistical or econometric techniques are not excuses for not thinking.

It is important to also notice that the measured impact is an *estimate*. In the instrumental variables case this is particularly important. Since the policy change may affect only a small fraction of the full sample, and since we are only using that part of the decision process that is independent of unobservables, the impact may be imprecisely estimated. This

implies that very large samples may be required to get reasonable precision in some situations. It all depends upon how “powerful” the instrument is — that is, roughly speaking, how much the source of randomness matters in the decision process.

The quality of the estimates obtained from IV methods depends on the quality of the instrument or instruments. A weak or invalid instrument may be worse than no instrument. Good instruments can be obtained in one of three ways: clever data collection, exploitation or creation of useful institutional variation, and randomization as described in Section 3.3. Obtaining good instruments is facilitated by careful planning at the program design and implementation stage (to produce that useful institutional variation) and at the time of evaluation design. It is also incumbent on the evaluator to use standard statistical tools to test the strength and validity of the instruments employed (see, e.g., Bound, Jaeger and Baker, 1995).

As was seen in the case of randomization on the margin, once the nature of instrumental variables estimation is understood, there are numerous opportunities to employ the technique in creative ways. This generates knowledge about the impact, or more correctly impacts, of government programs and can thereby improve public policy.

4.3. Longitudinal Methods

Longitudinal methods offer another way of dealing with situations in which there is selection into programs based on unobserved factors such as ability or motivation. These methods trade off analytic simplicity and additional data requirements for stronger assumptions about the nature of the unobservable factors than are required by IV methods. Moffitt (1991) provides a very accessible introduction to the econometrics of these methods.

The basic assumption underlying longitudinal methods is that selective differences between participants and non-participants are constant over time. If they are constant, then given two or more periods of data, they can be differenced out. The resulting estimator consists of the so-called “difference-in-differences” or “within” estimator. In this estimator, before-after changes in outcomes for participants are compared to before-after changes for non-participants. Two periods of data, one before and one after the period in which individuals decide whether or not to participate, suffice to produce an estimate. With additional periods of data, specification tests such as those employed in Heckman and Hotz (1989) can test the underlying assumption of stability in the unobservables.

The difference-in-differences estimator is widely used in evaluation work, although its performance in cases where it can be compared to experimental impact estimates is not encouraging. Heckman and Hotz (1989) reject it in many cases using data from the U.S. National Supported Work Demonstration. Heckman and Smith (1999) show that it performs very poorly, and is highly sensitive to the chosen before and after periods, using the experimental data from the U.S. National JTPA Study as a benchmark. In both these cases, the variation in participation is at the individual level; applications of difference-in-differences to variation at the subgroup or provincial level may perform better. Overall, we suggest caution in the use of these methods, and testing of their basic assumptions when possible. However, if longitudinal data are available at low cost, they are certainly worth calculating as one among several alternative non-experimental estimates as described in Section 4.5.

HRDC has produced some excellent research using differences-in-difference approaches that address some of the above concerns. For example, its 1997 report (available

on the web page) the “Canada Pension Plan Disability Insurance Benefits and Labour Supply and Well-Being of Older Workers,” which looks at a CPP-disability policy change that did not have a comparable QPP change, uses the QPP region as a comparison group to estimate the policy’s impact.

4.4. Experimental Economics

One way to evaluate programs that do not yet exist (always a problem) is to simulate expected outcomes. This is sometimes done using computer models – for example some HRDC evaluation projects involve microsimulations. Another relatively new and inexpensive approach in the evaluation context is to simulate new proposals by using people, instead of computers, to generate the outcomes. In this case policy proposals are simulated using techniques developed in experimental psychology and experimental economics. One recent such experiment was conducted by the Social Research and Demonstration Corporation (SRDC - 2001) for HRDC to look at the proposed “learn\$ave” program. In this case people from the potential target group for this educational savings program were asked to play a series of “games”, or simulations, using small amounts of real money — the average person gained \$123 and participated for an hour and a half. Ontario Hydro has funded research using this technique to evaluate potential consumer reactions to various changes involving alternative environmental policies.

While this is a relatively new (and relatively untested) approach for program evaluation, it offers some interesting possibilities. Like all of the other techniques mentioned it has limitations, but it may also help us to think carefully about issues that are not well addressed by other means. In our view, these methods merit further exploration, including

use in a context where results from a social experiment are also available, to which they can be compared. Studies that compare impact estimates generated by non-experimental econometric methods to impact estimates from social experiments, such as LaLonde (1986), Heckman and Hotz (1989) and Smith and Todd (2002) have provided important insights into what evaluation techniques work and which don't. Similar evidence is required before we rely too much on methods developed in experimental economics in evaluation contexts.

4.4. The Value of Multiple Methods within a Given Evaluation

Frequently there is a political demand for each evaluation to produce "one number". A social experiment frequently does produce a single impact estimate. In contrast, in a non-experimental context it is wiser to pursue several alternative evaluation strategies. At the end of the day, if they all produce similar estimates, this is very comforting. If they do not, then the evaluator has to figure out why they are different and also weigh the evidence in order to determine which estimates should receive the greatest weight in making overall conclusions about the program. The evidence to be weighed here consists of institutional evidence and evidence from the data collected for the evaluation on the plausibility of the assumptions that justify each of the non-experimental estimation methods employed.

As an example, consider the evaluation of the new Ticket to Work (TTW) program in the U.S. This program provides vouchers for rehabilitation and employment-related services to persons receiving Supplemental Security Income (SSI) and disability insurance (DI) in the U.S. Random assignment could not be used because of provisions in the law requiring everyone in the eligible population to be served. Instead, the present evaluation design outlined in Lewin Group (2001) will employ three alternative non-experimental strategies.

The first consists of comparisons within states between the period before and the period after implementation of TTW. The second consists of difference-in-differences estimates that compare states that implement TTW early to states that implement it late. The third consists of comparisons using matching methods of persons who use and do not use their vouchers within particular states following TTW implementation. Note that the third strategy estimates the impact of using the voucher rather than not using it, while the first two provide estimates of the impact of the *offer* of a voucher. However, the offer estimates can be rescaled to provide estimates of the impact of treatment on the treated. As this is a new program, it is not clear in advance which strategy will yield the most plausible estimates. Each one depends on different assumptions about how the implementation of the program works out, what happens to macroeconomic conditions during the evaluation period and on the process by which individuals decide whether or not to use their vouchers. Pursuing all three increases the likelihood that at least one set of estimates will prove compelling.

4.5. Outcome Variables

Thus far the discussion has involved a high level of abstraction. We have not mentioned exactly what variables are of interest, nor how to measure them. Earnings, employment, unemployment, re-employment rates and the like are usually the key economic variables in evaluating employment and training programs. There are good reasons to look at these variables. Increasing employment and earnings constitute the primary goal of most active labor market policies. Impacts on earnings form a natural input into cost-benefit analyses.

At the same time, non-economic outcomes such as well being or satisfaction can be of interest as well. Evaluating impacts on these outcomes raises no new econometric

problems. The only question is whether the additional information they provide is worth the cost of collecting and analyzing the information. This is a particular concern given that many of these variables are highly correlated with employment and income. Of course, in some cases the economic and non-economic outcomes may differ substantively. A program may make people feel satisfied, but lead to poorer labour market outcomes. Policy or political decisions must then weigh these conflicting findings.

5. General Equilibrium Effects

Another issue that we have not addressed so far concerns what we call general equilibrium effects. These represent the effects of programs on persons who do not participate in them. For example, a job placement program that helps one group of people find jobs, may simultaneously make job finding more difficult for another group. Alternatively, programs may have spillover effects if improving the labor market outcomes of participants benefits others (perhaps family members). These “external” costs and benefits are extremely difficult to identify and estimate, but they should be considered. It is important not to expect too much in the way of community-level effects from relatively small programs. Money spent looking for them may not be money well spent. In most cases, these effects cannot be estimated using random assignment. Estimating general equilibrium effects may require a separate evaluation component using different methods.

We now briefly consider a few examples of quality studies that have looked for, or are looking for, general equilibrium effects. First, we have two Canadian studies. As part of the Self-Sufficiency Project, Card, Robbins and Lin (1997) in a separate analysis looked for what the literature calls “entry effects”. These effects result when changes to a program

make it more attractive and thus induce more individuals to participate in it, or induce individuals already in the program to stay in it longer than would have been the case without the change. In the case of the SSP, the addition of the wage subsidy, available after one year on income assistance, was estimated to increase the fraction of new income assistance recipients who remained on the program for at least a year by a few percent.

In a related study, Morris and Michalopoulos (2000) look at the impact of impacts of the SSP's incentives for parents to return to work on various outcomes of their children. This is a particularly important and neglected externality, or general equilibrium, effect. Many programs that are targeted at adult members of the labour force have impacts on other members of the household that are commonly ignored in program evaluations.

In the U.S., Davidson and Woodbury (1993) looked for displacement effects in the U.S. Unemployment Insurance (UI) bonus experiments. In these experiments, the treatment consisted of the offer of a cash bonus to claimants who found a job early in their UI spells. Davidson and Woodbury (1993) estimate that the displacement of workers not in the experiment cancelled out about twenty percent of the employment impact of the program estimated in the experiment. In a study of tuition subsidy programs for university students, Heckman, Lochner and Taber (1998) find much larger general equilibrium effects. In their study, the partial equilibrium estimate of the impact of treatment on the treated is ten times larger than a general equilibrium impact that accounts for the decline in the relative wage of persons with a university degree that results from their increased supply. Overall, these studies, as well as a number of studies by European scholars, suggest that general equilibrium effects should not be ignored in the evaluation of large-scale social programs. Even in contexts where conducting a general equilibrium analysis is infeasible, estimates from the

literature can be used to determine how sensitive the cost-benefit performance of a program is to potential general equilibrium effects.

6. Facilitating Evaluation: Program Design, Program Implementation and Data Quality

6.1. Program Design and Implementation

As we have suggested a number of times already, the ways in which programs are implemented or managed can hinder or facilitate understanding their effectiveness.

Sometimes taking evaluation requirements into account in the planning of a program can allow very inexpensive, and also very high quality, analyses to be conducted.

Consider the following, by no means exhaustive, list of ideas. First, in the case of program design, lay out and enforce specific eligibility rules defined in terms of observable characteristics. This allows the construction of comparison groups known to be eligible for the program. Second, to the extent possible, have explicit rules for the process by which eligible persons come to participate in the program or in specific services in a multi-treatment program. This helps in modeling the participation process, which forms an important component of all non-experimental evaluations. Third, introduce randomization into the program where possible, as in the example of the Kentucky Unemployment Insurance profiling program given above. Fourth, allow variation in program parameters, rules and funding levels across jurisdictions within the program. For evaluation, it is even nicer if this variation is exogenous (or even random) but it is always useful. All of these types of institutional variation facilitate evaluation by inducing variation in program participation. In other words, this institutional variation can be used as instruments, as discussed in Section 4.2. Fifth, design the information system so that it collects and keeps rich information on

program participation and participant characteristics. Avoid system designs that overwrite old information with new information; disk space is cheap and the old information may prove useful in evaluation.

In the case of program implementation, we offer the following suggestions. First, start the evaluation process early and integrate it with the program implementation process. This may mean breaking the evaluation contract into two parts: evaluation design and evaluation execution. Starting the evaluation process early should allow for the collection of useful baseline data prior to program rollout. Third, and even better, run the program as a demonstration prior to full rollout, and allow enough time for really bad programs to be cancelled prior to full rollout. Running a demonstration provides valuable information on participation rates, daily operation, information systems and other program features that aids in the design of a credible evaluation for the full program. Fourth, stagger the rollout if possible to allow use of the variation in rollout dates as an instrument. Another way to accomplish the same thing is to stagger the rollout across subgroups of the eligible population. The late introduction subgroups can then act as a comparison group for the early introduction subgroups. This was done with some of the New Deal programs in the U.K., which were introduced for youth prior to being introduced for adults.

An example helps to illustrate some of these points. This (mostly) good example has already been mentioned – it is the Ticket to Work evaluation in the U.S. Although random assignment was ruled out due to universal service requirements, the program was designed in part with evaluation in mind. Two features stand out in this regard. The first is the fact that the evaluation effort was begun well before the program rolled out anywhere. This allows for baseline data collection and thereby for the use of before and after comparisons within

states. The second is the staggered rollout of the program across states. Combined with baseline data collection, this allows for difference-in-differences estimates between early and late rollout states. It also allows for some adjustments to the evaluation and data collection strategies in later states based on the experiences in early states.

6.1. Improving the Quality of Administrative Data

Administrative data are increasingly used in the evaluation of social programs around the world. They have a number of advantages over survey data. Generally, they are available for long time periods, allowing longitudinal methods to be used if desired or at least allowing analysts to condition on long histories of labor market outcomes and/or of program participation. Program treatments are also often (but not always) measured better in administrative data than they would be in surveys (see, e.g., the discussion in Smith and Whalley, 2001). Administrative data usually allow access to a whole population, which means that samples are large and survey non-response issues do not arise. Administrative data can also be much cheaper than survey data, particularly when marginal costs are compared.

At the same time, administrative data have some weaknesses for evaluation purposes, weaknesses that can in part be ameliorated through administrative changes. The key weakness in most administrative data sets is a lack of covariates. This makes them unsuitable for evaluation methods that assume “selection on observables” as discussed in Section 4.1 (though such methods are sometimes applied to them anyway). The obvious fix for this problem involves collecting some additional variables. In addition, in some cases a useful field exists, but caseworkers have no incentive to fill it in. Education provides an

example of this. It is a simple matter to force completion of these fields by caseworkers before they can proceed on to the other things they want to do. Matching administrative data sets also causes problems. Data entry errors in Social Insurance Numbers and other identifying information can cause mismatches, which reduces data quality and increases measurement error. Changing the data entry software to double-check these values upon entry could cheaply avoid some of these errors (and would likely benefit the other tasks for which the data are used).

6.2. Improving the Quality of Survey Data

In their research applying non-experimental methods to the data from the experimental evaluation of the U.S. Job Training Partnership Act (JTPA) evaluation, Heckman, Ichimura, Smith and Todd (1998) emphasize the importance of data quality in evaluations. This focus on data quality represents an attempt to bring some balance to a literature that has spent almost all of its time worrying about which econometric estimator to use, in the (vain) belief that the correct estimator could overcome low-quality data.

Improving the quality of the survey data used to do evaluations has a number of aspects, of which we highlight four. First, learn what variables to condition on in non-experimental evaluations and be sure to collect them in the data. This knowledge can be obtained by reading the literature and by running experiments in parallel with non-experimental data collection and using them to “calibrate” the non-experimental methods. Heckman, Ichimura, Smith and Todd (1998) found for adult men that there were much lower biases when they could condition on rich data on labor force histories just prior to program participation.

Second, increase response rates to more respectable levels. The U.S. Office of Management and Budget requires response rates of eighty percent. This level of response is not unattainable, despite growing respondent survey fatigue and refusal rates, though it may raise the cost of data collection.

Third, be sure to put participants and non-participants in the same local labor markets. For programs that operate in only a few local labor markets, this means gathering data on non-participants only from those local labor markets. It rules out the common strategy of using off-the-shelf national survey data to construct comparison groups for programs that do not operate nationally. The evidence on this in Heckman, Ichimura, Smith and Todd (1998) is clear, though. Failure to draw both groups from the same local labor markets results in substantial bias.

Fourth, be sure to measure the dependent variable in the same way for both participants and non-participants. Evaluations sometimes use administrative data on outcomes for one group and survey data for the other. Or, when off-the-shelf survey data sets are used for the comparison group, two different survey measures of earnings or employment may be compared. The evidence in the literature (e.g., Smith, 2001, and Moore, Stinson and Welniak, 2000) suggests that doing so often leads to incorporation of systematic measurement error in the impact estimates, because different methods of measuring outcomes such as employment and earnings lead to different values even for the same groups.

7. Facilitating Evaluation: Monitoring the Evaluators

Canada is fortunate to have a high-quality domestic evaluation industry. It can also call on the even larger community of experts outside its borders when the need arises. Nonetheless, even the best evaluation community can benefit from an institutional structure that encourages replication and independent review of evaluation findings. Such an institutional structure exists only in part in Canada at present. In this section, we provide some suggestions to help bring it about.

7.1. Public Use Data Sets

Subject to privacy concerns, a well-documented public use data set should be one of the products (“deliverables”) associated with every major evaluation. Such data sets allow independent verification of the official evaluation findings. Further, these data allow additional sensitivity analyses and the application of additional econometric methods beyond those in the official evaluation. They also encourage the production of additional research of interest to the government at little or no cost to it. Academics from tenured professors to lowly graduate students will jump at the chance to work with good data when it becomes available. In addition, as any academic doing empirical work knows, the possibility of future replication, and therefore of future public embarrassment if a mistake is found, is a powerful motivator to thought and care.

7.2. Independent or Peer Review and Publication of Results

Although almost all HRDC reports are available on its web site, which is an important element of public accessibility and awareness, HRDC should also encourage the publication of crucial evaluation results both in peer-reviewed journals, and in more popular outlets. This can be done by allowing for the time required for the preparation of journal articles in contract and data availability periods, and by making part of the payment conditional on publication of the evaluation results in such a outlet. Independent review by academic journals subjects technical aspects of the methods and interpretations of the official evaluation outside scrutiny. Furthermore, publication in policy and academic journals combined with more popular outlets exposes the evaluation findings to broader audiences and thereby allows more informed public debate on the merits of programs. It also helps the knowledge gained in individual evaluations to cumulate over time, which will result in improvements in program design and evaluation in the future. A full range of outlets should be pursued: from the academic to the popular.

Another, more modest, proposal also merits consideration. This is the inclusion of formal discussant comments from independent scholars in official evaluation reports. This was done in Westat's recent evaluation of the U.S. Employment Service program. For obvious reasons, these scholars should frequently be drawn from outside Canada, since the evaluation community (and the labor economics community more generally) in Canada is quite small. Doing this would add slightly to the cost of an evaluation, but would provide benefits both in terms of motivating the evaluators and in terms of informing the readers of the final evaluation report.

8. Cost-Benefit Analysis

Cost-benefit analysis is necessary for evidence-based policy. It exposes the full range of costs and benefits associated with a program by requiring their itemization and valuation. Individual studies may only provide inputs into the cost-benefit analysis, but a study that brings all the pieces together is necessary. A program of evaluation without a cost-benefit analysis is like a Canadian street corner without a Tim Horton's. It just isn't right.

8.1. Recent Developments in Cost-Benefit Analysis for Employment and Training Programs

Impact evaluations are, of course, just one input into a comprehensive cost-benefit evaluation. We have four main recommendations here, based on the recent literature. First, do a full-blown cost-benefit analysis. Do not play pretend by doing a "cost effectiveness" analysis that compares only one program to another or one service strategy to another. The costs of the program should be subtracted from its benefits, with both properly estimated and discounted into present value terms. The resulting number is then compared to an absolute standard, namely zero. Taxpayers work hard for the money that the government takes from them to pay for these programs; they deserve a serious accounting of how their dollars were spent.

Second, it is important in doing a complete cost-benefit analysis to consider multiple possible outcomes. Employment and training programs may have impacts on outcomes other than earnings and employment, such as participation in transfer programs, health, marital and family behavior and crime. The original Mathematica evaluation of the U.S. Job Corps program summarized in Mallar, Kerachsky, Thornton and Long (1982) is exemplary on this dimension. The National JTPA Study cited above and the recent experimental Job Corps

evaluation summarized in Burghardt, et al. (2001) also go beyond the standard earnings and employment outcome measures.

Third, it is important in doing a cost-benefit analysis to take account of the deadweight costs of taxation. As there is some debate in the literature about how big that cost is, the results of the cost-benefit analysis can be presented conditional on multiple assumptions about how large these costs are.

Fourth, evaluations typically have available only a couple of years of follow-up data, which implies that there is a need to extend the impact estimates outside of the data. The cost-benefit performance of a program may depend on whether or not the impacts persist after the period covered by the data. The literature provides some guidance here, but it is not definitive. As such, the results of the cost-benefit analysis can be presented conditional on multiple assumptions about the persistence of any estimated program impacts. Heckman, Hohmann, Khoo and Smith (2000) provide an example of a cost-benefit analysis that does this, as does Section 10 of Heckman, LaLonde and Smith (1999).

Fifth, as discussed in Section 5, an analysis should also take account of general equilibrium effects when possible. This may require a separate evaluation component or it may rely on estimates from the literature for similar programs. Once again, a sensitivity analysis including alternative estimates of the general equilibrium effects drawn from the literature may be in order.

8.2. Some Programs Do Not Work

An important part of doing a cost-benefit analysis is keeping in mind that the existing evidence suggests that employment and training programs are sometimes bad investments,

particularly for youth. That means, if you are the analyst doing the evaluation, you are sometimes going to bring bad news to the client. No one likes bad news, and no one likes to be the bearer of bad news. Nevertheless, the bad news is important here because it makes clear that the program must be fixed or the resources directed to other programs or returned to the taxpayers – the ultimate client in any program evaluation.

Just to drive the point home, we note two important evaluations that have found zero or negative impacts from employment and training programs. All three examples are based on experimental data. In the U.S. National JTPA Study, estimated impacts presented in Bloom, et al. (1997) and Orr, et al. (1996) on earnings were zero for female youth and negative or zero for male youth. This program provided very similar services, in a similar institutional setting, to those provided to unemployed youth in Canada. Long-term follow-up estimates for the U.S. National Supported Work Demonstration program reported in Couch (1992) show that for male youth, this expensive program – over US\$13,000 in 1997 dollars – providing on-the-job training and support services to high school dropouts, ex-addicts and ex-offenders, had no impact on earnings, even after seven years. Spending lots of money is no guarantee of programmatic success.

9. Summary and Conclusions

Program evaluation is a cornerstone of evidence-based policy and, more generally, of good government. In this paper we outline a number of ways in which to improve the integrity of the evaluation of employment and training programs in Canada. In particular, we highlight the following areas and policy recommendations:

- 1) Be clear about the policy question of interest. Be sure that the econometric evaluation methods and data collection strategies adopted provide an answer to that question, even

in a world where the impacts of programs vary across persons and wherein persons and program staff know this and make choices based upon it.

- 2) Use random assignment when possible. Frequent use of random assignment signals that a government is serious about evaluation and serious about basing policy on evidence. Infrequent use of random assignment sends the opposite signal. Keep in mind that randomization can aid in evaluation even without undertaking a traditional social experiment.
- 3) When random assignment is not used, a number of alternative non-experimental methods are available. The success or failure of these methods depends critically on decisions about the design and implementation of the program, and on the quality of the administrative or survey data used in the evaluation. Thoughtful choices about program implementation and design can create useful variation in participation across time, space or persons that allows for credible evaluations. Slick econometric methods will not, other than by chance, overcome weak data or careless program design and implementation.
- 4) General equilibrium effects of programs are important. Analyses of these effects require different methods, in general, than analyses of the impacts of programs on their participants. Funding such analyses will be worthwhile for large-scale programs. When a new analysis is impossible, the literature should serve as a guide in undertaking a sensitivity analysis of the cost-benefit performance of the program to likely levels of general equilibrium effects.
- 5) Institutional changes in the way evaluations get conducted in Canada could improve the quality of evaluations at low cost to the government. First, public use data sets should be created from all major evaluations in order that independent researchers can replicate and

expand on the official analysis. This approach has already yielded large dividends in the case of the SSP. Second, discussant comments from independent researchers could be included in the final reports of evaluations. Third, in the case of medium and large evaluations, one deliverable should be an article summarizing the evaluation methods and results in a peer-reviewed academic journal.

- 6) Cost-benefit analysis (not cost-effectiveness analysis) is the final step in program evaluation. Programs cost real money that would otherwise be used by taxpayers for their own ends. The taxpayers deserve a full and complete accounting of the success or failure of the programs operated on their behalf, not just a comparison to other programs or among different services within a given program. This includes taking into account the deadweight costs of raising the tax money to pay for the program. It also means a thorough sensitivity analysis to the duration of program impacts outside the period covered by the available data and to any possible general equilibrium effects. The literature suggests that some employment and training programs are not a wise use of government funds. Thus, a negative finding in a cost-benefit analysis is likely not a mistake, but rather very important news.

References

- Angrist, Joshua D. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants" *Econometrica* 66(2), 1998, pp. 249-88
- 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, pp. 1277-1366.
- Black, Dan, Jeffrey Smith, Mark Berger and Brett Noel. 2001. "Is the Threat of Reemployment Services More Effective than the Reemployment Services Themselves? Experimental Evidence from UI Claimant Profiling." Unpublished manuscript, University of Maryland.
- 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.
- Bound, John, David Jaeger and Regina Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430): 443-450.
- Burghardt, John, Peter Schochet, Sheena McConnell, Terry Johnson, Mark Gritz, Steven Glazerman, John Homrighausen and Russell Jackson. 2001. *Does the Job Corps Work? Summary of the National Job Corps Study*. Princeton, NJ: Mathematica Policy Research.
- Card, David. 2001. "The Returns to Schooling: Advance on some Persistent Econometric Problems" *Econometrica* 68(4): 1127-1160.
- Card, David, Philip K. Robbins, and Winston Lin. 1997. "How important are 'Entry effects' in financial incentive programs for welfare recipients?" Social Research and Demonstration Corporation, Ottawa.
- Couch, Kenneth. 1992. "New Evidence on the Long-Term Effects of Employment and Training Programs." *Journal of Labor Economics* 10(4): 380-388.
- Cragg, John. 1994. "Making Good Inferences from Bad Data." *Canadian Journal of Economics* (4): 776-800.
- Davidson, Carl and Stephen Woodbury. 1993. "The Displacement Effects of Reemployment Bonus Programs." *Journal of Labor Economics* 11(4): 575-605.
- Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-161.

- Heckman, James, Carolyn Heinrich and Jeffrey Smith. 2002. "Understanding Incentives in Public Organizations." *Journal of Human Resources*, forthcoming.
- Heckman, James, Neil Hohmann, Jeffrey Smith and 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 Evaluating 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. 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 Program." *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, pp. 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. 1998. "Evaluating the Welfare State." In Steiner Strom, ed., *Econometrics and Economics in the 20th Century: The Ragnar Frisch Centenary*. New York: Cambridge University Press for Econometric Society Monograph Series, pp. 241-318.
- Heckman, James and Jeffrey Smith. 1999. "The Pre-Programme Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies." *Economic Journal* 109(457): 313-348.
- Heckman, James, Jeffrey Smith and Nancy Clements. 1997. "Making the Most of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487-535.
- Heckman, James, Jeffrey Smith and Christopher Taber. 1998. "Accounting for Dropouts in Social Experiments." *Review of Economics and Statistics* 80(1): 1-14.
- Hum, Derek, and Wayne Simpson. 1993. "Economic Response to a Guaranteed Annual Income: Experience for Canada and the United States." *Journal of Labor Economics* 11(1, Part 2): S263-S296.

- 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.
- Lewin Group. 2001. *Evaluation Design for the Ticket to Work Program*. Falls Church, VA: The Lewin Group.
- Morris, Pamela, and Charles Michalopoulos. 2000. "The self-sufficiency project at 26 months: Effects on children of a program that increased parental employment and income". Social Research and Demonstration Corporation, Ottawa.
- 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.
- Moffitt, Robert. 1991. "Program Evaluation with Nonexperimental Data." *Evaluation Review* 15(3): 291-314.
- Moore, Jeffrey, Linda Stinson and Edward Welniak. 2000. "Income Measurement Error in Surveys: A Review." *Journal of Official Statistics* 16(4): 331-361.
- Orr, Larry, Howard Bloom, Stephen Bell, Fred Doolittle, Winston Lin and George Cave. 1996. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington DC: Urban Institute Press.
- Social Research and Demonstration Corporation. 2001. "Will the Working Poor Save to Invest in Human Capital? Insights from a laboratory experiment." *Learning What Works*, 1(2): 1-4.
- Smith, Jeffrey. 2001. "Measuring Earnings Levels Among the Poor." Unpublished manuscript, University of Maryland.
- Smith, Jeffrey and Petra Todd. 2002. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, forthcoming.
- Smith, Jeffrey, and Alexander Whalley. 2001. "How Well Do We Measure Public Job Training?" Unpublished manuscript, University of Maryland.

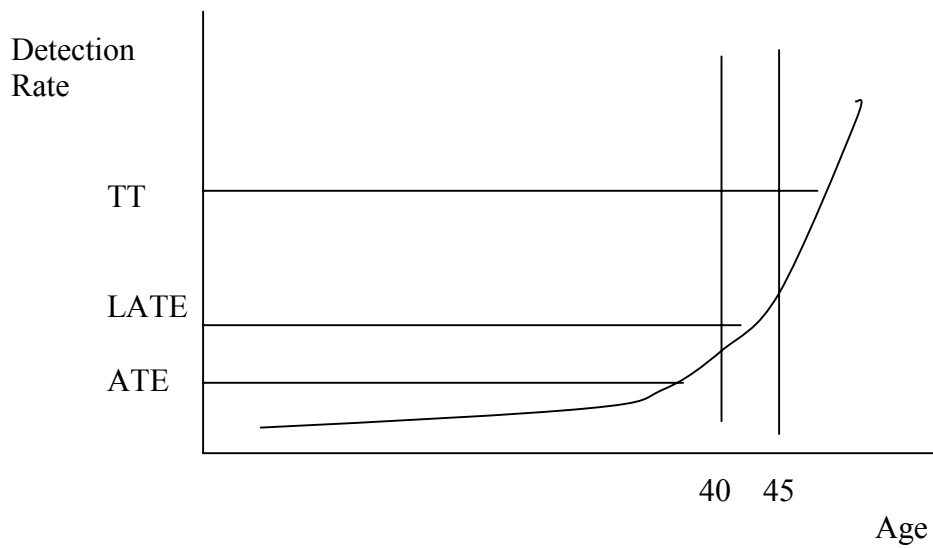


Figure 1 – Schematic of Mammography Detection Rates by Age