Quasi-Experimental Evaluation

Evaluation and Data Development
Strategic Policy
Human Resources Development Canada

January 1998

SP-AH053E-01-98
également disponible en français

Acknowledgement

Evaluation Services of HRDC has traditionally used a quasi-experimental research design (comparison group methodology) in estimating the impacts of an intervention. In order to ensure the suitability of the comparison group approach in subsequent evaluations of programs and services delivered under EI Part II, the attached study "Quasi-Experimental Evaluation" was carried out.

We would like to thank Dr. Robert Power of Power Analysis Inc. of London, Ontario and Dr. W. Craig Riddell, Department of Economics, University of British Columbia for researching and analyzing the evidence for and against comparison group design and for the preparation of the report on findings from a through review of the literature.

Series

Human Resources Development Canada (HRDC) has a long tradition in program evaluation. Over the years, the Evaluation and Data Development Branch (EDD) of HRDC has conducted numerous evaluations on a wide range of departmental programs. The expertise and experience acquired have permitted EDD to hone its evaluation techniques and approaches.

More recently, however, changes in the delivery nature of some HRDC programs have affected the way evaluation is conducted. Now, HRDC regional offices are more directly and actively involved in program evaluation.

In response to these changes EDD is proud to initiate a new series of publications entitled "Evaluation Tool Kit". The objective of the series is to build evaluation capacity and knowledge throughout HRDC and its partner organisations by providing pertinent information about designing, planning and conducting an evaluation. Existing information about particular topics will be synthesised and made readily accessible in short, precise reports which are adapted to the evaluation of human resources development programs.

The evaluation field is one of growing interest and progress. We invite you to provide us with comments you may have on this series or on a specific study.

The first study of the new series focuses on the quasi-experimental approach in evaluation.

Table of Contents

Quasi-Experimental Evaluation

Program manager: Why do we need to evaluate our program? We have a good handle on what's going on with our program and our clients, and we know we are very successful.

Evaluator: Because you never know if it was your program or something else that produced the success you are claiming.

Program manager: Of course we know it's our program. What else would cause all these people to find jobs so quickly?

Evaluator: Maybe the rapid economic expansion that we are now enjoying.

***Program manager*:** That's nonsense. Anyway, we know we have to have our program evaluated. But why do we need to go through the trouble of finding a comparison group to do an evaluation?

Evaluator: Let's say that six months after finishing your training program, 70% of the trainees are working. Would you consider that proof your program is a success?

Program manager: I'd say so, yes. We'd like to do better than 70% —in fact, we believe we are doing better than that — but I'd say that having 70% of our trainees working would make our program look very good, especially given the barriers many of our clients face when they come to us.

Evaluator: What proportion of these individuals might be working now if they hadn't gone through your training program?

Program manager: I'm not sure, but it wouldn't be as high as 70%, I can tell you that.

Evaluator: Well you don't really know that though. For all you know, 80% might be working now if they hadn't taken training.

Program manager: No way. You don't know how many obstacles our clients face when they come to us. We are providing a valuable service and are really helping our clients.

Evaluator: That may be so, but you haven't proven it. And the sponsors of the program need to know with certainty how successful you have been. They need to know they are getting bang for their buck.

Program manager: They are, I assure you.

Evaluator: Okay, let's say you are doing some good, that individuals who go through your program are indeed more likely to find a job than if they hadn't been trained. How much of an effect are you having? Would half of them have found jobs anyway? One-third? Two-thirds? You can't know that unless you do an evaluation that includes a comparable group of people who haven't taken your training?

Program manager: Even if half of them would have found a job without the training, isn't raising that proportion to 70% worth it?

Evaluator: I don't know. What did it cost for that incremental 20%? And how long will the effects of training last?

Program manager: Our program is well worth the money . . .

This contrived conversation illustrates well the different perceptions of program managers and evaluators when it comes to assessing the merits of a program. Managers live and work with the program every day: they care about their program and work hard to make it a success; they strongly believe they are doing a good job; and they understandably resent any implication that they are not.

Evaluators usually have no connection with the program, and, more importantly, no stake in its survival (which sometimes leads to an underestimation of the threat that evaluations can impose on program management and staff). They know that managers are heavily invested in their program and that a manager's assessment of the program — even one aided by reliable monitoring data — will not be accepted by program sponsors as a valid test of whether the program is meeting its objectives and is worth what it costs. And they know that many different factors that are unrelated to the design of the program can affect the outcomes of any social program, and can easily lead to unwarranted conclusions about the program.

This paper is based on the premise that only a good program evaluation can produce convincing evidence of a program's effectiveness in reaching its objectives, and that only an evaluation that includes a control or comparison group can provide estimates of program impact uncorrupted by the influence of other factors that also may affect outcomes. It summarizes the basics of evaluation research, focusing on what is usually the best practical approach — the "quasi-experimental design." It is written in non-technical language for managers who have little or no exposure to the field of evaluation, but includes more advanced treatments in appendices for those interested in the more technical aspects.

The paper begins with a brief introduction to evaluation, including a broad definition, and a summary of the two basic types of evaluation. It then moves to a discussion of evaluation design, starting with the reasons evaluators need to worry about appropriate designs. Common evaluation designs are introduced, with the pros and cons of each discussed. Designs without comparison or control groups are shown to fall well short of ideal in terms of ruling out extraneous

causes. Since any good evaluation is fundamentally a comparison between what happened to program clients and what would have happened had they not been in the program, one-group designs virtually preclude any serious summative evaluation.

Next, the report explains details of quasi-experimental evaluation design — its theoretical underpinnings and practical considerations. The concept of selection bias is explained and econometric procedures to control for it are introduced. Confidence in econometric estimates will be explored.

How to select comparison groups is the next consideration. The different techniques for matching are summarized, with an assessment of their relative strengths and weaknesses. Potential variables to use in conducting the matching are reviewed at this stage. Finally, with the general topic of quasi-experimental evaluation well explicated, we then proceed to apply the lessons to the more specific tasks of determining the best sampling techniques for drawing regional Employment Benefits and Support Measures (EBSM) comparison groups, and the best variables to use in the sampling.

1.0 Definition of Evaluation

There are a lot of different definitions of evaluation. Here is one of the best, because it touches on the most important aspects of evaluation:

> "Evaluation is a collection of methods, skills and sensitivities necessary to determine whether a human service is needed and likely to be used, whether it is conducted as planned, and whether the human service actually does help people" (Posavac and Carey, 1980, p.6).

This definition encompasses the two main types of evaluation: process and summative.

2.0 Two Main Types of Evaluation

Although the literature includes over a hundred different kinds of evaluation (see Patton, 1982), the vast majority boil down to two types: those that aim to determine if the program has been implemented as planned, and those that measure its success in achieving its objectives (i.e., its impact). The label most often associated with the first type is "process evaluation," although it is sometimes called formative evaluation. The latter type is known as "summative evaluation," also known as impact, outcome, or effectiveness evaluation.

Process Evaluation — *How is the program operating and how can it be made better?*
Process evaluations are directed at three key questions: (1) the extent to which a program is reaching the appropriate target population; (2) whether or not its service delivery is consistent with program design; and (3) what resources are being expended (Rossi and Freeman, 1993).

The main objective is to provide feedback to managers on whether the program is being carried out as planned and in an efficient manner. Guidance should be provided for modifying

the program to help ensure it meets its objectives. With this information, the program can be modified so it is carried out as planned, or the plan itself can be modified if it is found lacking.

Summative Evaluation — *Does the program achieve its objectives?* The purpose of summative evaluations is to assess the impact of the program; that is, ascertain the extent to which the program meets its objectives, and the needs of its target group. As well, it should provide advice for modifying the program so that it will better serve the needs of its clients and become more cost-effective (Stufflebeam and Shinkfield, 1985).

The remainder of this report will focus on the summative side, in particular summative evaluations with quasi-experimental designs. Before plunging into the arcane field of quasi-experimental design however, the paper sets a proper context through a discussion of the main types of summative evaluation designs.

3.0 Summative Evaluation Designs

There is no single correct evaluation design for impact evaluations. The goal is to come up with the best design possible under the circumstances. Almost all designs represent a compromise dictated by many practical considerations such as how much money and time are available, what the client considers compelling, how much a design might interfere with the normal operation of the program, and so on.

In deciding on the design, the credo is to maximize the credibility and usefulness of the findings. The evaluator must anticipate the kind of arguments that will be used to dismiss the findings.

In summative evaluations the key concern is to be able to ascribe the outcome to the program as opposed to innumerable other possible causes. In the vernacular of evaluators, this is known as "***internal validity.***" Every good summative evaluation is designed to minimize "threats to internal validity"; that is, they are designed to isolate the impact of the program from the impact of other potential causes. Campbell and Stanley (1971) identified seven threats to internal validity (Exhibit 1).

4.0 Types of Designs

There are dozens of possible designs to determine program impact. This section will summarize the most popular ones.

Because designs without comparison groups are very deficient in terms of internal validity — that is, they cannot normally rule out alternative explanations for program outcomes observed — ***all good summative evaluations include a comparison group****.* Nevertheless, single group designs are very common.

Exhibit 1 - Threats to Internal Validity

Threats due to real changes in the environment or in participants:

History: Changes in the environment that occur at the same time as the program and will change the behaviour of participants (e.g., a recession might make a good program look bad).

Maturation: Changes within individuals participating in the program resulting from natural biological or psychological development.

Threats due to participants not being representative of the population:

Selection: Occurs when assignment to participant or non-participant groups yield groups with different characteristics. Pre-program differences may be confused with program effect.

Mortality: Participants dropping out of the program. Drop-outs may be different from those who stay.

Statistical Regression: The tendency for those scoring extremely high or low on a selection measure to be less extreme during the next test. For example, if only those who scored worst on a reading test are included in the literacy program, they might be bound to do better on the next test regardless of the program just because the odds of doing as poorly next time are low.

Threats generated by evaluators:

Testing: Effects of taking a pretest on subsequent post-tests. People might do better on the second test simply because they have already taken it [1].

Instrumentation: Changes in the observers, scores, or the measuring instrument used from one time to the next.

Single Group Designs

The simplest but least satisfactory evaluation design is the posttest only design, symbolized, X O (where X is the program intervention — e.g., the training course — and O is a post-program observation such as annual earnings). Here, participants, having completed the program of interest, are surveyed to find out how well they are doing with respect to the behaviours or attitudes at issue. *This design cannot be used to credibly attribute any effects to the program,* for there is no objective basis to suppose that the program caused any changes. Indeed, because there is no information on the pre-program level of the variable(s) of interest, this design yields no information on change.

---

[1] Also, taking a pretest may sensitize participants to a program. Participants may perform better simply because they know they are being tested — the “Hawthorne effect.”

Whenever a program is supposed to bring about a change, before-and-after measures are a necessity with one-group designs. The simplest, the *pretest-posttes*t design symbolized as O_1 X O_2 (where X again denotes the intervention and O1 and O2 denote pre- and post-program outcome measures) requires a pretest of some sort before the program takes place (a reading test, for example), and a posttest after the program. This design is subject to most threats to internal validity. Most seriously, participants might have changed or some extraneous event may have brought about any observed difference between O_1 and O_2, so *no change can credibly be ascribed to the program.* For example, if an evaluation of this design showed that mean post-program earnings were lower than mean pre-program earnings, this should not be construed as proof that the program was defective: a recession may have caused the earnings drop.

Time series designs involve collecting data repeatedly about participants' situation at several times. Symbolically:

$$O_1 O_2 O_3 X O_4 . . . O_n$$

This design can be used to rule out (or at least quantify) regression and maturity as threats to internal validity. That is, any personal trends in the absence of the program can be accounted for and controlled. Advanced statistical procedures are required to isolate the effect of the program. History remains the key threat. Although supplementary data on the environment can help rule out events that can be identified, it is extremely difficult to identify — let alone quantify — all possible events that could have brought about the outcome observed. Consider the following example. Say Province A implemented a large-scale training program for its social assistance clients and observed the social assistance caseload statistics for several months before and after the intervention to see if the training program was lowering dependence on social assistance. But, at around the same time, Province B instituted its own policy change: a cutback in social assistance benefits for employable clients. That could precipitate an inflow of social assistance clients from Province B to Province A. Unless Province A knew about the policy change in Province B and took steps to measure its impact, the time-series evaluation could underestimate any positive impact of the training program.

In sum, one-group designs virtually preclude any serious summative evaluation. They are notoriously weak and easily dismissed, because it is generally impossible to rule out potential alternative explanations, especially key events that occurred while the treatment group was participating in the program (e.g., a recession). Two-group designs are required for sound evaluation.

Two Group Designs

A valid determination of impact requires comparing outcomes of a group of individuals who have participated in the program (treatment group) with an equivalent group of people who have not participated (control or comparison group). In theory, the best way to do this is by means of a *randomized experiment*, where individuals are assigned at random to the treatment or control group (Rossi and Freeman, 1993). Outcomes measures, chosen on the basis of

program objectives, are observed at some interval after the intervention ends, with any differences between groups attributable to the program: that is, the program can be said to have caused the observed differences. The design is symbolized as follows:

X O [participants]
O [non-participants]

Since randomization should remove — at least on average — any systematic differences between the groups, no pretest is needed. The impacts of the treatment can be measured simply by comparing the mean outcomes for treatment and control groups, with chance differences largely accounted for through standard statistical techniques (Greenberg and Wiseman, 1992).

The primary threat to this design — assuming the randomization process was carried out correctly — is non-random mortality or attrition (i.e., members of the participant group drop-out before completing the program, or members of participant and control groups cannot be located for follow-up purposes for reasons that are not random and therefore potentially systematically related to the impacts of the program). For this reason, a pretest is often given to both treatment and control groups (Mark and Cook, 1984, hold this is "essential"), so the effects of discontinuation from the program can be quantified and accounted for in the analysis.

Although experimental designs are as close to ideal as possible in theory, they are seldom practical. By far the most common constraints are program staff who refuse to comply because they consider randomized selection unethical or unacceptable, and evaluation timing: most often the evaluator enters the scene long after random assignment should have taken place. There are other problems as well. Most seriously, experimental methods are normally confined to a determination of the mean impact of the program; they cannot answer many other key policy questions, including the median impact of the program and the proportion of participants with a positive (or negative) impact from the program (Heckman and Smith, 1995)[2].

Given these constraints, quasi-experimental (non-experimental) models are frequently the only satisfactory way to proceed. There are different *quasi-experimental models*, but the most common and robust method involves constructing a comparison group of individuals who are comparable to participants. This can be done: by statistically controlling for differences between groups during data analysis; by matching participants and non-participants according to key traits (such as age, sex and education) believed to influence the outcomes of interest; or both[3].

---

[2] Appendix A briefly explores the on-going debate on the choice between experimental and quasi-experimentalmethods.

[3] Rubin (1979) showed that the use of both techniques — matching and statistical adjustment — was better than use of either technique alone (as reported in Dickinson et al, 1987).

The idea is to approximate random assignment as closely as possible by attempting to minimize or control for differences between the groups. Symbolically[4]:

$O_1 X O_2$ [participants]

$O_1 O_2$ [non-participants]

Here X is the program intervention, O_1 is a pre-program observation, and O_2 is a post-program observation. For instance, O_1 could be annual earnings in 1995, O_2 could be annual earnings in 1997, and X could be a 1996 training program.

Under a longitudinal quasi-experimental approach, for example, the evaluator compares the outcomes of two groups: program participants (the "treatment group") and non-participants (the "comparison group")[5]. "Outcomes," which relate to the objectives of the training program — finding a job, for example — are usually determined via a follow-up survey, conducted months or even years after program exit. The post-program outcome for each group is compared to pre-program status to determine if there has been any change, on average, within each group. Then, standard statistical procedures determine whether the change differs significantly between groups.

For example, say a baseline survey (see below) showed that half the participants and half the non-participants were working one year prior to applying for EI; and a follow-up survey found that 70% of trainees were working one year after the training, versus 60% of the comparison group one year after leaving EI. The increment for trainees is thus 20 percentage points, versus 10 for non-trainees. Simple statistical tests would determine if this difference is significant.

But, a finding of a statistically significant difference between groups does not necessarily imply that the difference was due to the program. *The analyst must demonstrate that the difference is attributable to the program.* That is, threats to internal validity must be ruled out. Unfortunately, the empirical evidence shows that participants are likely to be different from non-participants in ways that affect the outcome variables. Selection into most programs is non-random: those who volunteer to participate may be more motivated than those who do not, for example; and program administrators more often than not select those they feel will have the best chance of succeeding (i.e., the most talented), or conversely, select those most in need of the treatment.

Regardless of its source, *selection bias* affects the comparability of treatment and comparison groups. As long as all differences between the groups being compared are observable (e.g., personal traits), selection bias will not be a problem because statistical methods such as multiple

---

[4] Note that pre-program information is not strictly necessary for some comparison group designs — see section below on addressing selection bias — but it is always desirable.

[5] Appendix B presents an elementary mathematical representation of the quasi-experimental method.

regression analysis can control for the differences. Researchers do their utmost to match individuals in treatment and control samples to ensure observed characteristics are very similar, but they seldom know why a person is participating in a program. If any unknown (hence uncontrolled) feature of the person or program influenced the decision to participate, then the selection is non-random and differences between participants and non-participants may be incorrectly ascribed to the program.

No statistical method is likely to completely resolve the selection bias problem. Since it is impossible to anticipate all the factors that went into the decision to participate, the surveys and protocols cannot be designed to gather all relevant information. Quasi-experiments require analysis techniques that are much more complicated than those for true experiments. High-level statistics — "econometric models" — are required to deal with the differences between groups and isolate the effect of the program.

5.0 Examples of Selection Bias

A number of procedures are available for dealing with the potential problem of selection bias. In order to describe these methods it is helpful to use some concrete examples of the forms that selection bias frequently takes in evaluation research on education/training programs.

Example 1: Coop versus non-coop education

Cooperative education programs combine periods of formal education with periods of work experience in a systematic manner, while the more common non-coop educational programs provide formal education without organized periods of work experience. Many observers feel that the combination of education and work experience is likely to enhance the employability and earnings of coop graduates. Does empirical evidence support this belief? A simple way of answering this question would be to compare the post-graduation employment and earnings experiences of a sample of coop and non-coop graduates. Would this be convincing evidence of the impact of coop programs on employability and earnings? The answer in general is no.

The reason is that there may be reasons why those completing coop programs may have different levels of employability and earnings than those completing non-coop programs that are independent of the educational programs themselves. For example, if there are a limited number of coop programs (as is generally the case), these programs may be able to admit more qualified students, on average, than comparable non-coop programs. Similarly, the students who apply for coop programs may be more career-oriented, on average, than those applying to non-coop programs. Both these reasons would lead to the coop graduates having higher levels of employability and earnings even in the absence of completing a coop program. That is, if the coop graduates had instead completed non-coop programs, their employability and earnings would have been higher than the non-coop graduates. To some extent, higher levels of employability and earnings of the coop graduates are due to their being better, more qualified students and to their being more career-oriented than their non-coop counterparts.

Of course, it is also possible that the coop programs have a positive impact on the employability and earnings of graduates. If so, the total observed difference between coop and non-coop graduates consists of two components: one because coop programs attract more qualified and more career-oriented students (the selection effect), and one due to the impact of the program (the program impact effect).

This is the problem of potential selection bias. Coop programs are selecting the more qualified and career-oriented students who would have had higher employability even in the absence of the program. In these circumstances, a simple comparison of coop and non-coop graduates would over-estimate the true impact of the program. However, as discussed below, selection bias may be either positive or negative; that is, without accounting for selection bias, the estimated impact may be either above or below the true impact.

Example 2: Private schools versus public schools

Many of the same issues arise if one wishes to compare the outcomes (such as average grades in standardized examinations, high school completion rates, or post-secondary success) of private and public schools. Private schools may be able to select more qualified students, on average, than their public school counterparts. Similarly, the average student attending a private school may be more likely to have other attributes (such as having parents who place a higher value on education) than the average student attending public school.

For these reasons, a simple comparison of student outcomes in private and public schools is unlikely to give unbiased estimates of the impact of private schools on these outcomes. Some of the observed differences arise because private and public schools select students who differ systematically on such attributes as qualifications, parental wealth and family attitudes towards education.

Example 3: Impact of government-sponsored training programs

As will now be clear, similar issues of selection arise in the context of assessing the impacts of training programs. Those who undertake training are likely to differ systematically from those who do not take training. This could be due to differences between the trainees and non-trainees themselves; for example, those who are more educated or more labour market oriented may be more likely to apply for training. Alternatively, these differences could arise because of selection by program administrators, who may be more likely to enroll those who are most likely to benefit from training.

This example also helps make clear why selection bias may be either positive or negative. Suppose a program is designed to focus on the most disadvantaged among a particular population. In these circumstances, a simple comparison of trainees and non-trainees is likely to under-estimate the true impact of the program.

These three examples illustrate the point that virtually any program evaluation is likely to face

some potential selection bias. The reason is that participation in virtually all programs or interventions involve choices on the part of participants and non-participants and on the part of those administering the program. Because of such choices, the pools of participants and non-participants are likely to differ in systematic (or non-random) ways. If such differences between participants and non-participants are also related to the program outcomes, simple comparisons of participants and non-participants will give biased estimates of program impact. To deal with the very pervasive problem, a number of statistical methods have been developed.

6.0 Procedures for Addressing Selection Bias

Before describing the various procedures that have been developed and extensively used to deal with potential selection bias, two preliminary observations are noted. First, as the above examples make clear, selection into the program may be based on observable or unobservable factors. For example, in the case of comparing coop and non-coop programs, the qualifications of the coop and non-coop students may be observed (e.g. their high school grades) but the degree of career-orientation of the students may not. What factors are observed and what factors are unobserved will depend on the richness of the available data.

Controlling for selection into the program which took place on the basis of observable factors is straightforward. Thus the richer the available data (and thus the fewer the number of unobservable factors), the smaller will be the magnitude of selection bias due to unobserved factors.

The second observation is that although there are some unobserved factors which influence selection into the program (which will almost always be the case), this does not necessarily imply that simple comparisons of participants and non-participants will be subject to selection bias. Selection bias arises when the unobserved factors which influence participation/non-participation in the program also influence the impacts of the program.

To make this point clear, consider an extreme example. Suppose we wish to compare public and private schools in terms of their educational outcomes such as student performance on standardized tests. Suppose that, on average, private schools enroll more students with odd numbered birthdays than public schools. To the researcher, the attribute “having an odd/ even birthdate” is unobserved. This is a case of non-random selection; if students were randomly selected into public and private schools then the proportion of students with odd numbered birthdays would be approximately equal in the two school types. However, as long as having an odd numbered birthday does not influence the outcomes of interest (student performance on standardized tests), this non-random selection will not bias a simple comparison of student achievement in public and private schools.

For these reasons, methods for dealing with selection bias focus on the potential problems associated with factors which influence selection into the program which are (i) unobserved by the researcher/evaluator, and (ii) correlated with the program outcomes of interest. Unfortunately, in many evaluation studies there are a number of such factors meeting both

these conditions. For this reason, it is strongly advisable in virtually any evaluation study to address potential selection bias.

By their very nature, the factors that might give rise to selection bias are unobserved. In some cases, the methods described below will lead the evaluator to conclude that such potential sources of selection bias are not quantitatively important, and therefore do not lead to bias. This could be because unobserved factors leading to non-random selection into the program are not quantitatively significant in this particular program, or it could be because such factors are quantitatively significant but are not correlated with the program outcomes of interest. In other cases, the methods described below will lead the evaluator to conclude that selection bias is quantitatively important. In these cases, the methods also provide an estimate of the magnitude of the bias so that an estimate of the true impact(s) of the program on outcomes can be derived.

6.1 Two Step Adjustment Procedures

Two step (or two stage) procedures for addressing selection bias were developed by James Heckman and others in the late 1970s, and have become the most commonly used methods. In the first stage, the probability of participation in the program is analyzed. This analysis usually consists of a single equation model in which the dependent variable is the probability of participating in the program (an indicator variable which equals unity for program participants and zero for non-participants) and the independent variables are various factors that are believed to influence program participation/non-participation. The main purpose of the first stage is to obtain a correction factor (called the “inverse Mills ratio”) which is used in the second stage to take account of possible selection bias. As well, the estimates obtained in this first stage may be of interest in themselves in that they provide insight into the importance of the various factors that influence participation/non-participation in the program.

The second stage involves estimating program impact using a specified model (equation). The model includes:

- a “dependent variable,” which is the outcome the training program is supposed to affect, say earnings;
- several “independent” or explanatory variables, which are observed factors presumed to influence the outcome (e.g., age, sex, education);
- the “selection bias correction” variable (or inverse Mills ratio) obtained in the first stage; an indicator variable for participation/non-participation in the program; and
- a random error term to account for unobserved forces that may affect the outcome measure.

The model in words:

> Earnings = the effect of various observed factors + the effect of selection bias + the effect of the program + random error

(see Appendix C for the mathematical equation and further explanation)

Thus, the model isolates the impact of the program from other potential influences. *If the model is properly specified*, the addition of the "selection bias correction" variable removes this potential bias, thus giving unbiased estimates of program impact. We return below to the important issue of how to determine whether the model is properly specified.

A useful way to interpret this two step procedure is as follows. It is well known that omitting an important variable from a model will result in biased estimates of the coefficients on the variables included in the model. In the absence of a method for accounting for selection into the program, the estimation of the outcome equation omits an important factor — the determinants of program participation. The "selection bias correction" term obtained in the first stage provides an estimate of this factor, which is why including this term results in unbiased estimates (providing the model is properly specified).

One final observation relating to these two step procedures is in order. It is important to have one or more variables that influence selection into the program (i.e., that enter the first stage equation) but which do not influence the outcome(s) of the program (i.e., do not enter the second stage equation). Such variable(s) allow the separate identification of participation in the program and the outcomes or impacts of the program. Apart from the importance of having one or more such "identifying variables," the first stage participation equation and the second stage outcome equation may have many variables in common. The importance of these "identifying variables" also arises in the context of the method discussed next.

6.2 Instrumental Variable Methods

Selection bias arises because of the correlation between the indicator variable for participation/ non-participation in the program and the random error term in the outcome equation. The "instrumental variable" (IV) method to solving the selection bias problem, discussed by Heckman and Robb (1985) and Moffitt (1991) among others, centres on finding a variable (or variables) that influences selection into the program but does not influence the outcome of the program (and is thus not correlated with the random error term in the outcome equation). Because the instrumental variable is not correlated with the random error term, it can be used in the estimation without introducing bias. The formula for the IV estimator is given in Appendix C.

The search for IVs entails an in-depth investigation of the selection process. Personal characteristics of individuals would seldom suffice as instrumental variables because they are usually related to the outcome. For instance, level of education likely affects one's employability. Moffitt suggests that variation in the availability of treatment may yield a suitable variable. If a training program is available in one region but not another for reasons unrelated to the program's

intended outcome, region is a legitimate instrumental variable. This may be the case if the program is not available for political, bureaucratic or economic reasons.

In order to be a legitimate "instrument," the variable must be related to program participation/ non-participation but unrelated to the outcome(s) of the program. In some situations there may be numerous potential instrumental variables. In these circumstances, how should the analyst choose among these various potential IVs? The answer to this question is as follows. Each IV that is indeed unrelated to the outcome of the program (i.e. is uncorrelated with the random error term in the outcome equation) will yield unbiased estimates of the impact of the program. However, some IVs will yield more precise estimates of the impact of the program. Specifically, the more highly correlated is the IV with program participation/non-participation, the more precise will be the estimates of program impact. Thus the challenge in IV estimation is to find an instrumental variable that is highly correlated with program participation but uncorrelated with the outcome of the program. Unfortunately, it is often difficult to find variables that meet both these requirements, and therefore difficult to find good IVs among the many potential IVs.

As a way of further understanding the principles of IV estimation, consider the case of a program in which participation/non-participation is determined by random assignment. Suppose that each applicant is assigned as a participant P (non-participant NP) if the toss of a fair coin yields Heads (Tails). Then the indicator variable H (which equals unity if Heads and zero if Tails) is an ideal instrumental variable because H is perfectly correlated with the indicator variable for participation and because H is uncorrelated with the outcome of the program. Of course, in practice it is rare to have available such an ideal IV, but this example illustrates the characteristics one searches for when using this method.

This method thus has some features in common with the "identifying variable(s)" used in the two step method discussed above. In effect, both procedures require similar information. The main differences are: (i) the instrumental variable method is carried out in a single stage and does not therefore involve explicitly modelling the process of participation into the program; and (ii) the IV method produces estimates free of selection bias (if the model is properly specified) but does not provide an estimate of the magnitude of the selection bias, as is provided in the two step method[6].

Many examples of "instrumental variables" could be given. Moffitt (1991) gives the example of a government-funded health counseling program which, for reasons that are unrelated to the health needs of the populations in the two areas, funds the program in one area of a city but not in the other area. As a consequence, an indicator variable for the two areas of the city is unrelated to the health needs of the population in the two areas, but will influence the participation in the program. Another example, related to the assessment of public versus private schools discussed above, would be a measure of proximity to a private school for each of the students

[6] Also, the IV estimation procedure does not require the assumption of normally distributed random error terms in the first stage probit equation used in the two step procedure.

in the sample. Such proximity would be expected to influence the likelihood of attending a private school, but not the outcome of private schooling on student achievement. Appendix C has more on the IV method including equations.

6.3 Longitudinal Methods

The two step and instrumental variables methods discussed above can be implemented with post-program data alone (that is, cross-sectional data on participants and non-participants). However, if pre-program data on participants and non-participants are available, this should also be used and its incorporation in these methods will generally lead to more precise and more credible estimates of program impact. Longitudinal data follow the same individuals over two or more periods of time, and the longitudinal methods discussed here require at least one pre-program observation and at least one post-program observation on both program participants and non-participants.

The most common longitudinal estimator of program impact is the “fixed effects” or “difference-in-differences” estimator (sometimes also simply called the “differences” estimator). In the simplest case, in which there is one pre-program observation and one post-program observation, this approach proceeds as follows. First, take the difference between the post-program value of the outcome measure and the pre-program of the outcome measure for each participant and non-participant. This difference is thus a measure of how much change was observed in the outcome of interest between the period prior to the program and the period following the program. If the outcome of interest is earnings, as would be the case in assessing training programs, this would be the earnings gain or loss for each participant and non-participant. Next, take the difference between the average pre- versus post-program change for participants and the average pre- versus post-program change for non-participants (see Appendix C for equation). In the case of a training program, this is simply the difference between the average earnings gain or loss of participants and the average gain or loss of non-participants.

This simple estimator of program impact will be free of selection bias if selection into the program depends on unobserved person-specific “fixed effects,” that is, factors that are specific (or unique) to each individual in the sample, but are constant (“fixed”) over the time period of the analysis. For example, in the case of comparing coop and non-coop educational programs, such an unobserved factor could be how career-oriented the individual student is. If, as discussed previously, selection into coop and non-coop programs depends on how career-oriented the student is, and if this factor is not observed by the researcher (as will often be the case), the bias that would arise because of this unobserved factor will be removed by the use of the “difference-in-differences” estimator providing the extent of career-orientation of the student is constant over the sample period. Similarly, in the case of training programs, the “fixed effects” assumption is appropriate when unobserved person-specific factors such as ambition, labour force attachment and suitability for training are constant over time (but may vary across individuals). In many cases, such a “fixed effect” assumption will seem reasonable for such unobserved person-specific factors, although the validity of the assumption should be tested as discussed further below.

In summary, the difference-in-differences estimator will yield unbiased estimates of program impact — even in the presence of potential selection bias — when the source of potential bias is a correlation between participation/non-participation in the program and an unobserved factor which may differ across individuals but which is constant over time for each individual. If selection into the program takes this form, the simple difference-in-differences estimator is a straightforward way of dealing with selection bias.

More complicated longitudinal estimators are available for situations in which the assumption of constant or "fixed" person-specific effects is not appropriate. These generally require more than one pre-program and one post-program observation on each individual. For example, Moffitt (1991) discusses a "difference-in-differences in growth rates" estimator which is appropriate when the period-to-period change in the person-specific effect is constant over time. This estimator requires at least two pre-program and two post-program observations. Other types of longitudinal estimators are discussed in the context of training programs by Ashenfelter and Card (1985).

6.4 Specification Tests of Alternative Models

Three general classes of methods of dealing with selection bias have been outlined. Within each of these general classes, there are a number of variants on the basic procedure. The question which naturally arises is: Which of these methods should be employed, and under what circumstances?

In part, the answer to this important question is that the appropriate method should be determined by the researcher/evaluator, according to the circumstances of the program being assessed. A key aspect of being a good evaluator is being able to specify the model, including the methods for dealing with potential selection bias, appropriately. This aspect requires judgment, experience, and the ability to obtain information about the nature of the process by which participants are selected into the program.

Although factors such as judgment and experience are important, in most circumstances they are unlikely to be sufficient to enable the evaluator to know with reasonable certainty which method is most appropriate for dealing with selection bias. For this reason, it is important to employ a variety of specification tests which are available for determining which models or specifications are consistent with the data and which are not consistent. Examples of such specification tests are described in contributions by Heckman and Hotz (1989) and Moffitt (1991). Heckman and Hotz (1989) find, in the case of training programs, that use of a number of specification tests enables them to substantially narrow down the number of possible alternative forms which selection into the program may take, thus reducing substantially the range of possible estimates of program impact.

To date, such specification tests to assess the validity of alternative models have not been used as extensively as should have been the case. However, in part because of recent contributions to the evaluation literature, this use is increasingly becoming an important aspect of well-executed evaluations. An important side-effect of this trend is that evaluators are increasingly

required to devote thought and attention to the process by which selection into the program takes place (and therefore to the appropriate way to take account of potential selection bias), rather than to rely on a mechanical technique such as the Heckman two-step procedure.

6.5 Determinants of Program Participation

A final observation is that the more information that can be obtained about the process by which participants end up in the program and non-participants do not end up in the program, the more credible will be the estimates of program impact. As discussed previously, program participation depends on both observed and unobserved factors, and accounting for the influence of observed factors is much more straightforward and less uncertain than accounting for the role of unobserved factors. Thus acquiring richer data on the determinants of program participation is one of the most effective methods of dealing with potential selection bias.

In any evaluation there will always be some factors that are unobserved but which potentially could result in selection bias. In order to best take account of such possibilities, the greater the availability of rich qualitative information on the program and on participant and non-participant characteristics, the more able the evaluator is to choose the most appropriate method for addressing selection bias.

7.0 Selecting Outcome Measures

A critical step in a summative evaluation is to select the best measures for assessing outcomes. “An irrelevant or unreliable measure can completely undermine the worth of an impact assessment by producing misleading estimates” (Rossi and Freeman, 1993, p. 234).

Outcome measures relate to the impact the program is supposed to have, so suitable outcome measures focus on the program’s objectives. In general, HRDC training programs have all or a subset of the following goals: higher educational achievement; improved transition to the labour market with enhanced employability and earnings; reduced dependency on passive income support; and improved work attitudes. These outcomes are easily quantified and would be treated as dependent variables in the econometric models (since it makes sense to assess the program in terms of its intended effects). Thus, good outcome (post-program) variables for training programs include education level, employment status (i.e., working or not), time spent working or in school, annual earnings, months spent on social assistance, weeks spent on EI, and attitudes towards work, education and passive assistance.

It is important to note that pre-program measures of the outcome are highly desirable for any quasi-experimental evaluation. Very occasionally, the program’s management information system will have adequate pre-program data for participants and non-participants. For example, HRDC has complete and accurate data on EI use that are very useful for impact evaluations. HRDC’s Service and Outcome Measurement System (SOMS) also has good longitudinal data on employment status, and time spent in school or work. Unfortunately, reliable pre-program data on social assistance use and attitudes are very rare. Nevertheless, this information should be gathered if possible to conduct an impact evaluation. Some pre-program information

can be gathered during a follow-up survey, although problems of accurate recall and lost records render the data imprecise. Other pre-program information, especially attitudes, is impossible to reconstruct after the program.

Collecting pre-program data on the sample through a "baseline survey" is the best recourse. The fundamental purpose of a baseline survey is to establish the pre-program characteristics of participants and non-participants in support of a future summative evaluation. A good baseline survey would feature questions that aim to establish what the person was doing in terms of the outcome variables (work, school, earnings, social assistance, EI, and so on) before the program. It should explore pre-program events each year for at least two years before the program began: for example, the number of months on social assistance during 1995, 1996 and 1997. A good section on attitudes, and a section on demographics (especially if the administrative system is unreliable or non-existent) should also be included in the questionnaire. Finally, the baseline survey should ask for at least two contacts (family or friends) who can help locate the individual for follow-up purposes, because the target group for training programs tends to be very mobile.

8.0 Selecting a Comparison Group

The most important aim in constructing comparison groups is to match as closely as possible the observed factors that might predispose a person to be successful in terms of the outcome measures. A clear example would be the influence of pre-program education on earnings: there is a well-established positive correlation between education and earnings levels (Lalonde, 1995). Ideally, then, one would want the treatment and comparison groups to be similar in terms of pre-program education level. Evaluators would thus want a comparison group with means and distributions as close as possible to the trainee sample in terms of education.

Heckman et al (1995) claim to have developed an effective matching methodology based on several variables (as reported in Heckman and Smith, 1995b). Central to creating truly comparable comparison groups, they assert, is information on "labour force status transitions," especially transitions from employment to unemployment and from outside the labour force into unemployment. Other key matching variables according to these scholars are age, education, marital status and family income.

Geographic location and time period (e.g., both groups were on social assistance during a similar period) are other important matching variables (see Friedlander and Robins, 1995). Which variables to use depends on what information is available and the purposes of the program and of the evaluation.

Several different methods are used to accomplish the matching. Eligibility matching, cell matching and statistical matching procedures are popular.

In eligibility matching, cases are selected from a representative sample of the population that meet the eligibility requirements for the training program. A relevant example would be HRDC's longitudinal file, which includes all EI recipients and the training courses they have taken. Both

trainees and non-trainees could be selected from this database (presuming the training focused on EI recipients).

In cell matching, also called stratified matching, individual observations in both samples are divided into cells defined by the traits that predict the outcome variable. For example, one might pre-assign sample members to cells based on region, date of EI application, age, and so on. Cells with no trainees are eliminated, those with few cases are combined. Some studies then weight variables to equalize distributions between groups (Fraker and Maynard, 1987).

To construct a statistical match (e.g., via the closest neighbour technique) comparison group, an individual is chosen from the eligible population for each trainee based on the closeness-of-fit on predicted scores on predicted outcome, or on closeness-of-fit on traits correlated with the outcome (Fraker and Maynard, 1987). Background characteristics of each trainee are matched against those of each non-trainee; the non-trainee most closely resembling the trainee is selected.

Fraker and Maynard (1987) concluded that the general population is an inadequate source of comparison samples. Friedlander and Robins (1995) reported that statistical matching produced mediocre improvements in the accuracy of quasi-experimental estimates. Dickinson, Johnson, and West (1987) reported that the type of matching procedure (cell matching or statistical matching) did not affect results. Riddell (1991), in his review of training evaluations, agreed that the specific matching procedure used did not appear to affect impact estimates, although ". . . data bases which focus on members of the target groups may increase the accuracy of quasi-experimental estimates of programme impact."

Therefore, considering that individual characteristics can be controlled for in the model anyway, it seems safe to conclude that *the choice of matching procedure is not as important as ensuring the comparison group would meet the eligibility requirements for the program*.

9.0 Applying the Lessons: Choosing Comparison Groups for EBSM

Lessons from previous quasi-experimental evaluations can assist the process of drawing regional EBSM comparison groups:

- To date, there is little evidence that the matching procedure used makes much difference to the accuracy of estimates from quasi-experiments. Thus, the simplest course would be to ensure the comparison group meets the eligibility requirements for the training program. *At the very least the comparison group should be drawn from the same population as the treatment group*. EBSM programs, which are primarily aimed at EI recipients, could use the HRDC longitudinal file for choice of comparison group subjects, for example. The NESS database may also be a useful source of comparison group members.

- Individual components of EBSM have specific eligibility criteria that should be taken into account when selecting a comparison group. For instance, targeted wage subsidies are largely focused on those who face particular obstacles to employment such as disabilities. The sample frame for the comparison group should come from the subset of the EI population who share these obstacles.

- Moving one step beyond simple eligibility matching is advisable in the case of regional EBSM programs: it seems clear that comparison groups should at least come from the same region. In the case of Transitional Jobs Fund, the comparison group should be chosen from the same communities as the treatment group (i.e., those with unemployment rates of 12% or higher). If desired, and assuming the requisite information is available from the database used, other potentially important matching variables are labour force status changes, age, education, marital status, family income, and time period.

- Sampling strategy depends on the matching procedure employed. The easiest approach — and one that would be no less satisfactory than more complex ones judging by the literature — would be to first limit the population to those in the region who were on EI during the period that the program was in operation (plus any other eligibility factors unique to each EBSM component). Then a simple random sample could be selected. A more precise matching strategy would be more difficult to implement, but by no means impossible. For example, if one wanted to match by sex, age and education, a series of dummy variables (0-1) could be set up: men=0, women =1; under 30 years old=0, 30+=1; less than high school=0, high school gradute=1. A three-digit variable could then be computed for each participant and non-participant with the first digit representing sex, the second age, the third education (e.g., a 29 year old female high school graduate = 101). The software program could then be used to select at random, a match (or more than one match) for each participant.

- As for sample size, standard formulas should be employed, keeping in mind that they calculate margins of error for final sample size rather than the original sample size. Appendix D summarizes how to choose an appropriate sample size.

10.0 Conclusion

This paper has presented the basics of quasi-experimental evaluation design, with a focus on considerations for controlling selection bias and for choosing comparison groups. Quasi-experimental designs are usually the best approach to use for evaluating social programs because one-group designs are inadequate for establishing cause and effect, and two-group experimental designs are very often impractical. Complex analytical procedures are required to isolate the effect of the program from many other potential causes under this design, however.

Appendix A: True Experiments Versus Quasi-Experiments

There has been a lively debate for over a decade on whether true experiments or quasi-experiments represent the superior design. This appendix will present each side's argument.

The Experimental Camp

Those favouring the experimental approach question the validity of highly sophisticated non-experimental approaches for assessing the impact of training and employment programs because of the apparent difficulty in obtaining reliable estimates of the impact of labour market programs. They claim that economists using quasi-experimental techniques have had little success in isolating program effects (i.e., removing the "selection bias").

Scholars such as Ashenfelter and Card (1985), Barnow (1987), and LaLonde and Maynard (1987) contend that results of dozens of econometric studies were so varied that there may be no sound way to adequately measure program effects short of an experimental evaluation with random assignment to training or control groups.

The findings of Lalonde (1986) and Lalonde and Maynard (1987) were exceptionally damaging to the case for non-experimental designs. To assess the accuracy of such designs, they compared the results from a true experiment of the National Supported Work (NSW) Demonstration Project to those derived from various widely used non-experimental procedures to see if they could accurately estimate the true program impacts. They concluded that "there does not appear to be any formula [using non-experimental methods] that researchers can confidently use to replicate experimental results of the Supported Work Program." Using the same evidence, Fraker and Maynard (1987) concluded:

> This analysis demonstrated that results may be severely biased depending on the target population, the comparison group selected, and/or the analytic model used. More importantly, there is at present no way to determine a priori whether comparison group results will yield valid indicators of the program impacts (p. 216).

More recent work confirms the chief conclusions drawn from this work – that quasi-experimental estimators are biased and sensitive to minor changes in model specification. Friedlander and Robins (1995) assessed two conventional quasi-experimental strategies against experimental data from four social assistance reform experiments: comparing the treatment group in one locale to a comparison group in another locale; and comparing outcomes of the treatment group with a pre-program comparison group in the same area. They concluded that the non-experimental estimates were "usually quite different from the experimental estimates," especially for the comparison samples drawn from different areas. They also studied two statistical

techniques for improving the accuracy of quasi-experimental estimates: statistical matching to produce closely matched comparison groups; and "specification tests," to statistically assess the econometric model employed to determine if the estimates it yields are accurate[7]. Neither strategy markedly improved the accuracy of the non-experimental estimates.

Further support for the experimental method comes from Greenberg and Wiseman (1992), who stated that fifteen years of experimentation (e.g., income maintenance experiments, Supported Work Demonstration) have demonstrated that random assignment was "a methodologically superior approach to program evaluation," and provided evidence that such studies were feasible. Moreover, the work on the Omnibus Budget Reconciliation Act evaluations (by the Manpower Demonstration Research Corporation) convinced decision-makers at the U.S. Department of Health and Human Services — which usually funds evaluations of training programs — that random assignment should be used as the foundation of social assistance reform evaluations. MDRC's evaluations were generally considered excellent, helping to establish random assignment as the procedure of choice.

Another advantage of experimentation is that the results are understandable and convincing to policy makers (Burtless, 1995). Without the complicated qualifications associated with quasi-experiments, analysts can present straightforward experimental findings such as "the program raised annual earnings of participants by $1,000." This simplicity makes it more likely that policy makers will use the evaluation findings. "They do not become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid. Politicians are more likely to act on results they find convincing" (Burtless, 1995, p.67).

Finally, a National Academy of Sciences panel recommended the following conditions as necessary (but not sufficient) for quality research: *the use of random assignment to group*; reasonable operational stability of the program prior to final assessment; adequate sample coverage and low rates of sample attrition; outcome measures that well represent program objectives, both immediate and longer-term; and a follow-up period that allows time for program effects to emerge or decay (Gueron and Pauly, 1991).

These studies, among others, have convinced many evaluators that experimental estimators are better than quasi-experimental ones (e.g., Burtless 1995; and Friedlander and Robins 1995).

---

[7] For example, one can see if the model correctly predicts no differences in outcomes between trainees and non-trainees during the period before the program began. If the model finds statistically significant differences between groups before the program, it should be dropped since it failed the specification test.

The Quasi-Experimental Camp

Others, however, have contested the view that experimental designs are superior to non-experimental designs. In the forefront of this group are James Heckman and his colleagues (e.g., Heckman, Hotz, & Dabos, 1987; Heckman and Smith, 1995). They argue that with a sufficiently rich data set and appropriate econometric modeling techniques, it is possible to arrive at reliable estimates of impact.

To mount a serious challenge to the emerging consensus that experimentation was the better route, they had to surmount the conclusion of Lalonde and his associates that no available quasi-experimental method produced estimates close to those of the unbiased experimental estimates. Heckman and Hotz (1989) applied some simple specification tests to models estimated on data from the NSW experiment and found that the most inaccurate models could be rejected, leaving a subset of models that produced impact estimates close to those of the experimental results. Heckman and Smith (1995) presented some cogent arguments that undermine to some extent Lalonde's findings: sample sizes were too small and insufficient geographical data were available to place comparison group members in the same local labour market as participants; there was only one year of pre-program data available, ruling out potentially effective econometric strategies (and leaving the estimates subject to the "Ashenfelter Dip"[8]); the studies did not use a variety of standard specification tests; and there have been important advances in non-experimental methods since the these studies were done. For instance, Heckman and Smith (1995b) found that labour force dynamics (i.e., movements between employment, unemployment and into and out of the labour force), as well as other contributing factors such as age, education, marital status and family income, can be used to form comparison groups that are "virtually identical" to treatment groups, thereby controlling for selection bias.

Heckman and Smith (1995) also raise some serious objections to the experimental method, both theoretical and empirical:

- Randomization may alter the pool of persons eligible for the program or change the behaviour of participants, the so-called "randomization bias." For instance, in order to form a control group, occasionally the pool of potential participants must be expanded, usually by relaxing some eligibility criteria.

- If close substitutes for the experimental treatment are available, the assumption that the control group receives no treatment is annulled to the extent members take advantage of the training, generating a "substitution bias." For instance, some training programs consist of buying seats at community colleges; non-participants can choose to attend the same college course if they pay on their own or find another source of funding.

---

[8] This refers to the phenomenon that the earnings of training program participants tend to dip just before they enter training, because unemployment is often the impetus to take the course; thus, before and after difference estimators will tend to overestimate the effect of the program

- Experimental data cannot answer many of the questions of central interest to policy makers, including the median impact of the program and the proportion of participants with a positive (or negative) impact from the program. No parameters that depend on the joint distribution of outcomes in the treatment and control groups can be estimated. “Only if the evaluation problem is defined exclusively in terms of means can it be said that experiments provide a precise answer” (Heckman and Smith, 1995, p. 22).

- Experimental evaluations tell policy makers whether the programs work or not. They do not typically shed any light on why a program worked or did not work.

- Institutional factors make it hard to carry out random assignment at all in some cases and difficult to do an optimal job in others. Staff can subvert the process because the are opposed to random assignment. Costs can militate against making the selection at the optimal point in the decision process[9]. Multi-stage randomization, which is expensive and disruptive to the program, may be required to generate estimates of the impacts of differ-ent services.

- Complementary non-experimental analyses are often required to overcome the short-comings of experimental models.

Burtless (1995) agrees that experimental evaluations have such shortcomings, but counters that quasi-experiments also share some of them (e.g., quasi-experiments can also be costly and disruptive); plus they are “usually plagued by more serious statistical problems than those that occur in randomized trials.” Moreover, he adds, that nothing inherent to the experimental design precludes researchers from using non-experimental methods to analyze data.

Conclusion

The jury is still out as to whether quasi-experimental designs can adequately control selection bias. It is safe to conclude that experimental designs are superior in this critical respect. But the many problems associated with experiments render them impractical for many if not most evaluations. Quasi-experimental designs are often the best practical approach to take for evaluations of training programs.

[9] Randomization can occur at any step of the process: participant becomes eligible, becomes aware of the program and his/her eligibility for it, applies, is accepted, is assessed by staff, is assigned to particular services, begins receiving the services, and completes the program. The optimal place for randomization depends on the evaluation questions at issue. For determining mean impact, the assignment should take place as close as possible to the commencement of training to minimize attrition.

Appendix B: Brief Mathematical Representation of Quasi-Experiments

The problem of trying to evaluate the impact of a social program in a non-experimental setting may be represented as follows (Moffitt, 1991):

$$Y^{**}_{it} = Y^{*}_{it} + \alpha$$

$$\alpha = Y^{**}_{it} - Y^{*}_{it}$$

where

Y^{*}_{it} = level of outcome variable for person i at time t if he had not participated

Y^{**}_{it} = level of outcome variable for same person at same time if he participated previously

Evaluations aim to estimate α, the treatment effect. That is, we wish to estimate for those who have participated what Y would have been had they not participated. Clearly, we cannot know Y^{*}_{it} since these individuals ***have*** used the program. So we substitute Y^{*}_{it} of non-participants to estimate α:

$$\alpha = E(Y^{**}_{it} | d_i = 1) - E(Y^{*}_{it} | d_i = 0)$$

where $d_i = 1$ if person i has participated

$d_i = 0$ if person i has not participated

and | denotes "conditional on" so that the first term on the right hand side of the above equation is the mean value of Y for participants and the second term on the right hand side is the mean value of Y for non-participants.

In words, we estimate the treatment effect by estimating the expected value (E) of Y, say annual earnings, for those who have participated in a training program, and subtracting the expected value of Y for those who have not. Only if pre-program $E(Y^{**}_{it})$ for participants equals pre-program $E(Y^{*}_{it})$ for non-participants will there be no bias. But this will seldom be the case because of selection bias.

Appendix C:
Dealing With Selection Bias

Undoubtedly the most popular approach to dealing with selection bias is the Heckman (1979) *two-stage approach*. The first stage involves modeling selection into the program. Usually this takes the form of a single equation explaining program participation/non-participation[10]:

$$P = ßX + U$$

where P is a dummy variable (1 for participants and 0 for non-participants), X is a set of all observed factors that may account for participation in the program (e.g., age, sex), and U is a random error term which is assumed to be normally distributed to take account of unobserved factors that influence participation in the program. From this equation, the inverse of Mill's ratio is computed, which is then inserted into a second stage outcome equation to estimate program impact (usually via ordinary least squares):

$$Y = ßX + \alpha P + \delta M + U$$

where Y is the outcome of interest, X is a vector of observed variables, P is the participation dummy, and M is the inverse of Mill's Ratio. If the assumptions underlying the model are correct, the Heckman procedure removes the selection bias (δ), thereby producing an unbiased estimate of program impact (α). (The measure of program impact is the estimated coefficient on the indicator variable for participation/non-participation in the program.) If this equation were estimated by ordinary least squares without the inclusion of the selection bias correction term, the estimates would potentially be biased. However, if the model is properly specified, the addition of the "selection bias correction" variable removes this potential bias, thus giving unbiased estimates of program impact.

Another powerful means of controlling for differences between groups is called the *differences-in-differences* method. Longitudinal data are collected for key outcome measures 3/4 e.g., earnings, social assistance use. To account for the differences in the participant and non-participant samples a longitudinal estimator of program impact is employed; such estimators take account of the level of the outcome variable prior to and after the program, in contrast to cross-sectional estimators which use data on post-program outcomes alone. This estimator uses the pre- vs. post-program change in the outcome variable for non-participants as an

---

[10] In most cases, the equation is estimated as a "probit model" which is appropriate if the random term in this equation is normally distributed. (A probit is a measurement of probability based on deviations from the mean of a normal frequency distribution. It is analogous to multiple regression but with a dischotomous dependent variable.) However, two step prodedures for situations in which the assumption of normally distributed random terms is unlikely to hold are available (through not yet widely used).

estimate of the change that would have occurred for participants in the absence of the program. The estimated average program impact is then the difference between the pre- vs. post-program change in the outcome variable for participants and the pre- vs. post-program change in the outcome variable for non-participants. This permits a determination of the incremental impact of the program by controlling for biases caused by unobserved individual differences. A multivariate analysis can then show how the size of the differences-in-differences estimate of program impact varies according to various individual and program characteristics.

In equation form (Moffitt, 1991):

$$Y = E(Y^{**}_{it} - Y^{*}_{i,t-1} \mid d_i = 1) - E(Y^{*}_{it} - Y^{*}_{i,t-1} \mid d_i = 0)$$

where t = the posttreatment point, t-1 = pretreatment point, and

$Y^{*}_{it} - Y^{*}_{i,t-1}$ = change in Y^{*}_{it} from t-1 to t if treatment not received

$Y^{**}_{it} - Y^{*}_{i,t-1}$ = change in Y^{*}_{it} from t-1 to t if treatment received

Instrumental variable (IV) methods are widely used in situations in which ordinary least squares (OLS) estimates may be biased due to a correlation between one or more of the explanatory variables and the random error term in the model. In the evaluation/selection bias context, such potential bias arises because of the possible correlation between the participation/non-participation variable and the random error term in the outcome equation. This potential bias can be removed if one or more “instrumental variables” are available and included in the model.

The basic model of program outcome or impact is:

$$Y = \beta X + \alpha P + U \quad (1)$$

This can be written as:

$$Y = CW + U \quad (2)$$

where $C = (\beta\ \alpha)'$ and W = (X P) using matrix notation.

The least squares estimator of (2) is given by:

c = (W’W)-1 W’Y where (W’W)-1 is the inverse of the matrix (W’W). In general this estimator is biased because of the correlation between W and U (i.e. E{W’U} does not equal zero, where E{ } represents the expectations operator).

The IV estimator of (2) is given by:

c* = (Z’W)-1 Z’Y where Z is the matrix of instrumental variables. This estimator is in general unbiased because Z and U are uncorrelated, i.e. E{Z’U} equals zero if Z is an appropriate instrument.

Appendix D: Determining Appropriate Sample Size

What sample size should be chosen for a particular survey? It depends mainly on tolerable error, population size, the importance of particular subgroups, anticipated level of non-response, and how much money is available.

"Tolerable error" refers to the margin of error for the survey. Whenever the results of polls are reported in the news, the margin of error – e.g., plus or minus 3%, 19 times in 20 – is included. The margin of error tells the reader how accurate the poll's findings are. It is based on the "standard error," the measure of how much the sample mean differs from the population mean. The margin of error adjusts the standard error to account for any potential differences between the sample and the population via calculation of a "confidence interval" for the population mean. Traditionally, a 95% confidence interval is used (i.e., 19 times in 20). The tolerable margin of error is usually between 3% and 5% (much lower and the costs of the survey begin to rise dramatically).

The traditional formula for large population sizes is $n = 1.96^2\ p(1-p)/SE^2$, where n is sample size to be calculated, SE is the tolerable standard error, and p is the proportion having the characteristic being measured and (1-p) is the proportion who lack it (e.g., if 48% said yes, 52% must have said no). The 1.96 figure reflects the choice of a 95% confidence interval (in a normal distribution, 95% of the area under the curve is within 1.96 standard deviations of the mean). For example[11], if a margin of error of ± 3%, 19 times in 20 was tolerable, the following sample size would be required:

$$n = 1.96^2\ (.5*.5)/.03^2 = 1{,}068$$

Population size is a consideration only when it falls below 100,000 or so. Below that, something called the "finite population correction factor" must be used to determine sample size. The correction factor = $(N - n\ /N - /1)^{1/2}$, where N is the population size and n is the sample size.

Algebraically entering this factor into the sample size equation, yields:

$$n = (1.96^2\ p(1-p)N)/\ (1.96^2\ p(1-p)) + (N-1)SE^2$$

[11] By convention, p and 1-p are set to the most conservative level – .5 for each.

For example, if an evaluator wanted to learn how many EI clients to survey from a sample frameof 2,500, with an error rate of ± 3% at 95% level of confidence:

$$n = (1.96^2 * .25(2500))/ (1.96^2 * .25) + 2499(.0009) = 749$$

The sampling error associated with subgroups will be higher than that for the whole sample, because there are obviously fewer cases. A rule of thumb is that there should be a minimum of 100 individuals in any major subgroup that will be analyzed separately. This will achieve at least a ± 10% margin of error for each major stratum, the maximum tolerable (Rea and Parker, 1992).

Note that when choosing a sample size, there will always be some people in the sample who can't be located, or who will refuse to cooperate. Allowances must be made for anticipated non-response. For a *final* sample size of 1,000, with a 50% response rate, the *initial* sample size must be 2,000. Given a fixed budget, there is always a tradeoff between the initial sample size and the effort to reduce non-response. Too often a large initial sample is chosen and too little effort is expended in reducing non-response, with consequent effects on total error.

References

Ashenfelter, O. and D. Card (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics.* 67:648-660.

Barnow, B. (1987) The impact of CETA programs on earnings: a review of the literature. *Journal of Human Resources.* 22:157-93.

Burtless, G. (1995) The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives.* 9(2): 63-84.

Campbell, D.T. & J.C. Stanley (1971) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally & Co.

Dickinson, K.P., Johnson, T.R., and R.W. West (1987) An analysis of the sensitivity of quasi-experimental net impact estimates of CETA programs. *Evaluation Review.* 11:452- 472.

Fraker, T. and R. Maynard (1987) The adequacy of comparison group designs for evaluations of employment-related programs*. Journal of Human Resources.* 22:194-227.

Friedlander, D. & P.K. Robins (1995) Evaluating program evaluations: new evidence on commonly used nonexperimental methods. *American Economic Review*. 85(4):923-937.

Greenberg, D. & M. Wiseman (1992) What did the OBRA demonstrations do? In C.F. Manski & I. Garfinkel (Eds.) *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.

Gueron, J.M. & E. Pauly (1991) *From Welfare to Work.* New York: Sage. Heckman, J.J. (1979) Sample selection bias as specification error. *Econometrica* 47: 153- 161.

Heckman, J.J., V.J. Hotz, and M. Dabos (1987) Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review.* 11:395-427.

Heckman, J.J.& V.J. Hotz (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association.* 84:862-877.

Heckman, J. and R. Robb (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, edited by J. Heckman and B. Singer, 156- 246. Cambridge: Cambridge University Press.

Heckman, J.J & J.A. Smith (1995) Assessing the case for social experiments. *Journal of Economic Perspectives*. 9(2):85-110.

Heckman, J.J & J.A. Smith (1995b) Ashenfelter's dip and the determinants of participation in a social program: implications for simple program evaluation strategies. Unpublished manuscript, University of Chicago.

Heckman, J.J., H. Ichimura, J.A. Smith & P. Todd (1995) Nonparametric estimation of selection bias using experimental data. Unpublished manuscript, University of Chicago.

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

Lalonde, R.J. (1995) The promise of public sector-sponsored training programs. *Journal of Economic Perspectives*. 9(2):149-168.

Lalonde, R.J. and R. Maynard (1987) How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review,* II:428-451.

Mark, M.M. & T.D. Cook (1984) Design of randomized experiments and quasi-experiments. In L. Ruttman (ed.) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Moffit, R. (1991) Program evaluation with nonexperimental data. *Evaluation Review*, 15:291-314.

Riddell, C. (1991). Evaluation of manpower and training programmes: The North American experience. In OECD (Ed.) *Evaluating Labour Market and Social Programmes.* Paris:OECD.

Rubin, D.B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*. 74:318- 328.

Patton, M.Q. (1982) *Practical Evaluation*. Beverly Hills, CA.: Sage. Posavac, E.J. & R.G. Carey (1980) *Program Evaluation: Methods and Case Studies*. Englewood Cliffs, N.J.: Prentice-Hall Inc.

Rea, L.M. & R.A. Parker (1992) *Designing and Conducting Survey Research*. San Francisco: Jossey-Bass.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th ed.). Newbury Park, California: Sage.

Stufflebeam, D.L. & A.J. Shinkfield (1985) *Systematic Evaluation*. Boston: Kluwer Academic Publishers.