Reprint from:

EVALUATION REVIEW

"PROGRAM EVALUATION WITH NONEXPERIMENTAL DATA"

by Robert Moffitt

Books
Journals
Newsletters
University Papers
Annual Series

2455 Teller Road
Newbury Park
California 91320

Phone (805) 499-0721
Fax (805) 499-0871
Cable SagePub
Telex (510) 1000799

Sage Publications, Inc.

Statistical methods for program evaluation with nonexperimental data have been studied by economists and econometricians over the last 20 years. These methods are concerned with laying out the precise circumstances under which valid nonexperimental estimates of the effects of an intervention can be obtained, and then with methods for determining when and if those circumstances hold. This article provides a simple exposition of the methods of identification that have been developed and draws the lessons of those methods for future evaluation designs, data collection, and analysis.

PROGRAM EVALUATION WITH NONEXPERIMENTAL DATA

ROBERT MOFFITT
Brown University

Economists and econometricians have been studying statistical methods for program evaluation with nonexperimental data for at least 20 years. The major historical impetus for interest among economists was provided by the need to evaluate many of the social programs of the 1960s, particularly those designed to aid the low-income population with educational programs, training programs, and transfer benefits. Early studies by Goldberger (1972) and Cain (1975) were followed by many others, including those of Ashenfelter (1978) and the studies surveyed by Barnow (1987). A major shift in the econometric literature occurred with the introduction of "selectivity bias" methods (Gronau 1974; Lewis 1974; Heckman 1974), whose implications for program evaluation were first drawn by Barnow, Cain, and Goldberger (1980) and were later surveyed in textbook form by Maddala (1983). The most recent and most complete discussion of econometric methods for program evaluation has been provided by Heckman and Robb (1985a, 1985b).[1]

This article provides an exposition of these methods in relatively simple terms and without much of the technical language employed in the literature

AUTHOR'S NOTE: *This research was partially supported by the Panel on the Evaluation of AIDS interventions of the National Research Council. Comments on earlier versions of the article from James Heckman, V. Joseph Hotz, James Knickman, Charles Manski, and two anonymous referees are appreciated. All opinions and errors are those of the author.*

EVALUATION REVIEW, Vol. 15 No. 3, June 1991 291-314

and is designed to make the issues in the literature more accessible. The first issue discussed here is the precise delineation of the conditions under which estimates of the program impact derived from nonexperimental data are "valid" in a sense defined below. Following that, the article discusses how it can be determined whether or not those conditions are met. Finally, the article discusses the implications of the delineation and testing of the conditions for the design of future evaluations and the types of data that should be collected.

The outline of the article is as follows. In the next section, the nature of the evaluation problem is defined and the econometric solutions to that problem are presented. The third section provides discussion of methods of testing different assumptions, particularly in the manner recently developed by Heckman, Hotz, and Dabos (1987) and Heckman and Hotz (1989), and shows the importance of data availability in that testing procedure. The implications of the methods for future program evaluation are discussed in the final section.

IDENTIFYING PROGRAM IMPACTS WITH NONEXPERIMENTAL DATA

THE PROBLEM

Suppose that we wish to evaluate the effect of a particular intervention (i.e, a treatment) on individual levels of some outcome variable. Let Y be the outcome variable and make the following definitions:

Y_{it}^{*} = level of outcome variable for individual i at time t if he or she has not received the treatment

Y_{it}^{**} = level of outcome variable for same individual i at same time t if he or she has received the treatment at some prior date.

The difference between these two quantities is the effect of the treatment, denoted α:

$$Y_{it}^{**} = Y_{it}^{*} + \alpha \qquad [1]$$

or

$$\alpha = Y_{it}^{**} - Y_{it}^{*}. \qquad [2]$$

The aim of the evaluation is to obtain an estimate of the value of α, the treatment effect. The easiest way to think about what we seek in an estimate of α is to consider individuals who have gone through a program and therefore have received the treatment, and for whom we later measure their value of Y_{it}^{**}. Ideally, we wish to know the level of Y_{it}^{*} for such individuals—that is, we would like to know what their level of Y would have been had they not gone through the program. If Y_{it}^{*} could be known, the difference between it and Y_{it}^{**} would be a satisfactory estimate of α.[2]

The difficulty that arises does so because we do not observe Y_{it}^{*} directly, but only the values of Y_{it}^{*} for nonparticipants of the program. Define a dummy variable for whether an individual has or has not received the treatment:

$d_i = 1$ if individual i has received the treatment
$= 0$ if individual i has not received the treatment.

Then an estimate of α could be obtained by estimating the difference between Y_{it}^{**} and Y_{it}^{*} for those who did and did not go through the program, respectively:

$$\tilde{\alpha} = E(Y_{it}^{**} \mid d_i = 1) - E(Y_{it}^{*} \mid d_i = 0) \qquad [3]$$

where $E(Y_{it}^{**} \mid d_i = 1)$ is the expected, or average, value of Y_{it} of those who have received the treatment and $E(Y_{it}^{*} \mid d_i = 0)$ is the expected, or average, value of Y_{it} for those who have not received the treatment. Unfortunately, this is not what we wish to calculate, for we wish to calculate the difference between the expected value of Y_{it}^{**} for those with $d_i = 1$ and the expected value of Y_{it}^{*} that would have obtained for those with $d_i = 1$ as well—that is, the value of Y that would have arisen if those who did go through the program had not gone through it. That is, we would like to know

$$\hat{\alpha} = E(Y_{it}^{**} \mid d_i = 1) - E(Y_{it}^{*} \mid d_i = 1). \qquad [4]$$

The estimate $\hat{\alpha}$ in (4) is, in fact, the estimate that would be obtained if we had successfully administered a randomized controlled trial for the evaluation. For example, as individuals come in through the door of the program, they would be randomly assigned to treatment status or control status, where the latter would involve receiving none of the services of the program. At some later date we could measure the levels of Y for the two groups and calculate (4) to obtain an estimate of the effect of the program.[3]

When will the estimate we are able to calculate, $\tilde{\alpha}$, equal the estimate we would have obtained with a randomized trial, $\hat{\alpha}$? Comparison of (3) and (4) shows that the two will be equal if and only if the following condition is true:

$$E(Y_{it}^{*} \mid d_i = 1) = E(Y_{it}^{*} \mid d_i = 0). \quad [5]$$

In words, the two estimates of α are equal only if the expected value of Y_{it}^{*} for those who did not take the treatment equals the expected value of Y_{it}^{*} that those who did take the treatment would have had, had they not gone through the program.

The heart of the nonexperimental evaluation problem is reflected in equation (5), and an understanding of that equation is necessary to understand the pervasiveness and unavoidability of what is termed the *selection bias* problem when nonexperimental data are employed. The equation will fail to hold under many plausible circumstances. For example, if those who go through a health counseling program designed to encourage the adoption of better health practices happen to be those especially concerned with their health, and who have already begun adopting good health practices even before entering the program, they will be quite different from those who do not go through the program even prior to receiving any program services. Hence equation (5) will fail to hold because those who go through the program have different levels of Y_{it}^{*}, that is, different levels of good health behavior even in the absence of receiving any program services. The estimate of $\tilde{\alpha}$ will be too high relative to $\hat{\alpha}$, for the greater level of good health behavior observed for the treatment group subsequent to receiving services was present even prior to the treatment and is therefore not necessarily a result of the treatment itself. Those who are observed to have actually gone through the program are therefore a "self-selected" group out of the pretreatment population, and the estimate of $\tilde{\alpha}$ is contaminated by selectivity bias because of such self-selection. Put differently, the population of nonparticipants constitutes a "nonequivalent" comparison group.

The selection bias problem can also be thought of as an omitted variable or missing-data problem, in this case the omitted variable being Y_{it}^{*}. In the example just given, it may be that prior health practices can be an adequate proxy for Y_{it}^{*}, and hence inclusion of that variable will eliminate the bias, but this will not always be the case. The use of preprogram information on Y_{it} is discussed in detail in the next section.

The unavoidability of the potential for selectivity bias arises because the validity of equation (5) cannot be tested, even in principle, for the left-hand side of that equation is inherently unobservable. It is impossible in principle

to know what the level of Y_{it}^* for those who went through the program would have been had they not gone through it, for that level of Y_{it}^* is a "counterfactual" that can never be observed. We may know the pretreatment level of Y_{it} for those who later undergo treatment, but this will often not be the same as the Y_{it}^* we seek—for the left-hand side of (5), we need to know the level of Y_{it}^* for program participants that they would have had at exactly the same time as Y_{it}^{**} is measured, not at some previous time.[4]

Before discussing the solutions to this identification problem in the literature, it is important to point out that the object of the estimation—the true impact of the treatment on Y—may differ across persons. Equation (2), by omitting a subscript on α, implicitly assumes that the treatment impact is the same for all. An alternative is to replace α by α_i:

$$\alpha_i = Y_{it}^{**} - Y_{it}^* . \qquad [6]$$

To keep matters simple, equation (2) will be assumed in the analysis below rather than (6). However, it is important to note that equation (6) is not only more plausible than (2), but it has a critical bearing on the implications of any evaluation. Equation (6) is more plausible because it seems intuitive that many individuals will react to a particular treatment differently, for reasons that may be, in principle, measurable but that will often not be measured in the data available. The implications for program evaluation are critical because the estimate of program impact obtained in any particular evaluation will depend on which individuals, among all those in the population of interest, have been administered the treatment. For example, if an evaluation of a small program present in only a single local area of the country is conducted, and if the small size of the program has been achieved by admitting only high-impact individuals, the estimated treatment effect may differ considerably from that which would obtain if the program were implemented nationally, on a larger scale, and if, therefore, individuals with lower impacts were brought in. In general, the difference between (2) and (6) is of critical importance for the extrapolation of the results of a particular evaluation to other areas, other populations, national programs, or programs of a different scale.[5]

SOLUTIONS

There are three general classes of potential solutions to the selection bias problem (Heckman and Robb 1985a, 1985b).[6] Each defines circumstances under which the problem could be eliminated and what type of estimation

method would do so. The question is then whether it can be determined whether those circumstances hold, a question addressed in the next section. It is important to note that two of the three solution methods, the first and second in the order listed below, have important implications for evaluation design because they require the collection of certain types of data. These implications will be drawn out in the final section.

Solution Method 1: Identifying variables (Z's). The selection bias problem can be solved if a variable Z_i is available, or one can be found, that satisfies two conditions: (a) It affects the probability that an individual receives the treatment, but (b) it has no direct relationship to Y_{it}^* (e.g., no direct relationship to individual health practices in the example discussed previously). What is an example of such a Z_i? Suppose that a health counseling program is funded by the federal government and that the government funds the program in one neighborhood of a city and not in another neighborhood for political or bureaucratic reasons unrelated to the health needs of the populations in the two areas—and therefore unrelated to the health practices of the individuals in the two. If a random sample of the populations or subpopulations of interest were conducted in the two neighborhoods and if data on Y_{it} were collected (the data would include both participants and nonparticipants in the neighborhood where the program was funded) a comparison of the mean values of Y_{it} in the two would form the basis for a valid estimate of α.[7] The variable Z_i in this case should be thought of as a dummy variable equal to 1 in the neighborhood with the program and 0 in the other. The variable satisfies the two conditions given above—it obviously affects whether individuals in the two areas receive the treatment, because if $Z_i = 0$, no treatment is available, and it is unrelated to the level of Y_{it}^* in the two areas because the funding decision was made for reasons unrelated to health practices. This is a case of what is often termed a *natural* experiment, similar to an experiment inasmuch as the probability of having the treatment available is random with respect to the outcome variable under study as a result of natural variation.[8] This estimation method is also termed an instrumental-variable method in econometrics, where Z_i is the instrument. Indeed, the method of instrumental variables in econometrics is, in a fundamental sense, a generalization of the concept of a natural experiment.[9]

What is an example of an illegitimate Z_i? The same dummy variable just defined in the previous example would be illegitimate if the government funding decision were based not on political or bureaucratic decisions but on the relative level of health practices in the two areas, for example, if the health counseling program were placed in the neighborhood with the higher rate of illness. In that case, the dummy variable Z_i would not be independent of

Y_{it}^* – the presence of the program in a neighborhood would be associated with lower levels of good health behavior not because of a negative causal effect of the treatment but because of the reason for its placement.

Many other examples could be given as well. For example, it is possible that individual, rather than geographic characteristics could serve as legitimate instruments. If participation rates in a program differ for individuals with different age or educational levels, for example, variation in values of Y_{it} across those groups will provide a valid estimate of program impact, provided that those same characteristics do not directly affect Y_{it} in the absence of the program. However, for a Y defined as a variable measuring health practices, age and education are unlikely to satisfy the latter requirement. It is variation in the availability, rather than the actual receipt, of treatment across the population that is more likely to provide a legitimate Z in this case. Other examples include cases where Z_i is continuous rather than dichotomous. For example, if government funding levels of a health program are nonzero in all neighborhoods, but funding decisions are made for political and bureaucratic reasons, the level of funding itself defines a legitimate Z_i. In this case, rather than comparing the levels of Y_{it} for each different neighborhood with a different level of Z_i, some smoothing technique such as regression analysis could be used.

It is also important to note that a particular Z_i variable would still be legitimate if it were only partly unrelated to the value of Y_{it}^*, at least if that part could be isolated in the analysis. For example, if funding decisions are made by ranking areas by the level of Y_{it}^*, with funding starting at, say, the lowest Y_{it}^* (e.g., the lowest level of good health practices) and moving up the list, no legitimate Z_i could be based on funding. But if areas were grouped into categories – say, "high," "medium," and "low" Y_{it}^* – and if funding decisions were made for political or bureaucratic reasons within each of the categories, the variation in funding within category (i.e., conditional on category) would furnish a legitimate Z_i. Thus it is only necessary that some portion of the variation in the variable be isolated that satisfies the requirements for a legitimate Z_i.[10]

Provided that a legitimate Z_i is found, there are a variety of estimation techniques available.[11] Econometric practice typically employs linear regression formulations in which the influences of variables other than the treatment are controlled by least squares and where a formal instrumental-variables estimation procedure is used to remove the endogeneity of the treatment variable. Corresponding nonlinear procedures are available for nonlinear models. In some cases, particularly where the treatment impact varies across individuals in a random fashion (see equation [6]), other

two-stage methods such as the "lambda" technique may be used (Heckman 1979; Maddala 1983). In the absence of a distributional assumption – neither normality nor any other distribution has been assumed in any of the discussion thus far – it is the availability of a legitimate Z_i that permits the identification of the treatment impact in this estimation method as well.[12]

Clearly, the most important question is whether such a Z_i is available. But how is the investigator to know if a particular candidate for Z_i is or is not legitimate? This issue will be discussed in the next section.

Solution Method 2: Availability of cohort data. A second solution method requires the availability of "cohort," "longitudinal," or "panel" data, that is, data on the same individuals at several points in time before and after some of them have undergone the treatment. In the simplest case, data on Y are available not only after the treatment but also before, giving a data set with one pretreatment observation and one posttreatment observation for each individual, both participants and nonparticipants. In the more general case, three or more points in time may be available in the data.

The advantage of such data is that the past values of Y_{it} for an individual, those prior to the receipt of the treatment, may provide a good measure of the unobservable Y^*_{it}. Suppose, for example, that different individuals differ in their inherent healthiness, and that individuals select themselves into treatment on the basis of some permanent, unchangeable level of health (e.g., the least healthy are more likely to enroll in a health program). Then it may be that differences in past health practices of (future) participants and nonparticipants in a health program may adequately control for their differences in Y^*_{it}, the unobservable component of health. If so, the availability of longitudinal data can eliminate the selectivity bias that would be present in only a single cross section of data.

The use of such longitudinal data is sufficiently important to warrant an extended discussion. To illustrate this method, first consider the situation that would arise if data at two points in time were available, one before the treatment and one after it. Let "t" denote the posttreatment point and "t – 1" denote the pretreatment point. Then, analogously with the cross-sectional case considered previously,

$Y^*_{it} - Y^*_{i,t-1}$ = change in Y^*_{it} from t – 1 to t in the absence of having undergone the treatment

$Y^{**}_{it} - Y^*_{i,t-1}$ = change in Y^*_{it} from t – 1 to t if having undergone the treatment.

Then the effect of the treatment is α, and

$$Y_{it}^{**} - Y_{i,t-1}^{*} = (Y_{it}^{*} - Y_{i,t-1}^{*}) + \alpha . \quad [7]$$

Because $Y_{i,t-1}^{*}$ cancels out on both sides of (7), it is the same as (1) and therefore the true effect, α, is the same.

We can use the data on Y_{it}^{*} available from nonparticipants before and after the treatment to estimate the program effect as follows:

$$\tilde{\alpha} = E(Y_{it}^{**} - Y_{i,t-1}^{*} \mid d_i = 1) - E(Y_{it}^{*} - Y_{i,t-1}^{*} \mid d_i = 0) . \quad [8]$$

This estimator $\tilde{\alpha}$ is often called a "differences" or "change" estimator because it is computed by comparing the first-differenced values of Y for participants and nonparticipants. As before, a preferred estimate of the effect of the program would be obtained by a randomized controlled trial in which those wishing to undergo the treatment ($d_i = 1$) are randomly assigned to participation or nonparticipation status. With data on both pretreatment and posttreatment Y, the estimate of the program effect could be calculated as

$$\hat{\alpha} = E(Y_{it}^{**} - Y_{i,t-1}^{*} \mid d_i = 1) - E(Y_{it}^{*} - Y_{i,t-1}^{*} \mid d_i = 1) . \quad [9]$$

Unfortunately, with nonexperimental data the second term on the right-hand side of (9) is not measurable because, once again, we cannot measure Y_{it}^{*} for those who undergo the treatment.

The estimate we are able to obtain in (8) will equal that we could have obtained in the randomized trial, (9), if and only if

$$E(Y_{it}^{*} - Y_{i,t-1}^{*} \mid d_i = 1) = E(Y_{it}^{*} - Y_{i,t-1}^{*} \mid d_i = 0) . \quad [10]$$

Equation (10) is the key equation for the two-data-point case and is the analogue to equation (5) in the single-post-treatment-data-point case. The equation shows that a data set with a pretreatment and posttreatment observation will yield a good estimate of α if the change in Y_{it}^{*} from pre to post would have been the same for participants, had they not undergone the treatment, as it actually was for nonparticipants. Sometimes the change in Y_{it}^{*} is referred to as the "growth rate" of Y_{it}^{*}, in which case we may say that our nonexperimental estimate requires that the growth rate of Y for participants and nonparticipants be the same in the absence of the treatment.

Perhaps the most important point is that this condition may hold even though condition (5) does not. Equation (5), the condition that must hold for

the nonexperimental estimate in a single posttreatment cross section to be correct, requires that the *levels* of Y_{it} be the same for participants and nonparticipants in the absence of the treatment. Equation (10), on the other hand, only requires that the *growth rates* rates of Y_{it} be the same for participants and nonparticipants in the absence of the treatment, even though the levels may differ. The latter is a much weaker condition and will more plausibly hold in many situations.

The nature of the condition is illustrated in panels a and b of Figure 1. In panel a, the pretreatment levels of nonparticipants and participants, A and A′, respectively, are quite different—participants have a higher level of Y, as would be the case, for example, if those who later undergo a health program have higher levels of good health practices in the first place. From t – 1 to t, the level of Y for nonparticipants grows from A to B, as might occur if everyone in the population under consideration were increasing their level of good health practices even without participating in a program. The figure shows, for illustration, a growth rate of Y for participants from A′ to C, which is a larger rate of growth than for nonparticipants. The estimate of the treatment effect, $\tilde{\alpha}$, is also shown in the figure and is based on the assumption that, in the absence of undergoing the program, the Y of participants would have grown from A′ to B′—in other words, by the same amount as the Y of nonparticipants grew. Of course, this assumption cannot be verified because point B′ is not observed; it is only a counterfactual. But clearly the estimate in the figure would be a much better estimate than that obtained from a single posttreatment cross section, which would take the vertical distance between B and C as the treatment estimate. This would be invalid because equation (5) does not hold.

Panel b in Figure 1 shows a case where condition (10) breaks down. In that panel, a case is shown in which the Y of participants would have grown faster than that for nonparticipants even in the absence of treatment (A′ to B′ is greater than A to B). This might arise, for example, if those individuals who choose to undergo a health counseling program are adopting better health practices more quickly than are nonparticipants. In this case, the estimate of $\tilde{\alpha}$ is too high, because it measures the vertical distance between B″ and C instead of between B′ and C. Neither B′ nor B″ is observed, so we cannot know which case holds.

A major conclusion to be drawn from this discussion is that a superior estimate of program effect may, under certain conditions, be obtainable with more data.[13] Adding a single pretreatment data point permits the computation of an estimate of the treatment effect—the differences estimator in (8)—that

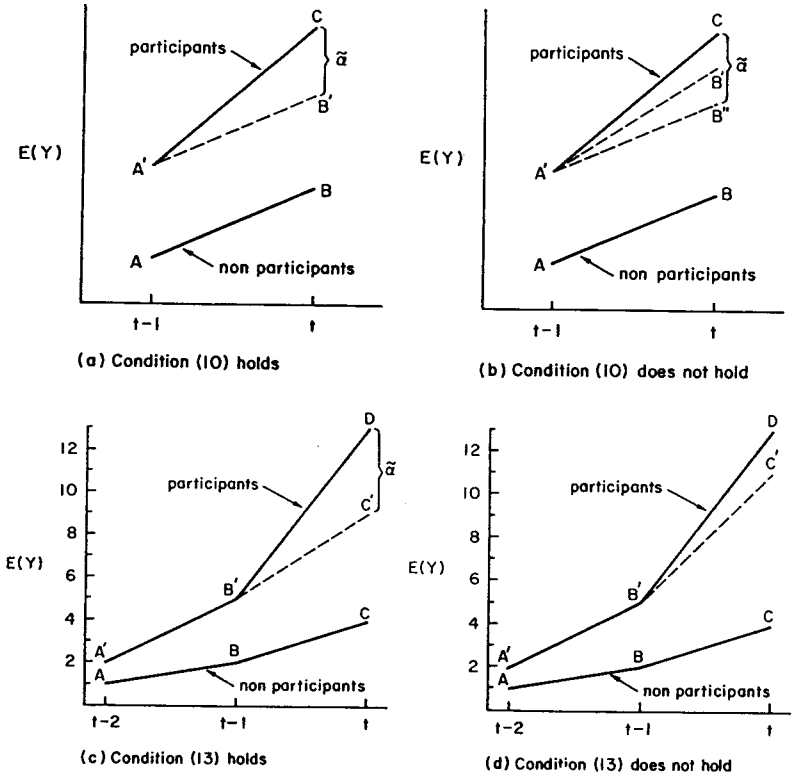

Figure 1: Alternative Trends for Participants and Nonparticipants

may be correct in circumstances in which the estimator using a single posttreatment is not. The importance of having additional data on the histories of Y, for example, stands in contrast to the situation faced when conducting a randomized trial, where, strictly speaking, only a single posttreatmen cross section is required. Thus is can be concluded that more data may be required for valid inference in nonexperimental evaluations than in experimental evaluations.

This point extends to the availability of additional pretreatment observations.[14] If, for example, levels of health are not constant and immutable, as discussed previously, but are instead changing over time, it may be that participants and nonparticipants select themselves into participation in a

program on the basis of the growth rates, not the levels, of Y^*_{it}. Additional pretreatment observations permit this issue to be examined. If an additional pretreatment observation is available at time $t-2$, for example, the estimate permitted in a nonexperimental study is

$$\tilde{\alpha} = E[(Y^{**}_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 1] - E[(Y^*_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 0], \quad [11]$$

whereas the estimate permitted in a randomized trial is

$$\hat{\alpha} = E[(Y^{**}_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 1] - E[(Y^*_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 1]. \quad [12]$$

This estimator in (11) is often termed a *differences-in-differences* estimator because it computes the treatment effect by comparing the "change in the change" of Y for participants and nonparticipants.

The two estimators will be equal if and only if

$$E[(Y^*_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 1] = E[(Y^*_{it} - Y^*_{i,t-1}) - (Y^*_{i,t-1} - Y^*_{i,t-2}) \mid d_i = 0]. \quad [13]$$

Equation (13) shows that a correct program impact estimate will be obtained only if the change in the growth rate of Y would have been the same for program participants in the absence of their having undergone treatment as it actually was for nonparticipants. Panel c of Figure 1 illustrates the situation when this condition holds. For both participants and nonparticipants, Y_{it} grows at an increasing rate over time as, for example, would occur if the adoption of good health practices were accelerating in the population. For nonparticipants, Y_{it} grows by 1 from $t-2$ to $t-1$ (A to B) and by 2 from $t-1$ to t (B to C). For participants, Y_{it} grows by 3 from $t-2$ to $t-1$ (A′ to B′) and by 8 from $t-1$ to t (B′ to D). The estimate of program effect, shown in the figure, is therefore 4 because it is assumed that the growth rate of Y_{it} for participants in the absence of the treatment would have accelerated by the same amount as it did for nonparticipants, namely, by 1 — from a growth rate of 3 between $t-2$ and $t-1$ to a growth rate of 4 between $t-1$ and t (B′ to C′). Panel d shows a case where condition (13) does not hold — there, the growth rate of Y_{it} for participants accelerates by more even in the absence of the treatment than it did for nonparticipants (B′ to C′ is greater than B to C).

The conclusion to be drawn from this discussion is that the availability of three points of data permits us to obtain an estimate of program effect that

may be valid in circumstances in which the estimate possible with two points of data is incorrect. For example, the application of the differences estimator in (8) to the data shown in panel c would give an incorrect estimate of program effect, for, in the absence of the treatment, the growth rates of Y_{it} for participants and nonparticipants between t – 1 and t (B′ to C′ and B to C, respectively) are not equal (e.g., because unobserved levels of health grow at different rates and individuals select themselves into program participation on the basis of their growth rates). In fact, this is the case illustrated in panel b, where the differences method gives an incorrect estimate. Thus, once again, the conclusion to be drawn is that more data permit the calculation of program effects that may be valid in circumstances in which the estimate available with less data is not.

An analogous implication holds if we consider four, five, or many points of preprogram data. More periods of data make possible estimates of treatment effects that are equal to those obtainable from a randomized trial under weaker and weaker conditions, thereby strengthening the reliability of the nonexperimental estimator. In the general case, a slight modification in the model allows us to write the estimate of the treatment effect as the following:[15]

$$\tilde{\alpha} = E(Y_{it}^{*} \mid d_i = 1, Y_{i,t-1}^{*}, Y_{i,t-2}^{*}, \ldots, Y_{i,t-k}^{*}) - E(Y_{it}^{*} \mid d_i = 0, Y_{i,t-1}^{*}, Y_{i,t-2}^{*}, \ldots, Y_{i,t-k}^{*}) \quad [14]$$

assuming that data are available for k pretreatment periods. This estimator will equal that obtainable in a randomized trial if and only if the following condition holds:

$$E(Y_{it}^{*} \mid d_i = 1, Y_{i,t-1}^{*}, \ldots, Y_{i,t-k}^{*}) = E(Y_{it}^{*} \mid d_i = 0, Y_{i,t-1}^{*}, \ldots, Y_{i,t-k}^{*}) . \quad [15]$$

This condition can be interpreted as requiring that the values of d_i and Y_{it}^{*} must be independent of one another conditional on the history of Y_{it}^{*} up to t – 1. Put differently, it must be the case that if two individuals are observed at time t – 1 who have exactly the same history of Y_{it}^{*} up to that time (e.g., the exact same history of health practices) – and who therefore look exactly alike to the investigators – they must have the same value of Y_{it}^{*} in the next time period regardless of whether they do or do not undergo the treatment. If, on the other hand, the probability of entering a health counseling program is related to the value of Y_{it}^{*} they would have had if the treatment were not available, the condition in equation (15) will not hold and the nonexperimental estimate will be inaccurate.

Solution Method 3: Parametric distributional assumptions on Y^*. In both solution methods discussed thus far, no parametric distributional assumptions have been placed on Y_{it}^* or on any of the other variables in the analysis. By implication, valid treatment impact estimates are obtainable by the use of nonparametric estimating procedures. The variable Y_{it}^* is unobservable in principle, as stressed at the beginning of the discussion; solution Methods 1 and 2 both provide valid impact estimates because an observed variable or set of variables is available that is independent of the unobservable Y_{it}^*. In the first solution method, a legitimate Z_i is one that is observable (obviously) and independent of the unobservable Y_{it}^*, whereas in the second solution method, treatment status (d_i) is independent of Y_{it}^* conditional on a particular observed Y history. The second method is, in fact, a special case of the first, where Z_i is defined as the treatment variation conditional on the Y history (i.e., the residual treatment variation).

A third solution to the selection bias problem can be achieved if the distribution of the unobservable Y_{it}^* conditional on d_i is known directly or can be determined with reasonable certainty. For example, if Y_{it}^* follows a normal, logistic, or some other distribution with a finite set of parameters, identification of a program effect free of selection bias is often possible.[16] It is often thought, for example, that certain inherent biological traits are distributed normally in the population. If it can be assumed with relative certainty that the distribution of Y_{it} in the *absence* of the treatment (i.e., Y_{it}^*) is so distributed, this approach is possible. Implicitly, the approach measures the effect of the treatment by deviations from normality of the distribution of Y_{it} within the sample of participants. A more formal method of thinking about estimation in this case can be obtained by considering a method-of-moments estimation method in which not only the difference in the means of Y_{it} for the participant and nonparticipant populations is used in the analysis but also the difference in the higher-order moments of Y_{it}. With knowledge of a particular parametric distribution for Y_{it}^*, these higher-order moments will take on particular functional forms that may permit the identification of the parameters of the distribution and therefore the degree of selection bias present.

Unfortunately, this method will not be especially useful for most interventions because we usually do not have firm prior knowledge regarding the distribution of the unobservable Y_{it}^* (i.e., the distribution of Y_{it} in the entire population in the absence of the treatment). Although it is fairly clear how outside information on the nature of the funding process can inform choice of potential Z_i variables, for example, and how outside information on the nature of the Y_{it} process can inform assumptions regarding similarity of Y_{it}

histories for participants and nonparticipants, it is unclear how outside information can be obtained to inform assumptions regarding the distribution of Y_{it}^*. In most cases, distributional information will be available only for self-selected groups of nonparticipants, or for participants and nonparticipants at a time prior to the intervention, neither of which provides necessarily accurate information on the distribution of Y_{it}^* because of the possibility of selection bias already discussed. The only population from which firm knowledge of the distribution of Y_{it}^* can be obtained is from a true control group, which is not available by assumption, or from a population group defined by a legitimate Z_i, in which case the Z_i can be used to identify the treatment impact rather than a distributional assumption on Y_{it}^*. Given these difficulties, this solution method will not be considered further.[17]

THE ROLE OF TESTING OF ASSUMPTIONS AND ITS RELATIONSHIP TO DATA AVAILABILITY

The discussion thus far has demonstrated that the availability of certain types of data—information on legitimate Z variables, or on individual histories—is related to the conditions that must hold, and the assumptions that must be made, to obtain an estimate of program effect similar to that obtainable in a randomized trial. However, the delineation of these conditions and assumptions is not particularly useful unless some means is developed for determining which, if any, of the conditions are met and which, if any, of the assumptions hold. Otherwise, the assumptions made by any particular investigator will be arbitrary and may differ across investigators. This section lays out the formal mechanism for conducting the necessary testing, based largely on the work of Heckman, Hotz, and Dabos (1987) and Heckman and Hotz (1989). The following section discusses the implications of these testing procedures for design strategy and data collection. The most important point will be that the type of data collection undertaken in any investigation bears importantly on the ability to test which of the conditions hold and which assumptions are met.

The formal answer to the question of whether and when assumptions can be tested is that "overidentifying" assumptions can be tested but that "just identifying" assumptions cannot be. An overidentifying assumption is one that could be dropped and a valid treatment estimate still obtained without it; a just identifying assumption is one that cannot be dropped because a valid treatment estimate could not be obtained without it. Whether an assumption

is overidentifying or just identifying depends on the data set available, for an assumption can be dropped—and therefore tested—if the available data are a bit more than are actually needed to estimate the model in question. As more data become available, many just identifying assumptions—which cannot be tested—can be turned into overidentifying assumptions—which can be tested.

The relationship between testing of assumptions and data availability is illustrated in Figure 2, which shows five different models that can be estimated on different data sets. The model at the top of the figure can be estimated on Data Set 1, while the two models below it can be estimated on a richer data set, Data Set 2, and the two models below that can be estimated on a yet richer data set, Data Set 3. At the top of the figure, it is presumed that the evaluator has a data set (Data Set 1) consisting of a single posttreatment data point with Y_{it} information, but no other variables at all—in particular, no Z_i variable is in the data set. The best the analyst can do in this circumstance is to compare the Y_{it} means of participants and nonparticipants to calculate $\tilde{\alpha}$ as in equation (4) above. This estimate will equal that obtainable from a randomized trial under the three assumptions shown in the box for Model I in the figure: that the missing Z_i is independent of Y_{it}^* conditional on d_i and that there is no selection bias in either levels of first differences. The first assumption is necessary to avoid omitted-variable bias, the bias generated by leaving out of the model an important variable that is correlated with both the probability of receiving the treatment and Y_{it}^*. Suppose, for example, that Z_i is a dummy for neighborhood location, as before. If location is an important determinant of health behavior, and if the probability of treatment also varies across areas, then not having a variable for city location in the data set will lead to bias because the estimate of program impact (the difference in mean Y_{it} between participants and nonparticipants) reflects, in part, interarea differences in health practices that are not the result of the treatment but were there to begin with. The second and third assumptions are necessary for the value of Y_{it}^* for nonparticipants to be the proper counterfactual, that is, for it to equal the value that participants would have had, had they not undergone the treatment.[18]

Modes II and III in the figure can be estimated if the data set contains information on a potential Z_i, like neighborhood location, but still only a single posttreatment observation on Y_{it} (Data Set 2). Each of these models requires only two assumptions instead of three, as in Model I, but each model drops a different assumption. Model II drops the assumption that there is no selection bias in levels—that is, it drops the assumption that (5) holds.

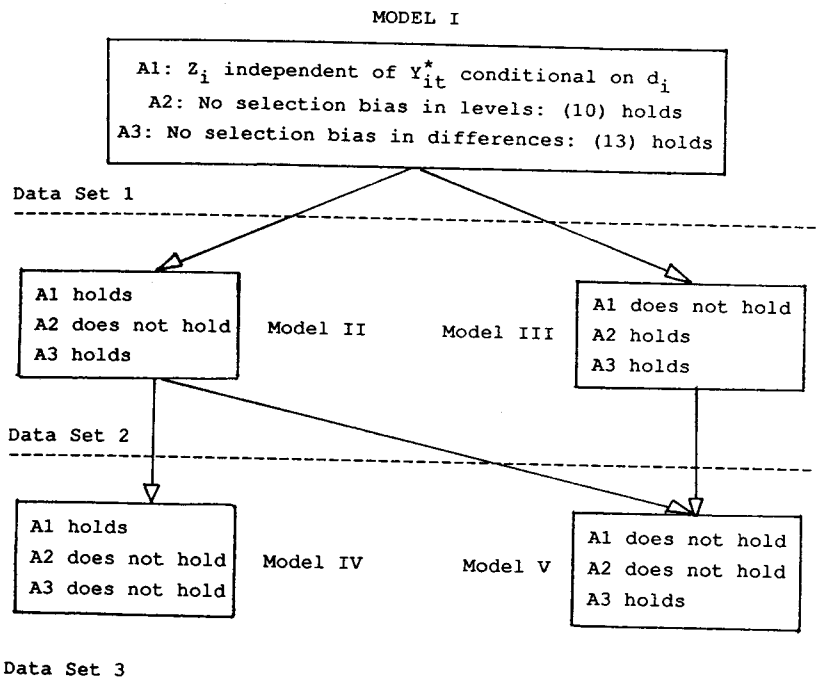

Figure 2: Estimable Models With Different Data Sets
NOTE: Data Set 1: Single postprogram, no Z_i. Data Set 2: Single postprogram, Z_i. Data Set 3: Preprogram and postprogram, Z_i.

This assumption can be dropped because a Z_i is now available and the instrumental-variable technique described earlier as Solution 1 can now be used alone to obtain a valid estimate of α. In this method, the values of Y_{it} for participants and nonparticipants in a given neighborhood are not compared to one another to obtain a treatment estimate – that estimate would be faulty because participants are a self-selected group. Instead, mean values of Y_{it} across neighborhoods are compared to one another, where the neighborhoods differ in the availability of the treatment and therefore have different treatment proportions (e.g., a proportion of 0 if the neighborhood has no program at all, as in the example given previously). For the treatment-effect estimate from this model to be accurate still requires the assumption that the Z_i is a legitimate instrument – that the differential availability of the program

across areas is not related to the basic levels of health behavior in each city (i.e., that Z_i and Y^*_{it} are independent).

Not only does Model II require one less assumption than does Model I, it also permits the testing of that assumption and therefore the testing of the validity of Model I. The test of the dropped assumption—that there is no selection bias in levels—is based on a comparison of impact estimates obtained from the two models. If the two are the same or close to one another, then it must be the case that there is, in fact, no selection bias in levels—because the impact estimate in Model I is based on participant-nonparticipant comparisons whereas that in Model II is not. If the two are different, then there must be selection bias—if the participant-nonparticipant differences within cities do not generate the same impact estimate as that generated by the differences in Y_{it} across different cities, the former must be biased because the latter is accurate (under the assumption that the Z_i available is legitimate).

Model III takes the opposite tack and drops the assumption that Z_i is legitimate but maintains the assumption that there is no selection bias in levels. The model estimates the treatment effect by making participant-nonparticipant comparisons only within areas, that is, conditional on Z_i. If there are neighborhoods where the program is not present at all, data on Y_{it} from those neighborhoods is not used at all, unlike the method in Model II. The Model III impact estimate will be accurate if there is no selection bias into participation, but it will also be accurate even if interarea variation is not a legitimate Z_i (e.g., if program placement were based on prior health need). In this case, a comparison of the impact estimate with that obtained from Model I—where participants and nonparticipants across areas were pooled into one data set and location was not controlled for because the variable was not available—provides a test for whether interarea variation is a legitimate Z_i. If it is not (e.g., if program placement across cities is based on prior health need)—the Models I and III will produce quite different treatment estimates, for Model I does not control for location but Model III does (Model III eliminates cross-neighborhood variation entirely by examining only participant-nonparticipant differences within neighborhoods). On the other hand, if neighborhood location is a legitimate Z_i (e.g., if program placement is independent of prior health need) then the two estimates should be close to one another.

The implication of this discussion is that Data Set 2 makes it possible to reject Model I by finding its assumptions to be invalid. This testing of Model I is possible because Data Set 2 provides more data than is actually necessary to estimate that model. Unfortunately, this data set does not allow the evaluator to test the assumptions of Models II and III necessary to assure

their validity. Each makes a different assumption – Model II assumes that Z_i is legitimate, whereas Model III assumes no selection bias to be present – and the estimates from the two need not be the same. If they are different, the evaluator must gather additional information.

Such additional information may come from detailed institutional knowledge – for example, knowledge of whether Z_i is truly legitimate (e.g., detailed knowledge of how programs are placed across neighborhoods). But another source of additional information is additional data, such as information on a preprogram measure of Y_{it}. For example, if Data Set 2 is expanded by adding a preprogram measure of Y (Data Set 3 in Figure 2), the assumptions of Models II and III can be tested by estimating Models IV and V shown in the figure. Each of these models drops yet another assumption, although a different one in each case. Model IV drops the assumption that there is no selection bias in differences but continues to make the assumption that Z_i is a legitimate instrument. The impact estimate is obtained by the instrumental-variable technique, as in Model II, but in this case by comparing the means of $(\overline{Y}_t - \overline{Y}_{t-1})$ across neighborhoods, thereby eliminating selection bias in levels if there is any. Model V drops the assumption that there is no selection bias in levels by applying the difference estimator in (8) but still assumes that there is no selection bias in differences.

Once again, the richer data set permits the testing of the assumptions that went into Models II and III and therefore makes possible their rejection. The arrows in the figure between models show which models can be tested against one another. A comparison of the estimates of Model IV to those of Model II provides a test of the third assumption (that there is no selection bias in differences); a comparison of the estimates of Model V and Model II provides a test of the first assumption (that Z_i is a legitimate instrument); and a comparison of the estimates of Model V and Model III provides a test of whether the second assumption holds (that there is no selection bias in levels). If each comparison indicates estimates that are similar to one another, the relevant assumption in the more restricted model (Model II or Model III) should be taken to be valid; when estimates differ, however, the assumption involved should be taken as invalid and the more restricted model should be rejected.

As before, Models IV and V now require certain assumptions for their impact estimates to be valid. The estimates required for each are different, but neither can be tested unless more information or more data are available. With Data Set 3, they are nontestable, just identifying assumptions. An additional preprogram data point or an additional Z_i variable would enrich that data set and would convert those assumptions to testable, overidentifying

assumptions. New models made possible by increasing the richness of the data set permit the evaluator to discard more and more assumptions and therefore obtain impact estimates that are more and more reliable. This strategy can be pursued until models are found that are not rejected by richer data sets.[19]

IMPLICATIONS FOR EVALUATION DESIGN AND DATA COLLECTION

The discussion in the second and third sections has many implications for design of program evaluations and for data collection efforts. First, the discussion in the second section indicates the directions that investigators must take to search for means of identifying program impacts, namely, toward the search for potential Z's and the investigation of longitudinal histories, both of which will require data collection. Second, the discussion in the third section provides the basis for a design strategy in which data collection of Z's and histories is expanded until valid identifying variables are found. This, in turn, provides important information for future evaluations because it indicates which types of data are necessary, and which are not, to obtain valid program impact estimates. Over the course of a sequence of investigations in the same area, knowledge can thus be built up within the evaluation community on the types of data necessary to obtain good estimates.

The research agenda posed by this evaluation strategy is ambitious in its scope. A search for Z's will likely require detailed investigations of the funding process, for example, and of the determinants of the allocation of programs to different groups. More generally, as many past observers have noted, the search for Z's requires a detailed investigation of the selection process itself (i.e., the process that determines who ends up in treatment and who does not). However, more than an investigation of why particular individuals do or do not enter treatment, the examples used here point more toward an investigation of the reasons for the availability of programs in different areas and to different groups. Such an investigation is likely to involve field work to determine why programs are available in some neighborhoods and cities and not in others, and to determine why different socioeconomic groups appear to have greater access than do others, even though two different groups appear to reside in an area where the program is available. For example, if a particular set of individuals has a low participation rate in a program because they are not served by appropriate public