

Evaluating Local Economic Development Policies:  
Theory and Practice

Prof. Jeffrey Smith  
Department of Economics  
University of Maryland  
[smith@econ.umd.edu](mailto:smith@econ.umd.edu)

Version of January 12, 2004

This chapter has benefited from discussions with and/or comments from Tim Bartik, Dan Black, Michael Lechner, Alistair Nolan, Miana Plesca, Elliot Stern and Alex Whalley, from reading the chapters by Tim Bartik and by Randy Eberts and Chris O'Leary (which were written before this one) and from discussions at the OECD conference in Vienna in November 2002.

## **1.0 Introduction**

Policies and programs undertaken to increase local economic development by governments and by private agencies may have positive effects, or they may not. In some cases, a lack of effects may result from poor program design or inadequate funding. In other cases, a lack of effect may result from the fact that the program really exists to funnel money to politically influential firms, individuals or groups, with the local economic development justification used as cover. When programs do not produce benefits in terms of local economic development, finding this out allows scarce funds to flow into other, more beneficial activities, or back to the long-suffering taxpayer. When programs do produce benefits, finding this out can generate political support for program persistence or even expansion.

Evidence on the efficacy of local economic development policies and programs comes from evaluations. This chapter presents an overview of the current literature on how to evaluate programs. The scholarly literature on program evaluation has advanced rapidly over the past fifteen years. For example, major developments in regard to “heterogeneous treatment effects” – different program impacts for different persons, firms, counties, cities or groups affected by a policy – affect both evaluation practice and how to think about evaluation design and interpretation. Similarly, important technical developments in non-parametric and semi-parametric methods allow much more flexible use of the available data, but at the same time create a demand for the high quality data that such methods require to produce reliable estimates. Social experiments have become routine (at least in the United States) in areas such as the evaluation of public employment and training programs. Unfortunately, evaluation practice, to a large extent,

remains mired in the 1970s. One of the main goals of this chapter is to provide a practical, relatively non-technical guide to these advances.

This chapter addresses some of the same issues as the chapters by Tim Bartik (2004) and by Randy Eberts and Chris O’Leary (2004), but with enough differences to make it a complement to, rather than a substitute for, those chapters. Five differences in particular deserve notice. First, this chapter devotes much more attention to the different econometric evaluation estimators in the literature, and provides a wealth of pointers into the rapidly expanding literature on the subject. A key theme of the chapter is the choice of an appropriate estimator given the available data, economic environment and institutional characteristics of the program being evaluated. Second, this chapter devotes more attention to the emerging literature on heterogeneous treatment effects, and how such effects influence evaluation design and interpretation. Third, this chapter worries more about the implications of general equilibrium effects for policy evaluations. Fourth, this chapter emphasizes that doing an evaluation may not make sense in all cases, particularly for smaller programs. Time spent reading the literature for good evaluations of similar programs may yield more useful results than a weak evaluation based on poor data completed by a poorly qualified evaluator using inappropriate methods. Finally, the perspective underlying this chapter is that evaluation, taken seriously, represents a method for ensuring that program managers further the goals of their principals – namely taxpayers and donors – rather than simply transferring resources to interested stakeholders, such as program operators, politically favored firms, or themselves. In practice, many low quality evaluations exist mainly to cover up exactly such behavior; for precisely this reason it is important to be very clear about what constitutes a good

evaluation and to design institutions that will reduce the flow of misleading, low-quality evaluations.

The remainder of the chapter is organized as follows. Section 2 describes the evaluation problem and discusses parameters of interest. Section 3 provides an overview of the theory of econometric program evaluation at a relatively non-technical level, and with plenty of pointers to the literature. Section 4 reviews the two leading serious alternatives to econometric program evaluation: participant self-evaluation and administrative performance standards. Section 5 discusses the practice of evaluation, in the broad sense of the choices facing an organization considering undertaking an evaluation, such as whether or not it is worth it to do an evaluation, who should do the evaluation, and how to make sure that it is any good. Section 6 concludes and restates the main themes of the chapter.

## **2.0 Programs and Parameters**

### *2.1 Types of Local Economic Development Programs*

Local economic development programs include a wide range of initiatives, from programs designed to improve the human capital of individual workers, to financial and in-kind subsidies to professional athletic teams, to enterprise zones, to tax subsidies designed to lure particular businesses, and on and on. Bartik (2004) presents a nice list in his Table 1a that includes a somewhat narrower set of activities than I have in mind here; Bartik (2003a) describes these policies in greater detail.

For this chapter, two dimensions of such programs hold particular relevance, as they shape choices regarding data collection and evaluation methods, as I describe in

detail below. The first dimension consists of the units directly treated by the intervention. Depending on the program, this could be individual workers, some or all firms in an area, cities, towns or districts, or entire states or countries. The second, related, dimension consists of the units that theory suggests the program will affect. In some cases, particularly for programs not expected to have much in the way of external effects, these two dimensions may coincide. For example, small-scale human capital programs may have little effect on individuals other than those receiving the additional human capital. In other cases, the two dimensions will not coincide. For example, a program may have positive effects on treated units and negative effects on untreated units, as when subsidizing one class of firms but not their competitors. In still other cases, programs may produce positive spillovers, as when a new park attracts new businesses and residents to an area, and increases property values in the surrounding neighborhoods.

## *2.2 Notation*

Popular and policy discussions of economic development programs often focus on their “effects,” as though the “effects” of a program represent a single well-defined entity. That programs have a variety of effects represents an important theme of this chapter. In the academic literature, this discussion falls under the heading of heterogeneous treatment effects. That literature discusses how the notion of a program’s effects changes and broadens when we consider that a program may have a different effect on each unit that participates in it and, in some cases, even on units that do not participate in it.

To make this point more clearly, I now introduce some very simple notation, which will serve to make meanings precise throughout the chapter. However, the chapter

is written so that it does not require an understanding of the notation to get the point; severely notation-averse readers can simply skim over it.<sup>1</sup>

Let  $Y$  denote some outcome variable. For an individual, it might be earnings, employment, or health status. For a firm it might be profits, sales, or employment. For a locality it might be population, or some measure of air quality or economic growth. Now imagine two worlds for each unit, one world where the unit participates in the program under study and one where it does not. We can imagine the unit's value of  $Y$  in each of those worlds, and we label the value in the world where the unit participates as  $Y_{1i}$  and the value in the world where the unit does not participate as  $Y_{0i}$ , where “ $i$ ” refers to a particular unit.

### *2.3 Parameters of Interest*

Using this notation, the effect of a program on unit “ $i$ ” is given by

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

In words, the literature defines the effect of a program on unit “ $i$ ” as the difference in outcomes between a world where that unit participates and a unit where it does not. The evaluation problem then consists of estimating whichever one of the two outcomes we do not observe in the data.

Many of the various parameters of interest defined and examined in the literature on program evaluation then consist of averages of the unit-specific impact ( $\Delta_i$ ) over

---

<sup>1</sup> Of course, adding some technical skills would not only represent personal development on the part of notation-averse economic developers, it might also improve their ability to promote local economic development.

various policy-relevant sets of units. The most common parameter of interest is the Average Treatment Effect on the Treated (ATET), or just “treatment on the treated” for short. This parameter indicates the average effect of the program on current participants. In terms of the notation just defined, it equals

$$\Delta^{TT} = E(\Delta_i | D_i = 1) = E(Y_{1i} - Y_{0i} | D_i = 1),$$

where  $D_i$  is a dummy variable for current participants, so that  $D_i = 1$  for units that participate in the program and  $D_i = 0$  otherwise, and  $E$  denotes the expectations operator, where the expectations are conditional on the condition to the right of the vertical bar (“|”). An estimate of the ATET, combined with an estimate of the average cost of the program per participating unit, allows a cost-benefit analysis of the question of whether to keep or scrap an existing program.

The Average Treatment Effect (ATE) represents a second parameter of potential interest. This parameter averages the effect of treatment over all of the units in a population, including both participants and non-participants. In terms of the notation,

$$\Delta^{ATE} = E(Y_{1i} - Y_{0i}).$$

The ATE answers policy questions related to universal programs – programs where every unit in some well-defined population participates. When considering making a voluntary program mandatory, policymakers need precise estimates of the ATET and the ATE, which may differ strongly if those units that choose not to participate in a voluntary program do so because it would have only a small, or even negative, effect for them. An example here would be taking a voluntary program of job search assistance or job training for displaced workers and making it mandatory (or nearly so by, for example, requiring participation in order to receive full unemployment insurance benefits).

A third category of parameters consists of Marginal Average Treatment Effects, or MATEs. A marginal average treatment effect measures the average effect of a program among a group at some relevant margin. For example, suppose that the program under consideration presently serves firms with fewer than 20 employees, and the proposal under consideration consists of expanding it to include firms with 21 to 30 employees. In this situation, the parameter of interest consists of the average effect of participation on firms with 21 to 30 employees. This parameter may differ from the ATET, which would give the average effect on existing participant firms (those with 1 to 20 employees), and could be either higher or lower, depending on the nature of the program treatment and its relationship to firm size. Comparing a MATE to the corresponding marginal cost of expanding (or contracting) a program provides a cost-benefit analysis for the program expansion or contraction. Note that a different MATE applies to each margin – expanding a program to include one set of units may yield different results than expanding it to include another set of units. A final category of parameters, called Local Average Treatment Effects, or LATEs, is discussed in Section 3.4.

The parameters presented so far may or may not capture general equilibrium effects, depending on the design of the analysis. General equilibrium effects are those other than the immediate effects on the treated units, and result from changes in the behavior of untreated units in response to the program. Such changes may occur directly (a firm with 11 employees fires one in order to become eligible for a program that serves firms with 10 or fewer employees) or indirectly through changes in prices, as when a tuition subsidy increases the supply of skilled workers and thereby lowers their wage.

Consider a state-level program that subsidizes training at a particular class of firms. Some states have the program and others do not. An estimate of the ATET on the firms receiving the subsidy will capture only the direct effects on the employment, productivity, sales, and so on for those firms. In contrast, an estimate of the ATET on the states adopting the subsidy will capture any general equilibrium effects at the state level, including reductions in employment at unsubsidized firms, but not general equilibrium effects that operate across state boundaries.

Bringing general equilibrium effects into the picture adds some additional parameters of interest. For example, we might now have some interest in what the literature calls the Average Effect of Treatment on the Non-Treated (ATNT). This parameter measures the average effect of the program on units that do not participate in it, either because they choose not to or because they do not meet the eligibility criteria. To see this, consider the case of a program that provides job search assistance to particular groups of workers. These workers now search more, and more intelligently, than before, and we would expect them to find jobs faster. But what happens to others in the labor market? First, some jobs that would have been filled by others now get filled by individuals who receive the job search assistance. As a result, they find jobs more slowly. Second, it may make sense for them to change their search intensity as well. Both factors lead a program that provides services to one group to have effects on other groups – effects that matter in assessing the value of the program. Calmfors (1994) provides a useful (and relatively accessible) introduction to general equilibrium issues in the context of active labor market policies.

### 3.0 Theory

This section provides a brief introduction to each of the main categories of econometric evaluation methods. Sections 3.1 to 3.6 each consider one category of method, and Section 3.7 considers how to choose among them. Although this chapter presents the various estimators as though they are dishes on a buffet, where the evaluator can choose which one to use based on its having a cool name, or its association to famous people, or its being the estimator *de jour*, in fact, estimator selection, properly done, must adhere to strict rules. Each of the categories of estimators examined in the following sub-sections provides the correct answer only under certain assumptions. An evaluator choosing an estimator must carefully consider the nature of the available data, the institutional nature of the program – particularly how participation comes about – and the parameters of interest. In some cases – and this constitutes another one of my themes – a lack of good data may mean that no estimator is likely to provide a correct answer, in which case the evaluator should simply stop and report this fact.

For readers wanting to learn more, the literature provides a number of other surveys of all or part of this material, ranging from the very non-technical to the very technical. At the less technical end, see Moffitt (1991, 2003), Winship and Morgan (1999), Smith (2000), Ravallion (2001), and Smith and Sweetman (2001). At a moderate technical level, see Angrist and Krueger (1999), Blundell and Costa Dias (2000,2002), and Heckman, LaLonde and Smith (1999), except for Section 7. For strongly technical presentations see Section 7 of Heckman, LaLonde and Smith (1999) and Heckman and Vytlacil (2004). Some standard econometrics texts contain presentations that emphasize

the issues focused on in the evaluation literature. In this regard, see Wooldridge (2002) at the undergraduate level and Green (2002) or Wooldridge (2001) at the graduate level.

### *3.1 Social Experiments*

Social experiments represent the most powerful tool in the evaluator's toolbox, but just as that favorite wrench may not make a good screwdriver, so social experiments serve the evaluator better in some contexts than in others. To see why evaluators like social experiments, consider a treatment with no external effects, and suppose that we seek to determine the impact of treatment on the treated. The primary problem in evaluation research (almost) always consists of non-random selection into treatment. Because of non-random selection into treatment, one cannot simply compare the outcomes of treated units with the outcomes of untreated units in order to determine the impact of treatment. In terms of the notation defined above, we cannot rely on the average outcomes of untreated units,  $E(Y_0 | D = 0)$ , to accurately proxy for the outcomes that treated units would have experienced, had they not been treated,  $E(Y_0 | D = 1)$ . Finding a good approximation to this counterfactual represents the tough part of estimating the treatment on the treated parameter (because the outcomes of participants,  $E(Y_1 | D = 1)$ , appear directly in the data). The problem of non-random selection into treatment is called the selection bias problem in the econometric evaluation literature. It is important to distinguish the classical selection bias problem of selection on the untreated outcome with non-random selection into treatment based on the effect of treatment. This latter type of selection has only recently received substantial attention in the literature; this type

of selection is what makes, for example, the mean impact of treatment on the treated different from the average treatment effect in programs that do not treat all eligible units.

Social experiments solve the selection bias problem by directly constructing the usually unobserved counterfactual of what participating units would have experienced, had they not participated. In particular, in a social experiment, units that would otherwise have received the treatment are randomly excluded from doing so. The outcomes of these randomly excluded units, under certain assumptions, provide an estimate of the missing counterfactual mean, given by  $E(Y_0 | D = 0)$ . This ability to obtain the counterfactual under what, in many (but not all) contexts represent very plausible assumptions, defines the power of experiments, and explains their attraction to evaluators.

As the virtues of experiments are fairly well known, and also extensively detailed in Bartik's chapter, I focus instead on some of the conceptual issues and limitations associated with experiments. The purport of this discussion is not to provide cover to those who want to avoid doing experiments because they wish to maintain an aura of uncertainty about the impacts of the programs they love (or benefit from financially, which often amounts to the same thing). Rather, it is to make it so that experiments do not get used when they do not or cannot answer the question of interest, and to make sure that they get interpreted correctly when they are used.

The first limitation of experiments is that they cannot answer all questions of interest. This limitation has three facets, which I cover in turn. First, randomization is simply not feasible in many cases. The evidence suggests that democracy increases economic growth, but we cannot randomly assign democracy to countries. Similarly,

political factors may prohibit randomization of subsidies to firms or randomization of development grants to cities and towns.

Second, experimental data may or may not capture the general equilibrium effects of programs. Whether or not they do depends on the units affected by any equilibrium effects and the units that get randomized in the experiment. If there are spillovers to units not randomized, as when a program for small firms has an effect on medium-sized firms, these effects will be missed. Similarly, positive spillovers will get missed in an evaluation that randomizes only treated firms, rather than randomizing at the locality level.

Finally, experiments provide the distribution of outcomes experienced by the treated, and the distribution of outcomes experienced by the untreated. They do not provide the link between these two distributions; put differently, experimental data do not indicate whether a treated unit that experienced a very good outcome would also receive a very good outcome had it not received treatment. In technical terms, an experiment provides the marginal outcome distributions but not the joint distribution. As a consequence, without further, non-experimental, assumptions, experimental data do not identify parameters that depend on the joint distribution of outcomes, such as the variance of the impacts. See Heckman, Smith and Clements (1997) for an extended discussion of this issue and a variety of methods for obtaining estimates of these parameters.

The second major limitation of experiments is that practical difficulties associated with the implementation of the experiments can sometimes complicate their interpretation. Readers interested in more general treatments of the implementation of social experiments should consult, e.g., Orr (1998), as well as the implementation reports

or summaries associated with major experimental evaluations, such as Hollister, Kemper and Maynard (1984), Doolittle and Traeger (1990), Newhouse (1994) and so on. The *Digest of the Social Experiments*, compiled by Greenberg and Shroder (1997), presents a comprehensive list of all the social experiments, along with pointers to details about their design, implementation and findings.

First, because an experimental evaluation tends to have a greater disruptive effect on local program operation than a non-experimental evaluation, experiments in decentralized or federal systems, such as those in the U.S. and Canada, often have problems with external validity, because of non-random selection of local programs into the experiment. This was an issue in the U.S. National JTPA study, where over 200 of the 600 local training centers were approached in order to find 16 willing to participate in the experimental evaluation. Other than trying to keep the experimental design relatively unobtrusive, and offering side payments (about US\$1 million was devoted to this in the JTPA Study), little can be done about this other than comparing the characteristics of participating and non-participating local programs and avoiding overly ambitious generalizations about the results.

Second, as described in, e.g., Heckman and Smith (1995), experiments may suffer from randomization bias. This occurs when individuals behave differently due to the presence of randomization. For example, if the units under study can undertake activities that complement the treatment prior to receiving it, they have less incentive to do so during an experiment, because they may be randomly excluded from the treatment. Note that randomization bias differs from Hawthorne effects. The latter occur when individuals being evaluated change their behavior in response to being observed, whether

in the context of an experimental or a non-experimental evaluation. Little empirical evidence exists on the importance of randomization bias.

Third, depending in part on the placement of random assignment within the process by which units come to receive the treatment, dropout within the treatment group may cause problems for the interpretation of the experimental impact estimates. Dropout here refers to a departure from the treatment after random assignment, perhaps because it appears less attractive once fully known. Randomly assigning units early in the participation process tends to increase dropout. As detailed in Heckman, Smith and Taber (1998), dropout is a common feature of experimental evaluations of active labor market policies. The usual responses take two forms. In the first, the interpretation of the impact estimate changes and becomes the mean impact of the offer of treatment, rather than of the receipt of treatment. In the second, the impact estimate gets adjusted using the method of Bloom (1984). See Heckman, LaLonde and Smith (1999), Section 5.2, for more details and a discussion of the origins of the adjustment.

Fourth, as discussed in detail in Heckman, Hohmann, Smith and Khoo (2000), in some contexts, control group units may receive a treatment similar to that offered the experimental treatment group from other sources. Their analysis considers the case of employment and training programs, and they show that, at least in the decentralized U.S. institutional environment, where many federal and state agencies offer subsidized training of various sorts, substitution is quite common. In this case, the outcomes of the control group do not represent what the treated units would have experienced had they not received treatment. Instead, they represent some combination of untreated and alternatively treated outcomes. The literature indicates three responses to substitution

bias. As with dropouts, one consists of reinterpretation of the parameter – this time as the mean difference between the treatment being evaluated and what the treated units would have received were that treatment not available, which will sometimes be no treatment and sometimes be some other treatment. The second response consists of adjusting the experimental mean difference estimate by dividing it by the difference in the fraction treated between the treatment and control groups. This represents a generalization of the Bloom (1984) estimator and requires that the substitute treatment have a similar impact to the treatment being evaluated. Finally, the third response consists of using the experimental data to do a non-experimental evaluation. See Heckman, Hohmann, Smith and Khoo (2000) for more details.

As discussed in, e.g., Smith and Sweetman (2002), variants of random assignment can sometimes overcome the political obstacles to experimental evaluation. One such design is the so-called “randomized encouragement” design, which applies to voluntary programs with participation rates less than 100 percent. Here, rather than randomizing treatment, an incentive to participate, such as an additional subsidy or additional information about the program, gets randomly assigned to eligible units. In simple terms, this strategy creates a good instrument (see Section 3.4 for more about instruments) by creating some random variation in participation. This method depends crucially on having an incentive that actually does measurably affect the probability of participating. The second alternative design consists of random assignment at the margin, as in Black, Smith, Berger and Noel (2003). They examine the effect of mandatory reemployment services on unemployment insurance recipients in the state of Kentucky. Individuals get assigned to the mandatory services based on the predicted duration of their

unemployment spell. Only individuals at the margin of getting treated get randomly assigned. This proved much less intrusive than full-scale random assignment and also satisfied the state's concerns about treating all claimants with long expected durations. In a heterogeneous treatment effects world, neither of these alternative versions of random assignment estimates the mean impact of treatment on the treated. However, the parameters they do estimate may have great policy interest, if policy concern centers on marginal expansions or contractions of the program.

In sum, experiments have enormous power, both because of their statistical properties and, not unrelated, because of their rhetorical properties. Policymakers, pundits and plebes can all understand experimental designs, something that is not so true of non-experimental methods such as matching, instrumental variables or structural general equilibrium models. In contrast to the present situation, where evaluators must constantly cajole and prod resistant agencies to undertake random assignment evaluations, in a well-ordered polity, government officials would bear the burden of making a case for not doing random assignment in the case of expensive or important programs that justify a full-scale evaluation and that do not fall into the inappropriate categories described earlier in this section.

### *3.2 Selection on Observables: Regression and Matching*

Selection on observables occurs when observed characteristics determine participation in a program but, *conditional on those characteristics*, participation does not depend on outcomes in the absence of participation. In such situations, conditioning on the characteristics that determine participation suffices to solve the selection bias problem.

In general, selection on observables has the greatest plausibility when the observed data contain variables that relate to all of the major factors identified by theory (and evidence on similar programs) as affecting both participation and outcomes. A simple example makes the point clear. Suppose that both men and women choose at random to participate in some training program, but that men choose to participate with a higher probability than women. Assume as well that women have better labor market outcomes without the training than do men (not an unrealistic assumption in the populations targeted by many training programs) and that the training has the same average effect on men and women. Simply comparing the outcomes of all participants to all eligible non-participants will understate the impact of the program, because this comparison conflates the impact of training with the effects of the over-representation of men in the program. Because men have worse labor market outcomes in the absence of training than women, the over-representation of men in the program will make this simple comparison a downward biased estimate of the impact of the program. If, instead, we separately compare male participants and non-participants and female participants and non-participants, we will obtain an unbiased estimate of program impact.

By far the most common way of taking account of selection into treatment on observable characteristics consists of using standard linear regression methods, or their analogs such as logit and probit models for limited dependent variables, and including the observables in the model. A standard formulation would look like

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i,$$

where  $Y_i$  is the outcome of interest,  $D_i$  is a dummy variable for receiving treatment (with  $\beta_D$  the corresponding treatment effect),  $X_{1i}, \dots, X_{ki}$  are the confounding variables and

where the regression would be estimated on a sample of treated and eligible non-treated units (which means you cannot use this approach for a treatment that reaches all eligible units). In a common effect world, provided the selection on observables assumptions holds,  $\beta_D$  estimates the common treatment effect. In a heterogeneous treatment effects world, it estimates the impact of treatment on the treated under fairly general assumptions.

Regression has the great advantages of familiarity and ease of interpretation and use. All standard statistical packages include it, and even some database programs. The coefficients have interpretations as partial derivatives or finite differences (though this becomes a bit more complicated in logit and probit models, some statistical packages now report marginal effects, which are close cousins to partial derivatives<sup>2</sup>). Despite these advantages, it is important to note that regression is not, in general, an “expedient and satisfactory net impact technique,” as claimed in the Eberts and O’Leary chapter in this volume. Whether or not regression produces consistent estimates depends on whether the selection on observables assumption holds, which in turn depends on the richness of the set of the variables available for inclusion in the regression and on the nature of the selection process in each context. As I emphasize in Section 3.7 below, no econometric evaluation estimator provides a general solution to the selection bias problem. In some cases regression will do so and in others it will not. One of the primary contributions of the evaluator consists in determining which case corresponds to the evaluation at hand, a task that requires more than soothing phrases.

---

<sup>2</sup> For example, Stata reports marginal effects equal to derivatives evaluated at the mean of the covariates, as opposed to the slightly more difficult, but technically preferable, procedure of calculating mean derivatives by taking the average of derivatives evaluated at the covariates values for each observation.

In addition to assuming selection on observables, standard linear regression analysis also imposes a linear functional form on the data, which may or may not correspond to the underlying population relationship. Including higher order terms in  $X_i$  relaxes this constraint, but this is rarely done in practice. Matching methods, which have received a lot of attention in the evaluation literature in the past few years, relax the linear functional form assumption inherent in the standard regression approach while maintaining the assumption of selection on observables.

The basic idea of matching is to directly compare individuals with exactly the same (or similar) values for the relevant confounding variables. This avoids any functional form restrictions. The easiest way to see this is to consider an example with only discrete  $X$  variables. In that case, the simplest version of the matching estimator consists of finding, for each treated unit, an untreated unit with identical values of the covariates. The impact estimate, which provides an estimate of the impact of treatment on the treated in a heterogeneous effects world, then consists of the mean outcome for the treated units minus the mean outcome for the matched untreated units. An estimate of the average treatment effect can be obtained by re-weighting the  $X$ -specific estimates by the distribution of the  $X$  in the population (rather than implicitly weighting them by the distribution of  $X$  in the treatment group, as matching normally does).

When the set of matching variables includes continuous variables, or a large number of discrete variables, exact matches become difficult.<sup>3</sup> In such cases, matching relies on a distance measure to determine which untreated observations should play a role in estimating the counterfactual for each treated observation. Put differently, the distance

---

<sup>3</sup> In the technical literature, this is called the “curse of dimensionality.” See Smith and Todd (2004) for further discussion.

metric converts distances on a vector of variables into a single number, which can then be used to match similar observations. With a single continuous variable, absolute differences in that variable can serve as the metric. With a richer covariate set, the choices include Mahalanobis metric matching, as in Rosenbaum and Rubin (1985), and propensity score matching, as developed by Rosenbaum and Rubin (1983). The propensity score is just statistical jargon for the probability of participation conditional on  $X$ .

The economics literature has tended to focus on propensity score matching because theory often provides guidance on which variables should affect the probability of treatment and because the probability of participation is often of independent interest and so would get calculated anyway. With propensity score matching, the estimated counterfactual for a given treated observation is based on the outcomes of untreated observations with similar probabilities of participation. In order to satisfy the assumption of selection on observables, the propensity score model should include all of the important variables that affect both participation and outcomes – not just all of the ones in the data set at hand.

Matching methods have the additional benefit, relative to standard regression methods, of focusing attention on the so-called “support” problem. The support problem arises when the data contain no similar untreated observations for some of the treated observations. In such situations, it is common when applying matching methods to simply drop such treated observations from the analysis, which can substantially affect the nature of the parameter being estimated if there are a lot of them. In contrast, standard regression methods will produce estimates even in the absence of comparison

units that look like the treatment units because the linear functional form fills in for the missing data. Put differently, the regression identifies the untreated outcome model in the region of the data where the untreated observations lie, and then projects it out into the region of the data where the treated units lie, thereby implicitly estimating the counterfactual.

Figure 1 helps illustrate this point. The horizontal axis represents the matching variable under the assumption that conditioning on this single  $X$  takes account of non-random selection into treatment. The vertical axis represents the outcomes. The two clouds of data represent treated and untreated units, respectively. Region B, in the center of the figure, constitutes the region of “common support”, where there are both treated and untreated observations with roughly similar values of  $X$  (close enough to be “good” matches). The support condition fails in Region C, on the right side of the figure, which includes treated observations but no untreated observations with similar values of  $X$ . No matching estimate can be constructed for the treated observations in Region C. In sharp contrast, there would be no problem including them in a regression analysis, whose linear functional form would project the conditional mean outcome without treatment estimated using the data on untreated units in Regions A and B out into Region C. In the end, projections of this sort may work out in a given context, but the evaluator would still like to know whether his or her estimates rely heavily on the linear functional form or not. Finally, Region A contains untreated units but no treated units. When estimating the impact of treatment on the treated, this poses no problem; matching ignores these units, as they are not required to construct the counterfactual for any of the treated units. Standard linear regression methods, on the other hand, make use of these observations to

help pin down the relationship between  $X$  and the outcome. See Heckman and Vytlačil (2001a) and Black and Smith (2004) for extended discussions of the support problem.

Whether matching on  $P(X)$  or on some other distance measure based on  $X$ , a variety of different methods exist for actually doing the matching when the data lack exact matches. These methods differ on several dimensions, of which the two most important are probably whether or not they use more than one untreated observation to construct the counterfactual for each treated observation, and whether or not they use each untreated observation to help construct the estimated counterfactual for more than one treated observation. The simplest form of matching – common in the applied statistics literature but not in the economics literature – is single nearest neighbor matching without replacement. In this case, a single untreated observation estimates the counterfactual for each treated observation, and untreated observations can only estimate the counterfactual for one treated observation, even if they are the closest untreated observation to many treated observations. This method has the disadvantages that it tends to throw out a lot of comparison observations and that, in data sets with only a few comparison observations, it tends to lead to bad matches, in the sense of pairing treated observations with untreated observations that do not look much like them. Dehejia and Wahba (1999) provide a useful discussion of this issue, with an empirical example that illustrates the dangers of matching without replacement.

Other versions of matching include kernel matching, nearest neighbor matching with more than one nearest neighbor (this increases the bias but lowers the variance), local linear matching and weighting by the inverse of the propensity score. See Smith and Todd (2004) for a relatively applied overview of the different methods, and, among

others, Heckman, Ichimura and Todd (1997,1998), Heckman, Ichimura, Smith and Todd (1998), and Hahn (1998) for the technical details. Imbens (2004) provides a readable survey of the recent technical literature while Heckman and Navarro-Lazano (2004) contrast propensity score matching methods with methods based on exclusion restrictions and explicate the differing role the estimated probability of treatment plays in each one. Frölich (2004) presents a Monte Carlo analysis that suggests that kernel matching (and another method called ridge matching) tend to outperform nearest neighbor matching, local linear matching and weighting by the inverse of the estimated propensity score. Lechner (1999,2000) provides some fine examples of matching in practice.

A number of important extensions to matching exist in the literature. For example, longitudinal data allow the combination of matching with the difference-in-differences methods discussed in Section 3.4. Heckman, Ichimura, Smith and Todd (1998) develop these methods. See Blundell and Costa Dias (2002) for a less technical introduction (that also includes the case of repeated cross-section data) and Eichler and Lechner (2002) for an application. In studies of programs that may create effects at the level of a town or small city, matched local area designs are sometimes undertaken. These designs typically share the problem of having too few treated and untreated areas for reliable statistical inference. See Long and Wissoker (1995) for an example.

In sum, matching represents an important extension of traditional linear regression approaches when the data support an assumption of selection on observables. Matching does not solve the problem of selection on unobservables, but does allow flexibility in conditioning and makes it easy to examine the support issue. Matching, whether semi-parametric (propensity score matching) or non-parametric (cell matching),

requires more data than traditional approaches that impose more structure on the problem. While the standard statistical packages do not yet contain routines to perform matching, several user-written routines exist for use with Stata.

### *3.3 Selection on Unobservables: Longitudinal Methods*

One very simple method for evaluating a development policy consists of examining the difference in the outcomes of the units affected by the policy before and after the policy comes into force. The implicit assumption underlying this simple strategy is that units subject to the policy change would have had the same outcomes as before, had the policy not intervened to change them. Though reasonable in some contexts, this assumption requires that treated units not select into treatment based on temporary changes in their outcomes. For example, if firms only choose to participate in a subsidy program when they are having a bad year, and if most bad years are followed by good years even without the subsidy, then a before-after comparison of the outcomes of participating firms will overstate the impact of the subsidy on firm performance by attributing to it the subsidy firms' normal return to good times. The before-after estimator also requires the absence of aggregate changes in outcomes due to the macroeconomy or other factors. If the economy heats up and all units do better, then the estimator will incorrectly assign the gains to the treatment.

In terms of the notation, let  $t$  denote a post-program period and  $t'$  a pre-program period. The before-after estimator is given by

$$\Delta_{BA} = E(Y_{1t} | D = 1) - E(Y_{0t'} | D = 1).$$

The estimator is consistent only if  $E(Y_{1t} | D = 1) = E(Y_{0t} | D = 1)$ , which is just the stability condition in the previous paragraph expressed in notation. Note that non-random selection of units into treatment does not pose any problem so long as the stability condition is satisfied, because the before-after estimator compares treated units to their earlier selves. The before-after estimator estimates the mean impact of treatment on the treated, and does not capture any general equilibrium effects unless they are included in the outcomes used to calculate the estimates (as with, e.g., city-level outcome variables). Interrupted time series designs, called event history analyses in the empirical finance literature, generalize the before-after estimator by including additional periods of data before and/or after the treatment of interest. This allows the researcher to control more extensively for pre-existing trends in outcomes.

Concerns about confusing treatment effects with general changes in the economy motivate the so-called “difference-in-differences” estimator. This estimator compares the before-after change of treated units with the before-after change of untreated units. In so doing, any common trends, which will show up in the outcomes of the untreated units as well as the treated units, get differenced out. In terms of our notation, the difference-in-differences estimator consists of

$$\Delta_{DD} = [E(Y_{1t} | D = 1) - E(Y_{0t} | D = 1)] - [E(Y_{0t} | D = 0) - E(Y_{0t} | D = 0)].$$

The common time trend assumption that justifies the estimator is given by:

$$E(Y_{0t} | D = 1) - E(Y_{0t} | D = 1) = E(Y_{0t} | D = 0) - E(Y_{0t} | D = 0).$$

Researchers most commonly estimate the difference-in-differences model using a relatively simple regression model, as in

$$Y_i = \beta_0 + \beta_T T_i + \beta_D D_i + \beta_{DD} T_i D_i + \varepsilon_i,$$

where  $T_i = 1$  in period “ $t$ ” and  $T_i = 0$  otherwise and where we omit the  $X$  variables (from the notation – not from the model) for simplicity. The coefficient  $\beta_T$  measures the effect of the common time trend, while  $\beta_D$  estimates the time invariant difference in untreated outcomes between the treated and untreated units, and  $\beta_{DD}$  provides the difference-in-differences impact estimate. In a common effect world (which is a world with homogeneous treatment effects),  $\beta_{DD}$  estimates the mean impact of treatment on the treated (and all of the other mean impact parameters). In a heterogeneous effects world,  $\beta_{DD}$  estimates the mean impact of treatment on the treated under fairly general assumptions. Like the before-after estimator, the difference-in-differences estimator generalizes to the case of more time periods either before or after the treatment or both. Blundell, et al. (2001) discuss and implement such estimators.

Panel data models constitute the most general version of these estimators. These models apply to data sets with multiple observations over time on many treated units (and perhaps some untreated units). A regression is run of the outcome variable of interest on exogenous covariates plus dummy variables for each unit and each time period. The unit dummy variables control for permanent differences in outcomes among units, just as in the simple difference-in-differences model. The time period dummies control for aggregate effects in each period. Panel models require some variation in the timing of the treatment; without such variation, the treatment effect cannot be distinguished from the aggregate time effects.<sup>4</sup> These models also require that the timing of treatment among

---

<sup>4</sup> The literature that uses panel models to evaluate the impact of the switch from Aid to Families with Dependent Children to Temporary Aid to Needy Families in the U.S. provides a good example of the dangers of using these models with very little variation in the timing of treatment. This literature relies on

units not depend on transitory changes in outcomes. This is not an innocuous assumption. As Heckman and Smith (1999) show in the context of a government job training program, individual participation depends critically on transitory labor market shocks, with the result that longitudinal estimators have a large bias.

In terms of the notation introduced earlier, the basic panel model has the following form:

$$Y_{it} = \beta_0 + \beta_D D_{it} + \mu_i + \mu_t + \varepsilon_{it},$$

where  $\beta_D$  is the panel data impact estimator,  $D_{it}$  is a time-varying indicator for treatment,  $\mu_i$  is a unit-specific intercept,  $\mu_t$  is a time-period-specific intercept and I again omit the  $X$  for simplicity. The time period intercepts soak up any common trends, while the unit-specific intercepts soak up time invariant differences between units. What is left, essentially, is a weighted average of before-after estimates for the different treated units.

A couple of examples illustrate how the models work. Evans and Topoleski (2002) evaluate the impact of Native American casinos on employment and other outcomes (such as crime and bankruptcy in the same county) using panel data methods. They construct panel data on outcomes for Native American tribes in the U.S. that do and do not have casinos. They combine this with data on the timing of casino openings for those tribes that have them. Because of the complicated legal structure surrounding casino gambling in the U.S., and the fact that both the state and federal governments play a role, there is a lot of variation in both the incidence and timing of casinos among tribes. Moreover, the variation in the timing of casino openings among tribes that have them is

---

limited monthly variation in the implementation of TANF across states – variation that seems likely to be related to the outcomes under study and therefore violates the assumptions that justify panel models.

plausibly unrelated to the temporal pattern of outcomes in the absence of the casinos. Put more simply, the timing depends on the vagaries of the political and legal systems, and not on changes in the tribal employment rate. The tribal dummy variables included in the model take account of permanent differences between tribes, and the year dummies take account of nationwide trends affecting all tribes. They find that casinos increase tribal employment to population ratios as well as increasing employment in the surrounding county (as well as crime and bankruptcies).

Coates and Humphreys (1999) provide another economic development application of panel data methods. They compile panel data on cities in the U.S., including the presence or absence in each year of professional football, baseball and basketball franchises and the construction of new stadiums and arenas. Using the same basic framework described above, they examine the effect of pro sports franchises and the construction of new sports facilities – almost always with substantial public money and almost always justified as engines of economic growth – on both the levels and growth rates of economic activity. They find little in the way of economic effects from public spending on pro sports teams, which comports with the remainder of the serious literature on the topic; see, e.g., the papers collected in Noll and Zimbalist (1997).

Overall, panel data methods represent a powerful tool when longitudinal data are available on treated and untreated units, when the timing of treatment varies among units, and when the timing of treatment is unrelated to the outcomes being studied, conditional on the included conditioning variables. It is this latter condition that sometimes gets ignored in the literature. Additional data allow the testing of this assumption in many contexts, see, e.g., Moffitt (1991) and Heckman and Hotz (1989). For readers interested

in the potential pitfalls of panel methods, the highly controversial literature on the impacts of state “right to carry” laws (these allow certain classes of U.S. citizens to carry concealed weapons), which relies almost entirely on such methods, digs deep into what can go right, and what can go wrong, with this approach. See, e.g., Lott and Mustard (1997), Black and Nagin (1998) and Ayres and Donohue (2002).

### *3.4 Selection on Unobservables: Instrumental Variables*

Methods based on instrumental variables (IV) or exclusion restrictions represent an alternative econometric strategy for dealing with selection on unobservables. An instrument, or exclusion restriction, is a variable that affects participation in the treatment but does not affect outcomes other than through its effect on treatment participation (conditional on the other variables included in the outcome equation). The “exclusion restriction” usage arises from the fact that such variables can be excluded from the outcome equation but included in the treatment equation. Unlike longitudinal methods, IV methods require only cross-sectional data, and they can potentially deal with selection on unobservables that vary over time. See Angrist and Krueger (2001) for a longer non-technical discussion of IV methods and see Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999) for more technical treatments.

A simple example with a binary instrument in a common effects world provides the basic idea. Consider two otherwise identical towns, one of which is close to a public training center and the other of which is far away from the same training center. Each town includes 100 eligible persons, of whom 50 have a car. The outcome in the absence of training equals 100 for all eligibles. The benefit of training (net of the opportunity cost

of participant time) equals 10 for everyone (a common effect world), while the cost for those in the near town to get to the training center equals 1. For the eligibles in the far town, the cost of transport to the training center equals 5 for those with a car and 15 for those without a car. The upshot of all this is that all of the eligibles in the near town take the training but that only the eligibles with cars in the far town take training. As a result, we can use location as an instrument, because it affects participation in the treatment – a different fraction of the eligibles participates in each town – but does not affect outcomes other than through its effect on treatment – everyone gets 100 in the absence of treatment in both towns.

The Wald estimator for binary instruments, given by

$$\Delta_{IV} = \frac{E(Y | Z = 1) - E(Y | Z = 0)}{\Pr(D = 1 | Z = 1) - \Pr(D = 1 | Z = 0)},$$

suffices for our simple example (though two-stages least squares could also be used and yields an equivalent estimate). In the Wald estimator,  $Z$  denotes the instrument and  $D$  denotes participation in treatment as before; in our simple example,  $Z = 1$  for eligibles in the near town and  $Z = 0$  for individuals in the far town. Plugging in the numbers from the example yields

$$\Delta_{IV} = \frac{[100 + (1.0)(10)] - [100 + (0.5)(10)]}{1.0 - 0.5} = \frac{5}{0.5} = 10,$$

which is the correct answer. This example illustrates the key point that by inducing variation in treatment receipt unrelated to outcomes in the absence of treatment, the instrument identifies the treatment effect through comparisons of the outcomes of groups with different values of the instrument.

When the instrument is continuous or multi-valued rather than binary, or when multiple instruments are available (which represents good fortune indeed), standard two-stage least squares methods replace the Wald estimator. In the two-stage estimator, the “endogenous” variable (participation in treatment) is first regressed on the instrument or instruments and any exogenous  $X$  variables in the model. In the second stage, the outcome is regressed on the predicted values from the first stage as well as the  $X$ . Intuitively, using the predictions from the first stage, rather than the participation dummy itself, omits variation in participation not resulting from factors unrelated to outcomes in the absence of treatment. A small literature in econometrics explores less restrictive semi-parametric evaluation estimators based on exclusion restrictions, although these estimators have seen very little use in applied work. See Newey, Power and Walker (1990) and Blundell and Powell (2001) for further discussion.

Instruments are a wonderful tool when available, but where do they come from? Deliberate creation represents the most direct way of obtaining instruments. One way to think about social experiments is that they are devices to create good instruments; this includes the randomized encouragement design described in Section 3.1. See Heckman (1996) for more on this view of experiments. A second form of deliberate instrument creation consists of theory combined with clever data collection. For example, Card (1995) adds data on the distance to the nearest college or university to a standard individual-level data set and then uses distance as an instrument for years of schooling. The final, and in practice the most important, form of data collection combines theory with institutional knowledge. Institutional changes that seem unrelated to the outcomes of interest (or at least whose timing is not related to them, as with, e.g., some court

decisions) but affect participation in a treatment can provide good instruments. The bus strike used to provide variation in pre-natal care in Evans and Lien (2002) provides such an example. Differences in program management choices or in program intensity across jurisdictions constitute a potential source of instruments in many economic development contexts. In general, evaluators should think of finding instruments as a side-benefit of thinking about the economics of a given evaluation problem and of learning about the relevant institutions, both necessary activities in their own right.

The bivariate normal selection model of Heckman (1979) is closely related to the instrumental variables model. The bivariate normal model, as its name suggests, assumes that the error terms in the outcome and participation equation have a bivariate normal distribution. Under this assumption, Heckman (1979) provides a two-stage estimator that estimates the impact of treatment on the treated in a common effect world.<sup>5</sup> Formally, the bivariate normal model does not require an exclusion restriction – the functional form assumptions suffice to identify the parameter of interest. However, it is well known in the literature (see, e.g., the survey by Puhani, 2000) that the model tends to instability without an exclusion restriction. As such, evaluators should avoid this estimator unless they have a good instrument at hand. More generally, as Heckman and Robb (1985) note, this estimator makes stronger assumptions than the IV estimator; if the common effects assumption is plausible in a given context, the IV estimator is preferred.

Things become more complicated in a world with heterogeneous treatment effects, particularly when those treatment effects are correlated with the instrument. To see this, return to the simple example of the two towns sharing one training center

---

<sup>5</sup> The model can also be estimated in one stage using full information maximum likelihood methods. While there is an efficiency gain from doing so, the two-stage version may be more robust to misspecification, as it relies less strongly on the joint normality assumption.

considered above. Suppose that half the persons eligible for the program in each town have a benefit of 10 from the program, while half have a benefit of five. Suppose too that transport costs are now homogenous, so that everyone in the near town has a cost of zero to getting to the training center, while everyone in the far town has a cost of seven. In this version of the story, everyone in the near town again participates in training, along with half of the eligibles in the far town. Now, however, instead of the eligibles in the far town with low transport costs participating, it is the eligibles in the far town with the impacts of 10 who participate. For them, the impact of 10 exceeds the transport cost of seven, while for the eligibles with an impact of five in the far town the transport cost of seven makes it not worth their while to participate.

How do these changes in the story of the two towns affect the estimate produced by the Wald estimator? The mean outcome in the near town is now  $107.5 = (100 + (0.5 * 10) + (0.5 * 5))$ . The mean outcome in the far town equals  $105 = 100 + (0.5 * 10)$ . The probability of participation in the near town equals 1.0 and that in the far town equals 0.5. Thus, the Wald estimate equals  $(107.5 - 105) / (0.5) = 2.5 / 0.5 = 5.0$ . This estimate might seem surprising, given that most of those who participate (two thirds to be exact) receive an impact of 10.

The key feature of this version of the story lies in the correlation between the impacts (conditional on participation) and the instrument. In the near town, the mean impact among participants is 7.5, while in the far town it is 10. When a binary instrument such as this one is correlated with the heterogeneous treatment effects conditional on participation, the Wald estimator no longer estimates the mean impact of treatment on the treated. Instead, it estimates what Imbens and Angrist (1994) call a Local Average

Treatment Effect (LATE). In simple terms, the Wald estimator now estimates the mean impact on those units who change their participation status in response to the change in the value of the instrument. The units who change their participation status when the instrument changes in the story of the two towns are those with an impact of five; they participate in the near town but not in the far town. Thus, the LATE in this case equals five.

Note that the other standard treatment parameters do not equal five in this case. For example, the mean impact of treatment on the treated equals  $(2/3) * 10 + (1/3) * 5 = 25/3 = 8.33$ . The average treatment effect equals 7.5, because by assumption half of the eligibles in each town have an impact of 10 and half have an impact of five. The relationship between the treatment on the treated and the LATE is instructive here. The treatment on the treated parameter exceeds the LATE because the inframarginal participants – those who participate regardless of the value of the impact (in this case, those with an impact of 10) have higher average impacts than the marginal participants, as economic theory would predict. The three parameters differ in this case and they answer different policy questions of interest. Heckman (1997), Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999) discuss the binary instrument case in greater detail.

The paper by Sweetman, Warburton, McPhee and Warburton (2003) provides a nice applied example of LATE estimation. In their study, the instrument consists of the person who makes the final decision on disability benefit cases in the Canadian province of British Columbia. When the person making the decision changes, and the acceptance probability increases, the authors can estimate a very interesting LATE – namely the

impact of receiving disability payments on the marginal candidates whose applications would get rejected under one regime but get accepted under the other. They cannot, of course, estimate the impact of receiving disability insurance payments on the inframarginal applicants whose cases would be approved under both regimes.

Things get a bit more complicated with continuous instruments, or with multiple instruments, or with continuous (or multi-valued) treatments in a heterogeneous treatment effects world. See Angrist and Imbens (1995) and Heckman and Vytlačil (1998) for discussions of multi-valued treatments. The bivariate normal model also generalizes to the heterogeneous treatment case, and has received a lot of attention in the literature. In this case, the additional structure provided by the strong distributional assumptions on the error terms allows the estimation of numerous LATEs, as well as the ATE and the treatment on the treated parameter. See Heckman, Tobias and Vytlačil (2003) for further details and an empirical example. Recently, a semiparametric version of the bivariate normal selection model has appeared in the literature. A full description of this model is beyond the scope of this paper; see Heckman and Vytlačil (2001b) for details.

In sum, instruments represent a powerful tool for deriving compelling estimates of the impacts of local economic development programs when they are available. To date, IV methods have seen little use in the economic development literature; as such, clever researchers who seek out novel sources of exogenous variation likely have some low-hanging empirical fruit to harvest that would add to both to our knowledge of the impacts of economic development policies and to our knowledge of the performance of these estimators in practice.

### 3.5 Discontinuity Designs

Discontinuity designs apply to treatments allocated using a particular variable or set of variables, with the treated and untreated units distinguished by a sharp break in the value of the variables. Thus, we might imagine a program that provides remedial education to pre-school students based on their score on a standardized test, with those below a certain score receiving the treatment and those above it not. Or, grants to small towns might be made available to only those towns with populations below 15,000. In general, suppose that treatment is provided to units with a value of some variable  $Z$  greater than a cutoff  $C$ .

Absent additional assumptions, discontinuity designs estimate the mean effect of treatment on units located at the cutoff  $C$  by comparing the outcomes of units just above the cutoff to the outcomes of units just below the cutoff. In terms of our notation, the discontinuity design estimates

$$\Delta_{DIS} = E(Y_1 - Y_0 | X \approx C).$$

For example, in the case of the grants to small towns, the discontinuity estimator estimates the effect of the grants on towns with populations of approximately 15,000 by comparing the outcomes of towns with populations just under 15,000 with the outcomes of towns with populations just over 15,000. The exact form of the comparison depends on the particular discontinuity estimator selected.

Of course, in a common effect world, the impact estimated in the discontinuity design generalizes to other units. Adding the assumption that the untreated outcome is linear in covariates leads to the so-called “regression discontinuity” design, which can be estimated using standard regression methods. Depending on the application, discontinuity designs may capture general equilibrium effects. Whether they do or not

depends on the extent to which outcomes in the treated units influence outcomes in the untreated units. For example, in the case of the grants to small towns, the estimator will pick up any general equilibrium effects captured by town-level outcome variables, but will be biased by migration from large to small towns in response to the grants.

The key to the discontinuity design is that units (or others) must not be able to manipulate  $Z$  so as to cause certain units to be treated and others not. Some treatments will meet this requirement and others may not. The example given above of a treatment rule based on a standardized test with no subjective component is an example of the former. A policy that provides benefits to firms with fewer than five employees provides an example of the latter. A sufficiently attractive benefit will cause some firms with six or more employees to cut their payroll down to five. These firms will almost certainly be a non-random sample of firms with six or more employees. As a result, comparisons at the margin between firms with five and six employees no longer estimate the parameter of interest.

The literature provides some useful theoretical discussions. See, e.g., the related section in Heckman, LaLonde and Smith (1999) as well as, at a more technical level, Hahn, Todd and van der Klaauw (2001). The latter paper highlights the tradeoff between bias and variance associated with deciding how much weight to assign to observations at some distance from the cutoff point. Applications include van der Klaauw (2002), who looks at the effect of financial aid on university admissions, Black (1999), who looks at housing values along the borders of school attendance areas (within school districts) and Pence (2003), who looks at the effects of state mortgage regulations along state borders.

### *3.6 General Equilibrium Methods*

Four types of econometric evaluation methods seek to directly estimate general equilibrium effects. First, as noted in the preceding sections, standard partial equilibrium evaluation methods can capture general equilibrium effects in cases where the unit of analysis incorporates these effects. Thus, for example, if a treatment is randomly assigned to some towns and not to others, and the general equilibrium effects occur within towns rather than between them, then comparing town-level outcomes between towns randomized into the experiment and towns randomized out of the experiment will capture the general equilibrium effects.

The second method consists of traditional (some might say “old style”) multiple equation models that link various aspects of local economic development together. These models are similar in spirit to the multiple equation macroeconomic models used by banks and others to generate short-term economic forecasts, but with more specific case study assumptions built in regarding the particular local economic development context. These models have fallen out of favor in the academic literature because they violate many of the rules of sound econometric practice, as the equations often consist of one endogenous variable regressed on several others, with no instruments in sight. The theoretical basis underlying the multiple-equation system typically lacks much in the way of formal structure. The presence of many equations often makes it difficult to see where estimated effects come from and, in practice, the models often require some subjective input to produce reasonable numbers. In short, these models are hard to like, but the demand for numbers, combined with the lack of simple alternatives keep them in play for practical applications.

The third method is what I call the “magic multiplier” method. This method plays a leading role in evaluations of transit projects and sports infrastructure investments, where the consulting firm performing the evaluation usually has a clear idea of the desired result in advance. In this method, some measure of direct impacts is constructed in a more or less reasonable way. The direct impacts then get multiplied by the magic multiplier, which is supposed to capture all the spillover effects. The particular values chosen for the magic multiplier in a given application typically have little formal justification, which indeed is one of their attractions to those performing the evaluation and their clients. See the papers in Noll and Zimbalist (1997), as well as Crompton (1995), for more details.

The fourth approach to directly estimating general equilibrium effects consists of specifying a structural general equilibrium model of the relevant economy that includes the program, calibrating or estimating the parameters of the model, and then simulating the model with and without the program.<sup>6</sup> Structural general equilibrium models make very strong assumptions about how the different elements of the economy affect one another. Given the complexity of these models, the analyst has no choice but to treat very simply all but those aspects of the economy most relevant to the issue at hand. While the strong assumptions and structure represent a disadvantage in one sense, they represent the strength of these models as well. All the assumptions are clear and on the table, and whatever effects emerge from the model can be traced to particular aspects of the model economy and thereby back to the underlying theory. Because of their complexity and because of the expense associated with these models, undertaking the

---

<sup>6</sup> The debate regarding the relative merits of calibration and estimation of structural equilibrium models lies well beyond the scope of this paper. See Hansen and Heckman (1996), Kyland and Prescott (1996) and Sims (1996) for three relatively non-technical presentations of different views of the debate.

development of such a model makes sense only for large programs (either in terms of expenditure or in terms of the fraction of the relevant units directly treated), where the answer matters a lot and where we expect important general equilibrium effects.

The literature contains a handful of such evaluations; I highlight four notable examples here, all drawn from the field of active labor market policy. Davidson and Woodbury (1993) construct a general equilibrium search model that allows them to evaluate the general equilibrium effects of the U.S. Unemployment Insurance (UI) bonus programs.<sup>7</sup> These programs, whose partial equilibrium impacts were estimated by a series of social experiments, paid UI claimants a cash bonus if they found work within the first 11 weeks of their benefit claim. Davidson and Woodbury's (1993) model indicated that from 30 to 60 percent of the partial equilibrium impact gets cancelled out in general equilibrium due to displacement and changes in the optimal search effort of unemployed workers not eligible for the bonus. Lise, Seitz and Smith (2003) modify Davidson and Woodbury's (1993) model to evaluate the Canadian Self-Sufficiency Program, a wage subsidy to long-term income assistance recipients who find full-time work. They find that general equilibrium effects, including changes in optimal search effort and in the distribution of wages, change the social cost-benefit performance of the program from positive to negative. Blundell, Costa-Dias and Meghir (2003) use general equilibrium methods to examine a wage subsidy program in the UK and find that taking account of general equilibrium effects makes a big difference to their findings. Finally, Heckman, Lochner and Taber (1998, 1999) consider the general equilibrium effects of a \$500 per year tuition subsidy to university attendance in the U.S. They find that taking account of changes in equilibrium skill prices means that the estimated general

---

<sup>7</sup> See Meyer (1995) for an overview of UI-related policy experiments.

equilibrium impacts are smaller than the partial equilibrium impacts by a factor of ten. In each of these cases, taking general equilibrium effects into account plays an important role in getting the correct answer about the effects of a policy change.

### *3.7 Choosing Among Alternative Non-Experimental Evaluation Methods*

When feasible, an experimental evaluation will provide the most compelling evidence on the effectiveness of local economic development programs. When an experiment is not feasible, the evaluator must choose among the alternative non-experimental evaluation methods summarized in Sections 3.2 to 3.6.

The lucky analyst enters the process at the beginning, and can influence the program design and implementation as well as the data collection with a specific estimation strategy (or strategies) in mind. In this happy situation, the analyst can build in two or three alternative evaluation strategies, thereby providing multiple lines of evidence and allowing for the (not unlikely) possibility that one of them will not work out in practice. The unlucky analyst enters at the end of the process, and must try to choose an evaluation method that fits data and institutions chosen by others.

The literature provides a lot of guidance to both the lucky and the unlucky analyst, guidance that frequently gets ignored in evaluation practice. Most of this guidance comes from a growing list of papers that use experimental data sets to benchmark the performance (usually in terms of bias) of alternative non-experimental estimators in different contexts defined by the available data and the institutional setup of the program, where the latter in turn determines the nature of the process by which some units come to receive treatment. This literature includes LaLonde (1986), Fraker and

Maynard (1987), Heckman and Hotz (1989), Bell, Orr, Blomquist and Cain (1995), Friedlander and Robins (1995), Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1998), Heckman and Smith (1999), Dehejia and Wahba (1999,2002), Agodini and Dynarksi (2001), Glazerman, Levy and Myers (2003), Michalopoulos, Bloom and Hill (2004) and Smith and Todd (2004). The literature also includes a simulation study, Section 8.3 of Heckman, LaLonde and Smith (1999) that examines the performance of alternative non-experimental estimators for various data generating processes.

The early parts of this literature framed the question of interest in terms of finding the one true estimator – the magic bullet that would slay the beast of selection bias in every context. More recently, the literature has realized that this is not the right question, because there is no magic bullet. As described above, different non-experimental evaluation strategies make different assumptions about the nature of the selection process and about the available data. When those assumptions hold, a given estimator will provide consistent estimates of certain parameters of interest. When they do not, it will not. Thus, rather than looking for one single estimator that works universally, the literature now seeks the mapping, or relationship, between the institutions and data available in a given context (and the parameter of interest) and the choice of a non-experimental evaluation strategy. Sometimes, as in Hui and Smith (2003), the data available in a given context do not support *any* estimator.

The literature that makes use of experimental benchmarks teaches a number of somewhat obvious (though not so obvious that they have not been ignored in many published papers and even more unpublished ones) but important lessons. A few

examples serve to demonstrate the value added by this line of research. The evidence from the series of papers using the National Supported Work Demonstration experimental data show that the handful of variables (age, race, education and lagged annual earnings) available in the U.S. Current Population Survey (CPS) do not suffice to control for selection into a program that served a highly disadvantaged population including ex-convicts and ex-addicts. The “selection on observables” strategies described in Section 3.2 require rich data on observable determinants of participation and outcomes. Heckman, Ichimura, Smith and Todd (1998) demonstrate the importance of drawing comparison group members from the same local labor markets as participants when evaluating active labor market programs. They also show that using outcomes measured in different ways for treated and untreated units (such as administrative data for one and survey data for the other) can lead to outcome differences that look like selection bias but really constitute systematic measurement differences. Heckman and Smith (1999) show that when individuals select into a treatment based on transitory labor market shocks, as they do in most active labor market policies, longitudinal estimator strategies such as those described in Section 3.3, which assume selection based on permanent differences, fare quite poorly. Heckman and Hotz (1989) demonstrate the value of statistical specification tests in contexts with rich enough data to allow their use.

While the specific lessons from the literature derive mainly from active labor market policies, the general lessons hold when choosing a non-experimental estimator for all sorts of economic development programs. A sound evaluation will pay close attention to the nature of the institutions that determine selection into treatment. These institutions determine the nature of any selection bias, and thereby the plausibility of particular non-

experimental evaluation strategies. A sound evaluation will also pay close attention to the fit between the available data and the particular evaluation method employed. Matching methods make no sense without rich data. Instrumental variable methods make no sense without a good instrument. Longitudinal methods make no sense with cross-sectional data or when selection into treatment depends on transitory, rather than permanent, characteristics. In short, a sound evaluation builds on economic theory, econometric theory and existing evidence in choosing a non-experimental evaluation strategy that matches the data and institutions present in a given context.

#### **4.0 Alternatives to Econometric Evaluation**

##### *4.1 Participant Self-Evaluation*

Evaluators could save a lot of time, money and econometric effort if program participants could reliably evaluate a program directly. In the case of a program that treats individuals, this consists of asking participants, following their participation, whether or not the program made them better off and, if so, how much. A simple survey question replaces all the econometric issues discussed in Section 3. Indeed, many evaluations of U.S. employment and training programs include such questions, and the performance standards for the U.S. Workforce Investment Act include customer satisfaction measures; see U.S. Department of Labor (2000). A similar procedure could apply to firms as well, with the relevant officer chosen to answer the question as in Bartik (2004). Survey methods could even be used to get at some sorts of general equilibrium effects, as is done in the literature on the valuation of environmental amenities; see, e.g., Portney (1994).

For example, local residents could be asked how much they value their new small business incubator, even if they do not make use of it themselves.

While participant self-evaluation sounds good in theory, surprisingly little direct evidence exists regarding its ability to get the correct answer. In order for participant self-evaluation to yield valid impact estimates, respondents have to correctly estimate the unobserved counterfactual of what would have happened to them had they not participated, and then compare it to their realized experience as participants.

Heckman and Smith (1998) present some direct evidence on the validity of participant self-evaluation that I reproduce here in Table 1. Experimental treatment group members in the U.S. National Job Training Partnership Act Study were asked in a follow-up survey 18 months after random assignment if they thought the program benefited them (see table notes for exact question wording). Table 1 shows the fraction of each of the four demographic groups in the experiment – adult males, adult females, male youth and female youth – who answered yes. It compares these to the experimental impact estimates on self-reported earnings over the 18-month period between random assignment and the follow-up interview. Table 1 reveals little correlation between the fractions self-reporting that they benefited from the program and the experimental earnings impact estimates, which suggests that program participants do not do a good job of constructing the counterfactual necessary to determine whether or not the program made them better off. See Heckman and Smith (1998) or Smith and Whalley (2004) for further discussion of the evidence from the JTPA experiment.

The indirect evidence also suggests trouble for participant self-evaluation. First, the discussion in Section 3 makes it clear that researchers have great difficulty

constructing reasonable estimates of average counterfactuals – probably an easier task than estimating the counterfactual for a single person. The literature on behavioral decision theory suggests that individuals have all sorts of cognitive problems with less difficult tasks such as making consistent decisions when choices are presented in different ways, and that most people are poor “intuitive statisticians.” Survey effects may also cause trouble. Respondents may not want to risk offending an interviewer by saying that a program did not help them (or they may not want to admit to themselves that they wasted their time and energy on it!) See, e.g., Bradburn, Sudman and Wansink (2004) in regard to interviewer effects of this sort. Finally, as Bartik notes in his chapter in this volume, for programs involving monetary transfers (or valuable in-kind transfers), respondents can have a direct interest in the program continuing, and so may report a behavioral response even when one does not exist in order to keep the goodies coming.

While the limited available evidence argues against relying on participant self-evaluation for impact analysis, this does not preclude gathering useful information from participants about other aspects of their participation (or, indeed, from gathering useful information from eligible non-participants about why they chose not to take part). For example, participants will likely have a good sense of the quality of service they received and of the amount of red tape involved in participation.

#### *4.2 Alternatives to Econometric Evaluation: Performance Standards*

Administrative performance standards represent another potentially inexpensive alternative to impact analysis. Performance standards typically consist of quantitative measures of program outputs (the number of checks sent out on time) or outcomes (how

many of the trainees found a job within a month after finishing the training program). In terms of our notation, they generally consist of functions of  $Y_1$ . They have grown in popularity as part of the “reinventing government” movement of the 1990s – see, e.g., Osborne and Gaebler (1992) – and now pervade the U.S. government as a result of the Government Performance and Results Act (GPRA) of 1993.

Performance standards have many uses, and in some contexts they tell you all you need to know. For example, the primary mission of the U.S. Social Security Administration (SSA) consists of sending out checks (or making direct deposits) to the correct people at the correct time and in the correct amount. A performance measure that gives the fraction of the time that this happens tells much (if not all) of what needs to be told about how well SSA performs. Their task consists of an outcome, rather than an impact, and so is well suited to management using outcome based performance measures. Of course, the checks SSA sends out will have behavioral impacts of interest to economists and policy-makers – such impacts are not well captured by performance measures based on outcome levels.<sup>8</sup>

In contexts where program impacts represent the main object of concern, reliance on performance standards as a proxy for impact estimates requires evidence of a systematic relationship between the two. To make the problem concrete, consider a job training program that uses “entered employment rates” – such as the fraction of trainees employed 90 days after leaving the program, as its performance measure. This measure is a version of  $Y_1$ . In order for the performance measure to provide a useful proxy for the program’s impact – that is, for the difference it makes in the employment rate relative to

---

<sup>8</sup> I have taken this example from Wilson (2000), a book well worth reading for those interested in why managing and evaluating government programs proves so difficult in practice.

what would have occurred had the participants not participated in the program – the performance measure must be positively correlated with the program impact. In this case, this means that employment after leaving the program must be correlated with the change in employment status induced by the program. In one extreme case, that where no one finds employment without the program, the correlation equals one and the performance measure equals the impact. More generally, there is no particular reason why this condition should hold.

A program that is most effective for those clients least likely to find employment on their own may exhibit a negative relationship between employment rates and impacts. To see this, consider the extreme case of easy-to-serve clients who would always find employment on their own and hard-to-serve clients who never would, but do so half the time when they participate in the program. In one month, the program serves half of each type of client. Its employment rate is 0.75 because the easy to serve all find employment and so do half of the hard to serve, while its impact is 0.25, reflecting the fact that it only benefits hard to serve clients. In another month, it serves only hard-to-serve clients. In that month, its employment rate is 0.50 but its impact is also 0.50, because none of the hard-to-serve clients would have found employment on their own. For this program, the employment rate performance measure is negatively related to program impacts, and so provides a poor proxy for program impacts (and a strong incentive for program managers to cream skim by serving only easy-to-serve clients).

The evidence on this question for U.S. employment and training programs, such as that presented in Heckman, Heinrich and Smith (2001) and Barnow (1999) and summarized in Barnow and Smith (2003) suggests that common performance measures

used in that context, such as entered employment rates, do not correlate very well, if at all, with program impacts.

Very similar issues arise in other program contexts when performance gets judged according to outcomes rather than impacts. For example, government programs that subsidize commercial research and development are often judged based on the fraction of projects that pay off, which provides an incentive for program operators to choose projects that would have been funded anyway without the subsidy, rather than funding projects with (potentially) large spillovers but low private benefits. It is the latter type of project that provides the economic justification (as opposed to the political justification) for the subsidy. See Wallsten (2000a) for a popular discussion of the U.S. Small Business Innovation Research (SBIR) program and Wallsten (2000b) for a more academic discussion. Wallsten's research suggests that each dollar of the funds provided under the SBIR program crowds out a dollar of private research funds. This results from a focus on choosing projects likely to succeed in the market, as this is the metric used to judge program success. His findings indicate that the program has no impact on the total amount of research undertaken. Overall, the literature suggests that performance standards do not represent a general substitute for econometric impact evaluation.

## **5.0 Practice**

This section briefly considers some important practical issues associated with evaluating local economic development policies. The first subsection focuses on when not to do an evaluation. The second subsection focuses on how to choose an evaluator and then how to evaluate an evaluation once completed, and the third highlights some important issues

in cost-benefit analysis. This section builds on the discussion in Smith and Sweetman (2001).

### *5.1 When Not to Do an Evaluation*

Evaluations consume time and resources. As such, evaluations, like the programs being evaluated, should go forward only when their benefits are likely to exceed their costs.

This subsection outlines a number of situations in which an evaluation will likely not pass a standard cost-benefit test and where, as a result, the money that would be spent on evaluation would be better spent on other things.

The first situation where an evaluation is a bad idea is when money is short and other basic administrative functions are not in order. Before doing an evaluation, program operators should have a clear idea of which units participate in their program, whether or not the participating units are eligible for the program and, in a voluntary program, how the participating units compare to the population of all eligible units. They should also know how much money the program is spending, what it is being spent on, and which treated units the money is being spent on. Collecting and examining all this information represents a basic fiduciary duty on the part of the program operator acting as an agent to the long-suffering taxpayer (or to those who donate to a non-profit organization that sponsors economic development projects). These fiduciary duties should come before attempts at evaluation; after all, a program that is not under control in terms of eligibility and costs seems unlikely to produce much in the way of impacts.

The second situation where evaluation can be skipped is when the impact is known in advance. This can happen for two reasons. First, there may already be a

substantial, high quality evaluation literature for a particular type of program. For example, thanks to a long series of experimental evaluations at the state level, the literature has a pretty clear idea of the effects of both mandatory and voluntary job search assistance programs on single mothers on welfare in the United States. See, e.g., Gueron and Pauly (1991) and Bloom and Michalopoulos (2001). Additional evaluations, unless they cover a non-trivially different treatment modality or population, probably do not justify their cost. The second way for the answer to be known in advance is when programs really exist just to transfer money to some favored individuals, firms or other interests, with the economic development justification serving as a useful distraction for a bored public and an inept media. Subsidies to particular large firms seeking to locate a new plant represent an important example of such programs. From a national point of view, such bidding wars between states and localities can do no better than have a zero impact, and will have a negative one to the extent that they cause geographic misallocations of production. Indeed, the European Union forbids competition of this type among its member nations.<sup>9</sup> These programs also undermine the rule of law, an important determinant of long-run growth at the international level, by treating some firms differently just because of their size, mobility, or political connections. A bit of cynicism, combined with simple economic theory and careful attention to where the money goes, usually suffices to identify such circumstances. Sadly, exposing them to the light of day does not always (or even often) result in their disappearance.

The third situation where an evaluation may not represent a good investment arises when the samples available for the evaluation would lack the size required for statistical inference. For example, a program that subsidizes only five firms will not

---

<sup>9</sup> See the discussion and references in Bartik (2003b).

provide a clear statistical picture of program impacts using any of the methods outlined in Sections 3.1 to 3.6. The econometric methods outlined here require a substantial number of treated and untreated units in order to identify impacts statistically distinguishable from zero. Limited sample sizes can also arise from budgetary limitations in situations with a large number of units potentially available for an evaluation, but with substantial costs (e.g., survey costs) of adding each unit to the data. Statisticians have developed formal methods for determining the number of observations required to detect an impact of a given size with a certain degree of confidence under various assumptions about the variance of the outcome measure and other characteristics of the evaluation context. Applying these methods constitutes a “power analysis”; such an analysis should precede the decision about whether to proceed with an evaluation except in cases with very small or very large sample sizes. See, e.g., Maxwell (2000) for a discussion of power analysis and additional citations.

The fourth situation that calls for not doing an evaluation arises when the data do not exist to support an evaluation, or when high quality data cannot be obtained at a cost within the available budget. For example, many relatively expensive evaluations of major government programs often rely on survey data with surprisingly (or even shockingly) low response rates. In such situations, the evaluation must devote additional attention to issues of selective non-response, which cast doubt on the consistency and generality of the findings. In such situations, it might be better to spend enough money to get a reasonable response rate (say, 80 percent) or to use an alternative data source, such as administrative data. The latter are, of course, no panacea. Data quality has a low priority at many agencies, which means an evaluation must devote substantial resources

to ex post cleaning of the data, perhaps discarding some fields entirely. See, e.g., Hotz and Scholz (2002) for further discussion of administrative data. In general, there exists some reservation data quality level below which an evaluation becomes valueless.

A lack of evaluation expertise constitutes the final situation when no evaluation may dominate some evaluation. If the organization potentially undertaking the evaluation lacks access to the funds or the personnel to carry it forward, or to evaluate it when done, then it should generally not attempt the evaluation, particularly if the literature contains relatively strong evaluations of similar programs operated in similar contexts. Weak evaluations do not justify the money they cost, which leads directly to the topic of the next subsection, which concerns how to choose an evaluator and how to evaluate the evaluator's evaluation.

### *5.2 Choosing and Evaluating an Evaluator*

Anyone can declare herself an evaluator and seek contracts for performing evaluations. In practice, many individuals (and collections of individuals in firms or project-specific coalitions) perform evaluations, including economists, statisticians, psychologists and sociologists. Some do evaluation on the side, in addition to teaching and/or academic research; others do evaluation full time. Some know all the latest econometric methods while others can only run regressions. The particular evaluation context and the budget loom large here, so I offer only a few general observations.

First, experiments are harder than you think; if you want to do one and have not done one before, hire a firm that knows how to do it. Such firms include MDRC, Mathematica and Abt Associates, to name a few. Second, different types of evaluators

have different characteristics, which should be matched to the needs at hand.

Professional evaluation firms cost the most, but they have a lot of experience and deliver a very polished product on time and, generally, within budget. Academics, on the other hand, tend to cost a bit less, and sometimes know more econometrics, but have a lower probability of finishing on time and a bit less polish.

Third, some evaluators will take your money and give you back an embarrassing mess. For example, the paper by Gregory (2000), published (for unknown reasons) in the journal *Evaluation*, suggests an evaluation centered on a variant of the “sites of oppression matrix”. In this approach to evaluation, key stakeholders sit around and contemplate all the ways in which life has treated them poorly (or at least those ways somehow related to the program) and the evaluator then writes about the filled-in matrix. Rather obviously, such exercises represent an entertaining diversion for the stakeholders and easy money for the evaluator, but yield no insight about the impact of the program. Fourth, sometimes you can get the econometric part of an evaluation done at low cost if you provide an academic researcher with interesting data that they can use to write articles for publications in scholarly journals. Indeed, many if not most published evaluations of social programs were not paid for by the agencies operating the program they evaluate.

Evaluations, like programs, require evaluation. Some evaluations are very good while others are very weak. Not all agencies that commission evaluations have the internal staff expertise to undertake such evaluations. Even if they do, external quality checks may add substantial value to the evaluation and also increase its credibility. A number of methods exist for incorporating external feedback and review into the

evaluation process. Large-scale evaluations often include a technical review panel of experts who provide feedback at critical stages, such as the design report and the draft impact analysis. In smaller evaluations, a single outside expert may play this role by providing comments on drafts of various reports. Once an evaluation is complete, feedback is still useful to guide readers in determining how much weight to place on the results obtained. Encouraging publication in peer-reviewed journals is one way to accomplish this; and the knowledge that the final product will eventually be sent out for peer review provides an incentive for quality throughout the evaluation process. Inclusion of reviewer comments as an appendix to the published final report, as in Jacobson and Petta (2000) plays a similar role.

### *5.3 Cost-Benefit Analysis*

Evidence-based policy builds on a foundation of serious and thorough cost-benefit analyses of various policy alternatives. Cost-benefit analyses, in turn, rest on a foundation of high-quality econometric program evaluations. Many trees have given their lives for books on cost-benefit evaluation. This subsection does not attempt a general treatment, but instead highlights a few key issues that have received too little attention in the literature. The discussion here draws on the discussion in Section 10.2 of Heckman, LaLonde and Smith (1999).

Impact evaluations commonly generate net impact estimates for a short time, usually a few months or years. Yet, at least in certain contexts, such as human capital development programs or infrastructure investments, we expect impacts to persist for some time. In such contexts, two related issues arise. The first issue is how to project the

estimated benefits outside the period of the data. While theory or evidence from other evaluations of similar programs with longer follow-up periods can play a role in guiding this decision, in the end the best course will likely consist of constructing estimates of the cost-benefit performance of the program assuming multiple plausible durations for program benefits. The recent experimental Job Corps evaluation does not do this, and as a result, particularly in its executive summary provides a somewhat misleading guide to policy. That evaluation presumes as a base case that program impacts last (essentially) forever and concludes on that basis that the program easily passes a cost-benefit test (see Table 3 in Burghardt et al. (2001) and the surrounding discussion). Yet because of the high cost of this program, and the relatively short period of post-program data collection, without the assumed future benefits the program would have a positive gross impact (which represents a major achievement relative to most government employment and training programs for youth) but would fail the cost benefit test miserably.

The second issue is what discount rate to use for future benefits. A small literature exists that attempts to estimate optimal social discount rates under various assumptions (see, e.g., the discussion and references in Liu (2003)). Once again, reporting cost-benefit estimates for multiple plausible rates seems best.

Another issue in cost-benefit analysis concerns the deadweight cost of taxation, called the “excess burden” in the public finance literature. This number measures the cost to the economy of the marginal tax dollar including (ideally) both the direct costs of operating the tax system and the indirect costs of the distortions induced by the tax system. The literature offers a wide variety of estimates of this cost, ranging from only a few cents to well over one dollar per dollar of tax revenue. See, e.g., Browning (1987)

and Snow and Warren (1996) for further discussion and evidence. Once again, given the uncertainty in the literature, presenting multiple estimates based on different values seems the best course (and ignoring the deadweight costs altogether, as too many evaluations do, seems the worst course).

In the case of each of the cost-benefit issues considered here, presenting multiple estimates that rely on different assumptions does two important things. First, it allows readers with different prior beliefs than the evaluator about these issues to see the cost-benefit estimates under his or her preferred assumptions. Second, it highlights to policymakers the range of cost-benefit estimates consistent with the data. This forces them to confront uncertainty about program performance and at the same time provides a sense of the robustness of any recommendations concerning the policy.

## **6.0 Conclusion**

The resources devoted to local economic development programs have valuable alternative uses. Econometric program evaluations play a key role in determining when to continue with economic development programs and when to shut them down. In this chapter, I have emphasized five key themes in evaluating such programs.

First, and perhaps foremost, I have emphasized the importance of reading the literature. We have learned a lot about how to do econometric policy evaluations in the last two decades, but this knowledge has not yet affected evaluation practice to the extent that it should. The knowledge we have gained includes both advances in methods, as well as advances in practice, including the use of administrative data and clever identification strategies.

Second, there is no magic bullet. No econometric evaluation estimator provides consistent estimates for all (or even most) possible combinations of data, institutions and parameter of interest. Regression does not do this, matching does not do this, the bivariate normal model does not do this, difference-in-differences does not do this, and IV does not do this. The search for such an estimator, which animated the literature for many years, has now come to an end, replaced by a more sensible research program designed to identify the mapping between characteristics of the data and institutions and the parameter of interest to the estimators likely to yield consistent answers.

Third, heterogeneous treatment effects matter. They affect the choice and interpretation of econometric evaluation estimators. They imply careful thought about the exact parameter of interest required to answer a particular policy question. The conceptual literature is advancing rapidly in this area, but has already revolutionized how evaluators think about what they do.

Fourth, general equilibrium effects matter in many evaluation contexts, particularly when considering local economic development programs, which generally aim to create such effects. The potential presence of general equilibrium effects has important implications for the joint decision regarding which econometric evaluation estimator to employ and the unit of analysis for the evaluation. Depending on the unit of analysis, some estimators will miss or, worse still, be biased by, general equilibrium effects. Pick the unit of analysis too large, and program effects get lost in the shuffle; pick the unit of analysis too small and general equilibrium effects get missed. General equilibrium effects seem to attract bad evaluation practices as well, particularly the use of magic multipliers in evaluations of public infrastructure investments.

Finally, not every program will benefit from an evaluation. Before proceeding with one, some thought, some power calculations, and an informal cost-benefit analysis of the evaluation itself will help to sort out situations where an evaluation represents a sound investment from situations where it represents a waste of time and money.

## Bibliography

- Agodini, Roberto and Mark Dynarski. 2001. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Mathematica Policy Research Working Paper No. 8723-300*.
- Angrist, Joshua and Guido Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*. 90(430): 431-432.
- Angrist, Joshua and Alan Krueger. 1999. "Empirical Strategies in Labor Economics." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Volume 3A*. Amsterdam: North-Holland. 1277-1366.
- Angrist, Joshua and Alan Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives*. 15(4): 69-85.
- Ayres, Ian and John Donohue. 2002. "Shooting Down the More Guns, Less Crime Hypothesis." NBER Working Paper No. 9336.
- Barnow, Burt. 1999. "Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act." *Journal of Policy Analysis and Management*. 19(1): 118-141.
- Barnow, Burt and Jeffrey Smith. 2003. "Performance Management of U.S. Job Training Programs." Unpublished manuscript, University of Maryland.
- Bartik, Timothy. 2003a. "Local Economic Development Policies." W.E. Upjohn Institute Staff Working Paper No. 03-91.
- Bartik, Timothy. 2003b. "Thoughts on American Manufacturing Decline and Revitalization." *Employment Research*. 10(4): 1-4.
- Bartik, Timothy. 2004. "Evaluating the Impacts of Local Economic Development Policies On Local Economic Outcomes: What Has Been Done and What is Doable?" In this volume.
- Bell, Stephen, Larry Orr, John Blomquist and Glen Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs*. Kalamazoo: W.E. Upjohn Institute for Employment Research.
- Black, Dan and Daniel Nagin. 1998. "Do Right-to-Carry Laws Deter Violent Crime?" *Journal of Legal Studies*. 27(1): 209-219.

- Black, Dan and Jeffrey Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching." *Journal of Econometrics*, forthcoming
- Black, Dan, Jeffrey Smith, Mark Berger and Brett Noel. 2003. "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System." *American Economic Review*. 93(4): 1313-1327.
- Black, Sandra. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics*. 114(2): 577-599.
- Bloom, Dan and Charles Michalopoulos. 2001. *How Welfare and Work Policies Affect Employment and Income: A Synthesis of Research*. New York: Manpower Demonstration Research Corporation.
- Bloom, Howard. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*. 82(2): 225-246.
- Blundell, Richard and Monica Costa Dias. 2000. "Evaluation Methods for Non-Experimental Data." *Fiscal Studies*. 21(4): 427-468.
- Blundell, Richard and Monica Costa Dias. 2002. "Alternative Approaches to Evaluation in Empirical Microeconomics." *Portuguese Economic Journal*. 1(1): 91-115.
- Blundell, Richard, Monica Costa Dias and Costas Meghir. 2003. "The Impact of Wage Subsidies: A General Equilibrium Approach." Unpublished manuscript, University College, London.
- Blundell, Richard, Monica Costa Dias, Costas Meghir and John Van Reenan. 2001. "Evaluating the Impact of a Mandatory Job Search Assistance Program." IFS Working Paper No. WP01/20.
- Blundell, Richard and James Powel. 2001. "Endogeneity in Semiparametric Binary Response Models." CEMMAP Working Paper No. CWP05/01.
- Bradburn, Norman, Seymour Sudman and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design*. Jossey-Bass.
- Browning, Edgar. 1987. "On the Marginal Welfare Cost of Taxation." *American Economic Review*. 77(1): 11-23.
- Burghardt, John, Peter Schochet, Sheena McConnell, Terry Johnson, Mark Gritz, Steven Glazer, John Homrighausen and Russell Jackson. 2001. *Does the Job Corps Work? Summary of the National Job Corps Study*. Princeton, NJ: Mathematica Policy Research.
- Calmfors, Lars. 1994. "Active Labor Market Policy and Unemployment – A Framework for the Analysis of Crucial Design Features." *OECD Economic Studies*. 22(1): 7-47.

- Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In Louis Christofides, Kenneth Grant and Robert Swidinsky, eds., *Aspects of Labour Market Behavior: Essays in Honor of John Vanderkamp*. Toronto: University of Toronto Press. 201-222.
- Coates, Dennis and Brad Humphreys. 1999. "The Growth Effects of Sport Franchises, Stadia and Arenas." *Journal of Policy Analysis and Management*. 18(4): 601-624.
- Crompton, John. 1995. "Analysis of Sports Facilities and Events: Eleven Sources of Misapplication." *Journal of Sports Management*. 9(1): 14-35.
- Davidson, Carl and Stephen Woodbury. 1993. "The Displacement Effects of Reemployment Bonus Programs." *Journal of Labor Economics*. 11(4): 575-605.
- Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94(448): 1053-1062.
- Dehejia, Rajeev and Sadek Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics*. 84(1): 151-161.
- Doolittle, Frederick and Linda Traeger. 1990. *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.
- Eichler, Martin and Michael Lechner. 2002. "An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt." *Labour Economics*. 9(2): 143-186.
- Evans, William and Diana Lien. 2002. "The Benefits of Prenatal Care: Evidence from the PAT Bus Strike." Unpublished manuscript, University of Maryland.
- Evans, William and Julie Topoleski. 2002. "The Social and Economic Impact of Native American Casinos." NBER Working Paper No. 9198.
- Eberts, Randall and Christopher O'Leary. 2004. "Evaluating Training Programs: Impacts at the Local Level." In this volume.
- Fraker, Thomas and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs." *Journal of Human Resources*. 22(2): 194-227.
- Friedlander, Daniel and Philip Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review*. 85(4): 923-937.

- Frölich, Markus. 2004. "Finite Sample Properties of Propensity Score Matching and Weighting Estimators." *Review of Economics and Statistics*. Forthcoming.
- Glazerman, Steven, Dan Levy and David Myers. 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science*. 589: 63-93.
- Greenberg, David and Mark Shroder. 1997. *Digest of Social Experiments, 2<sup>nd</sup> Edition*. Lanham, Maryland: Rowman and Littlefield.
- Greene, William. 2002. *Econometric Analysis, 5<sup>th</sup> Edition*. Upper Saddle River, NJ: Prentice Hall.
- Gregory, Amanda. 2000. "Problematizing Participation: A Critical Review of Approaches to Participation in Evaluation Theory." *Evaluation*. 6(2): 179-199.
- Gueron, Judith and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage.
- Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*. 66(2): 315-331.
- Hahn, Jinyong, Petra Todd and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*. 69(1): 201-209.
- Hansen, Lars and James Heckman. 1996. "The Empirical Foundations of Calibration." *Journal of Economic Perspectives*. 10(1): 87-104.
- Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*. 47(1): 153-161.
- Heckman, James. 1996. "Randomization as an Instrumental Variable." *Review of Economics and Statistics*. 78(2): 336-341.
- Heckman, James, Carolyn Heinrich and Jeffrey Smith. 2001. "The Performance of Performance Standards." *Journal of Human Resources*. 37(4). 778-811.
- 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 Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*. 84(408): 862-874.

- Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*. 66(5): 1017-1098.
- Heckman, James, Hidehiko Ichimura and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*. 64(4): 605-654.
- Heckman, James, Hidehiko Ichimura and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*. 65(2): 261-294.
- Heckman, James, Robert LaLonde and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Volume 3A*. Amsterdam: North Holland. 1865-2097.
- Heckman, James, Lance Lochner and Christopher Taber. 1998. "General Equilibrium Treatment Effects: A Study of Tuition Policy." *American Economic Review*. 88(2): 281-386.
- Heckman, James, Lance Lochner and Christopher Taber. 1999. "General Equilibrium Cost-Benefit Analysis of Education and Tax Policies." In Gustav Ranis and Kakshmi Raut, eds., *Taxes, Growth and Development: Essays in Honor of Professor T.N. Srinivasan*. New York: Elsevier. 291-349.
- Heckman, James and Salvador Navarro-Lozano. 2004. "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics*. Forthcoming.
- Heckman, James and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press for Econometric Society Monograph Series. 156-246.
- Heckman, James and Jeffrey Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*. 9(2): 85-110.
- Heckman, James and Jeffrey Smith. 1998. "Evaluating the Welfare State." In Steiner Strom, ed., *Econometrics and Economic Theory in the 20<sup>th</sup> Century: The Ragnar Frisch Centennial*. Cambridge University Press for Econometric Society Monograph Series. 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 Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts." *Review of Economic Studies*. 64(4): 487-537.
- Heckman, James, Jeffrey Smith and Christopher Taber. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics*. 80(1): 1-14.
- Heckman, James, Justin Tobias and Edward Vytlačil. 2003. "Simple Estimators for Treatment Parameters in a Latent Variable Framework." *Review of Economics and Statistics*. 85(3): 748-754.
- Heckman, James and Edward Vytlačil. 1998. "Instrumental Variable Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling when the Return Is Correlated with Schooling." *Journal of Human Resources*. 33(4): 974-987.
- Heckman, James and Edward Vytlačil. 2001a. "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling." *Review of Economics and Statistics*. 83(1): 1-12.
- Heckman, James and Edward Vytlačil. 2001b. "Local Instrumental Variables." In Cheng Hsiao, Kimio Morimune and James Powell, eds., *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*. Cambridge: Cambridge University Press.
- Heckman, James and Edward Vytlačil. 2004. "The Econometric Evaluation of Social Programs." In James Heckman and Edward Leamer, eds., *Handbook of Econometrics, Volume 6*. Amsterdam: North-Holland. Forthcoming.
- Hollister, Robinson, Peter Kemper and Rebecca Maynard. 1984. *The National Supported Work Demonstration Project*. Madison: University of Wisconsin Press.
- Hotz, V. Joseph and Karl Scholz. 2002. "Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data." In *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press. 275-315.
- Hui, Shek-Wai and Smith, Jeffrey. 2002. "The Labor Market Impacts of Adult Education and Training in Canada." Report Prepared for Human Resources Development Canada.
- Imbens, Guido. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika*. 87(3): 706-710.
- Imbens, Guido. 2004. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Reivew." *Review of Economics and Statistics*, forthcoming.

- Imbens, Guido and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*. 62(4): 467-476.
- Jacobson, Lou and Ian Petta. 2000. "Measuring the Effects of Public Labor Exchange (PLX) Referrals and Placements in Washington and Oregon." Workforce Security Occasional Paper No. 2000-06, Employment and Training Administration, U.S. Department of Labor.
- Kydland, Finn and Edward Prescott. 1996. "The Computational Experiment: An Econometric Tool." *Journal of Economic Perspectives*. 10(1): 69-85.
- LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*. 76(4): 604-620.
- Lechner, Michael. 1999. "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification." *Journal of Business and Economic Statistics*. 17(1): 74-90.
- Lechner, Michael. 2000. "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany." *Journal of Human Resources*. 35(2): 347-375.
- Lechner, Michael. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption." In Michael Lechner and Friedhelm Pfeiffer, eds., *Econometric Evaluation of Labor Market Policies*. Heidelberg: Physica. 43-58.
- Liu, Liqun. 2003. "A Marginal Cost of Funds Approach to Multi-Period Public Project Evaluation: Implications for the Social Discount Rate." *Journal of Public Economics*. 87(7-8): 1707-1718.
- Lise, Jeremy, Shannon Seitz and Jeffrey Smith. 2003. "Equilibrium Policy Experiments and the Evaluation of Social Programs." Unpublished manuscript, University of Maryland.
- Long, Sharon and Douglas Wissoker. 1995. "Welfare Reform at Three Years: The Case of Washington's Family Independence Program." *Journal of Human Resources*. 30(4): 766-790.
- Lott, John and David Mustard. 1997. "Crime, Deterrence and Right-to-Carry Concealed Handguns." *Journal of Legal Studies*. 26(1): 1-68.
- Maxwell, Scott. 2000. "Sample Size and Multiple Regression Analysis." *Psychological Methods*. 5(4): 434-458.

- Meyer, Bruce. 1995. "Lessons from the U.S. Unemployment Insurance Experiments." *Journal of Economic Literature*. 33(1): 91-131.
- Michalopoulos, Charles, Howard Bloom and Carolyn Hill. 2004. "Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" *Review of Economics and Statistics*, forthcoming.
- Moffitt, Robert. 1991. "Program Evaluation with Nonexperimental Data." *Evaluation Review*. 15(3): 291-314.
- Moffitt, Robert. 2003. "Remarks on the Analysis of Causal Relationships in Population Research." Unpublished manuscript, Johns Hopkins University.
- Newey, Whitney, James Powell and James Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review*. 80(2): 324-328.
- Newhouse, Joseph. 1994. *Free for All: Lessons from the Rand Health Insurance Experiment*. Cambridge, MA: Harvard University Press.
- Noll, Roger and Andrew Zimbalist. 1997. *The Economic Impact of Sports Teams and Facilities*. Washington, DC: Brookings Institution.
- Orr, Larry. 1998. *Social Experiments: Evaluating Public Programs with Experimental Methods*. PLACE: Sage Publications.
- Osborne, David and Ted Gaebler. 1992. *Reinventing Government: How The Entrepreneurial Spirit is Transforming the Public Sector*. Boulder, CO: Perseus.
- Pence, Karen. 2003. "Foreclosing on Opportunity: State Laws and Mortgage Credit." FEDS Working Paper 2003-16.
- Portney, Paul. 1994. "The Contingent Valuation Debate: Why Economists Should Care." *Journal of Economic Perspectives*. 8(4): 3-17.
- Puhani, Patrick. 2000. "The Heckman Correction for Sample Selection and Its Critique." *Journal of Economic Surveys*. 14(1): 53-68.
- Ravallion, Martin. 2001. "The Mystery of the Vanishing Benefits: An Introduction to Evaluation." *World Bank Economic Review*. 15(1): 115-140.
- Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*. 70(1): 41-55.

- Rosenbaum, Paul and Donald Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *American Statistician*. 39: 33-38.
- Sims, Christopher. 1996. "Macroeconomics and Methodology." *Journal of Economic Perspectives*. 10(1): 105-120.
- Smith, Jeffrey. 2000. "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies." *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 136(3): 247-268.
- Smith, Jeffrey and Arthur Sweetman. 2001. "Improving the Evaluation of Employment and Training Programs in Canada." Unpublished manuscript, University of Maryland.
- Smith, Jeffrey and Petra Todd. 2004. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, forthcoming.
- Smith, Jeffrey and Alexander Whalley. 2004. "Are Participants Good Evaluators? Evidence from the Job Training Partnership Act" Unpublished manuscript, University of Maryland.
- Snow, Arthur and Ronald Warren. 1996. "The Marginal Welfare Cost of Public Funds: Theory and Estimates." *Journal of Public Economics*. 61(2): 289-305.
- Sweetman, Arthur, William Warburton, Rob McPhee and Rebecca Warburton. 2003. "Disability Status: Impacts on Health and Welfare Dependence." Unpublished manuscript, Queen's University.
- U.S. Department of Labor. 2000. "Core and Customer Satisfaction Performance Measures for the Workforce Investment System." Training and Employment Guidance Letter No. 7-99. Washington, DC: Employment and Training Administration.
- Van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Design." *International Economic Review*. 43(4): 1249-1287.
- Wallsten, Scott. 2000a. "The R&D Boondoggle." *Regulation*. 23(4): 12-17.
- Wallsten, Scott. 2000b. "The Effects of Government-Industry R&D Programs on Private R&D: The case of the Small Business Innovation Research Program." *Rand Journal of Economics*. 31(1). 82-100.
- Wilson, James. 2000. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.
- Winship, Christopher and Stephen Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology*. 25: 659-706.

Wooldridge, Jeffrey. 2001. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey. 2002. *Introductory Econometrics: A Modern Approach, 2<sup>nd</sup> Edition.* Mason, Ohio: South-Western College Publishing.

Table 1 Self-Assessments of JTPA Impact: Experimental Treatment Group  
(National JTPA Study, 18-Month Impact Sample)

	Adult Males	Adult Females	Male Youth	Female Youth
<b>Full-sample percentages</b>				
Percentage who self-report participating	61.63 (0.81)	68.10 (0.68)	62.62 (1.29)	66.29 (1.09)
Percentage of self-reports participants with positive self-assessments	62.46 (1.04)	65.21 (0.85)	67.16 (1.59)	71.73 (1.29)
Overall percentage with positive self-assessments	38.49 (0.81)	44.41 (0.73)	42.06 (1.32)	47.55 (1.16)
<i>Percentage of self-reported participants with a positive self-assessment by primary treatment received</i>				
None (dropouts)	48.89 (2.07)	51.44 (1.85)	58.90 (3.33)	61.56 (2.79)
Classroom training in occupational skills	74.10 (2.15)	73.47 (1.36)	72.73 (3.60)	75.28 (2.30)
On-the-job training at private firm	75.13 (2.18)	78.90 (2.14)	71.00 (4.56)	75.00 (4.04)
Job-search assistance	59.57 (2.27)	59.80 (2.18)	68.09 (3.94)	68.94 (4.04)
Basic education	62.96 (4.67)	56.55 (3.84)	70.97 (4.09)	78.44 (3.19)
Work experience	66.67 (9.83)	68.75 (5.84)	82.76 (7.14)	73.17 (7.01)
Others	58.47 (3.65)	66.40 (2.98)	62.50 (4.77)	77.98 (3.99)

Source: Table 8.11 of Heckman and Smith (1998).

Notes: (1) Reported proportions are based on responses to the question “Do you think that the training or other assistance you got from the program helped you get a job or perform better on the job?” This question was asked only of self-reported participants within the treatment group. The overall fraction of positive self-assessments assumes that self-reported non-participants would have provided negative self-assessments.

(2) The primary treatment is the one in which the trainee participated for the most hours according to the administrative records of the JTPA sites. Most trainees received only one service; few received more than two. See Smith (in press) for a detailed discussion. Note that for some self-reported participants the JTPA administrative records indicate that no services were received.

(3) Estimated standard errors in parentheses.

Figure 1

