The Review *of* Economics *and* Statistics

VOL. LXXX FEBRUARY 1998 NUMBER 1

ACCOUNTING FOR DROPOUTS IN EVALUATIONS OF SOCIAL PROGRAMS

James Heckman, Jeffrey Smith, and Christopher Taber*

Abstract—This paper explores issues that arise in the evaluation of social programs using experimental data in the frequently encountered case where some of the experimental treatment group members drop out of the program prior to receiving treatment. We begin with the standard estimator for this case and the identifying assumption upon which it rests. We then examine the behavior of the estimator when the dropouts receive a partial "dose" of the program treatment prior to dropping out of the program. In the case of partial treatment, the identifying assumption is typically violated, thereby making the estimator inconsistent for the conventional parameter of interest: the impact of full treatment on the fully treated. We develop a test of the identifying assumption underlying the standard estimator and consider whether exclusion restrictions produce identification of the mean impact of the program when this assumption fails to hold. Finally, we discuss alternative parameters of interest in the presence of partial treatment among the dropouts and argue that the conventional parameter is not always the economically interesting one. We apply our methods to data from a recent experimental evaluation of the Job Training Partnership Act (JTPA) program.

I. Introduction

IN RECENT years, social experiments have gained popularity as a method for evaluating social and labor market programs. A common problem affecting social experiments is that persons randomly assigned to the experimental treatment group often drop out of the program under study prior to receiving some or all of the treatment.[1] In the presence of dropouts, the usual experimental mean difference estimator provides an estimate of the mean impact of the assignment to treatment rather than of the mean impact of the treatment itself.

In this paper we consider the evaluation of social programs using experimental data when there are dropouts from the program. We begin with an instrumental-variable estimator variously attributed to Mallar et al. (1980), Smith et al. (1984, p. 251), and Bloom (1984).[2] This estimator is commonly used to produce estimates of the mean impact of treatment on the treated in experiments with dropouts. In experiments in which the dropouts receive none of the treatment prior to dropping out, this estimator performs very well in estimating the impact of full treatment on the fully treated. However, in the commonly encountered case where the dropouts receive some treatment prior to leaving the program, the estimator will generally not provide consistent estimates of the mean impact of treatment on the treated, which is often the parameter of interest in conducting evaluations.

This paper considers the case of partial treatment. In it, we (1) define the key identifying assumption underlying the estimator; (2) show why in the partial treatment case this assumption is likely to be violated; (3) discuss the economic content of the parameter in this case; (4) develop statistical tests of an implication of the identifying assumption; and (5) explore alternative instrumental-variable approaches to the identification of treatment effects when dropouts receive partial treatment. In the final section of the paper we analyze data from a recent experimental evaluation of the Job Training Partnership Act (JTPA) program (see Bloom et al. (1993)). The instrumental-variable estimator was employed in the JTPA evaluation even though many of the dropouts received some JTPA services prior to dropping out.

II. The Problem Posed by Dropouts

Consider an experimental evaluation in which persons who apply and are accepted into a program are assigned randomly into a treatment group eligible to receive the treatment and a control group ineligible to receive it. Assuming that everyone in the treatment group receives the treatment, the mean impact of the treatment on some outcome Y is defined as

$$\Delta = E(Y_t) - E(Y_c) \tag{1}$$

where Δ is the mean impact, E denotes mathematical expectation, $E(Y_t)$ is the mean outcome in the treatment group, and $E(Y_c)$ is the mean outcome in the control group.[3] An unbiased estimator of Δ is

$$\overline{\Delta} = \overline{Y}_t - \overline{Y}_c \tag{2}$$

Received for publication September 26, 1994. Revision accepted for publication January 17, 1997.

* University of Chicago, University of Western Ontario, and Northwestern University, respectively.

We have benefited from comments received at a Demography Workshop at the University of Chicago in January 1994, sponsored by the Population Research Center and chaired by V. J. Hotz. This research was supported by NSF SBR-91-11455, NSF-SBR-93-09525, and a grant from the Russell Sage Foundation.

[1] We are assuming that persons drop out of the program, but not out of the experimental data. Thus we do not discuss attrition bias in this paper.

[2] Heckman and Robb (1985, sec. 3.8) developed this estimator independently in the context of adjusting for contamination bias. In later work, Angrist and Imbens (1991) and Dubin and Rivers (1993) rediscovered the estimator, apparently independently.

[3] We leave implicit the conditioning on the event that all persons in the sample have applied to and have been accepted into the program.

where the overbar denotes sample mean. Estimates based on equation (2) can be constructed using post-random-assignment outcome data on those assigned at random.

Now suppose that some persons in the experimental treatment group actually do not receive the treatment.[4] Define d to be an indicator of dropout status (d for dropout), so that $d = 1$ for a treatment group member who drops out of the program and $d = 0$ for a treatment group member who receives the treatment. Let $E(Y_t|d = 1)$ denote the mean outcome of the treatment group members not receiving the treatment (the dropouts), let $E(Y_t|d = 0)$ denote the mean outcome of those receiving the treatment (the participants), and let k_t denote the probability of dropping out of the program for members of the treatment group. Then the mean outcome in the treatment group may be decomposed in the following way:

$$E(Y_t) = k_t E(Y_t|d = 1) + (1 - k_t)E(Y_t|d = 0). \tag{3}$$

Random variable d is in principle also defined for members of the control group, so that $d = 1$ if a control group member would have dropped out had he or she been a member of the treatment group, and $d = 0$ is the complementary event. The control group mean outcome can then be decomposed into two components:

$$E(Y_c) = k_c E(Y_c|d = 1) + (1 - k_c)E(Y_c|d = 0) \tag{4}$$

where $E(Y_c|d = 1)$ denotes the mean outcome of those control group members who would have been dropouts had they been in the treatment group, $E(Y_c|d = 0)$ denotes the mean outcome of those control group members who would have been participants had they been randomized into the treatment group, and k_c denotes the probability that a control would have dropped out of the program. Decomposition (4) is not empirically operational because we do not observe d for control group members.

When there are dropouts from an experiment, the estimator $\overline{\Delta}$ defined in equation (2) provides an estimate of the mean impact of the *availability* of the treatment[5] rather than an estimate of the mean impact of "full treatment on the fully treated," where this latter quantity is defined as

$$\Delta_p = E(Y_t|d = 0) - E(Y_c|d = 0). \tag{5}$$

The parameter Δ_p is widely regarded to be of greater interest than the mean impact of treatment availability. It is informative about the difference in outcomes obtained given receipt of full treatment and given receipt of no treatment for persons who actually received the full treatment. It is not the same as the effect of full treatment on all initial enrollees into the program unless dropping out is random with respect to potential outcomes.[6] For an evaluation of an ongoing program in which participants routinely drop out, it is not clear that Δ_p should be the only parameter of interest.

Data on the post-treatment outcomes of those in the treatment and control groups are not enough to estimate Δ_p because $E(Y_c|d = 0)$ is not known. Data on observed choices and outcomes alone provide no way to sort the control group into those who would and those who would not have been dropouts. Additional assumptions or additional data are required.

III. An Identifying Assumption and an Estimator That Implements It

An identifying assumption made by Mallar et al. (1980), Smith et al. (1984), and Bloom (1984), among others, is that the mean outcome of the dropouts in the treatment group is the same as the mean outcome of the persons in the control group who would have been dropouts, had they been in the treatment group. More formally, it is assumed that

$$E(Y_t|d = 1) = E(Y_c|d = 1). \tag{6}$$

Only equality of the conditional means is required. Variances and other parameters of the conditional distributions $(Y_t|d = 1)$ and $(Y_c|d = 1)$ need not be equal. To see how this assumption produces identification of Δ_p, first note that random assignment implies that

$$k_t = k_c = k. \tag{7}$$

Combining this fact with assumption (6) allows equation (4) to be solved for $E(Y_c|d = 0)$,

$$E(Y_c|d = 0) = \left(\frac{1}{1 - k}\right)E(Y_c) - \left(\frac{k}{1 - k}\right)E(Y_t|d = 1). \tag{8}$$

Note that each component of the right-hand side of equation (8) can be identified from the available data. Simple algebra using equations (5)–(8) reveals that

$$\Delta_p = \frac{E(Y_t) - E(Y_c)}{1 - k} = \frac{\Delta}{1 - k}. \tag{9}$$

[4] Note that this is not an attrition problem. We assume that everyone who is assigned randomly remains in the sample, but that some treatment group members drop out of the program and so fail to receive some or all of the treatment.

[5] This parameter is often referred to as the "intent to treat" parameter in the biostatistics literature (see, e.g., Efron and Feldman (1991)).

[6] That is, Δ_p corresponds to the mean impact of tretment on all initial enrollees only when (Y_t, Y_c) is statistically independent of d. Technically, all that is required is mean independence, namely, that $E(Y_t|d) = E(Y_t)$ and $E(Y_c|d) = E(Y_c)$.

The natural method of moments estimator replaces population parameters by their sample counterparts,

$$\bar{\Delta}_p = \frac{\bar{Y}_t - \bar{Y}_c}{1 - \hat{k}_t} = \frac{\bar{\Delta}}{1 - \hat{k}_t} \tag{10}$$

where $\hat{k}_t$ is the fraction of the treatment group members who drop out. Writing the estimator in this form shows that it scales up the estimate of the mean impact of treatment availability into an estimate of the mean impact of full program participation by assuming a zero impact of the program on the dropouts. The asymptotic variance for this estimator is[7]

$$\text{var}(\bar{\Delta}_p) = \frac{\text{var}(\bar{Y}_t) + \text{var}(\bar{Y}_c)}{(1-k)^2} + \left[\frac{E(\bar{Y}_t) - E(\bar{Y}_c)}{(1-k)^2}\right]^2 \frac{k(1-k)}{n_t} \tag{11}$$

where n_t is the sample size of the treatment group sample. If k is assumed to be known, the second term vanishes.

While the instrumental-variable estimator is widely used, the standard errors reported in empirical implementations of it typically do not account for the estimation of k. In results not reported in detail in this paper, we show using data from a recent experimental evaluation of the JTPA program that ignoring sampling error in k is not empirically important in that evaluation, which had sample sizes well in excess of 1000 for each demographic group (see Heckman et al. (1994)).

While the literature confines its attention to mean impacts, both the identifying assumption (6) and the resulting estimator can be generalized to allow the full distribution of outcomes of control group members who would have participated in the program had they been in the treatment group to be obtained. Letting $F_t(y|d = 1)$ be the distribution of $(Y_t|d = 1)$ and letting $F_c(y|d = 1)$ be the distribution of $(Y_c|d = 1)$, the distributional analog to assumption (6) is

$$F_t(y|d = 1) = F_c(y|d = 1). \tag{12}$$

This assumption is clearly stronger than assumption (6). It is not needed to construct the estimator $\bar{\Delta}_p$ given in equation (10). However, many economic models that justify assumption (6) will also justify assumption (12). When assumption (12) fails, there is no a priori reason to expect assumption (6) to hold except by coincidence. Under assumption (12) the outcome distribution for members of the control group who would have received treatment had they been in the treatment group is given by

$$F_c(y|d = 0) = \frac{F_c(y) - kF_t(y|d = 1)}{1 - k}. \tag{13}$$

Assumption (12) is useful for two reasons. First, if one is interested in some aspect of the distribution of Y_c given $d = 0$ other than just its mean, it will be identified from equation (13). Second, assumption (12) provides a testable restriction that we now investigate.

IV. Testing the Strengthened Version of the Identifying Assumption

In this section we describe the strongest testable restriction implied by assumption (12). The appendix develops a test of the restriction. As discussed previously, while the estimator defined in equation (10) requires only the weaker assumption (6), this assumption is implied by assumption (12) and is unlikely to hold in its absence except by coincidence. Thus it is of interest to test restrictions on the distribution of outcomes implied by assumption (12) for persons in the control group who would have participated (received full treatment) had they been in the treatment group.

The restriction that we examine follows from equation (13), which is an immediate consequence of assumption (12). For assumption (12) to hold, $F_c(y|\ d = 0)$ must be a legitimate distribution function, that is,

$$\frac{F_c(y) - kF_t(y|d = 1)}{1 - k} \tag{14}$$

is a proper cumulative distribution function (cdf).

Condition (14) imposes a number of testable restrictions on the component distributions, such as

$$0 \leq \frac{F_c(y) - kF_t(y|d = 1)}{1 - k} \leq 1 \tag{15}$$

for all y. Another restriction is that the constructed cdf should be monotonically increasing in y.

A weakness of our proposed strategy is that any test of assumption (12) based on restriction (14) may not be consistent. (A test is consistent if, as the sample size becomes large, the power of the test goes to 1.) The problem arises because assumption (12) is sufficient for restriction (14) but not necessary; that is, restriction (14) may hold even when assumption (12) does not. Any test based on restriction (14) will have no power against alternatives that are consistent with it but not with assumption (12). Thus rejection of restriction (14) constitutes rejection of assump-

[7] The derivation is a straightforward application of the delta method (see, e.g., Angrist and Imbens (1991) or Heckman et al. (1994)).

tion (12), but acceptance of restriction (14) is not necessarily evidence in support of assumption (12).[8]

Note that in general the power of the test will depend on k. When k is close to 1, the restriction is very strong. However, as k becomes smaller, the restriction becomes weaker.

While in some cases, tests of restriction (14) have no power, the restriction is the strongest testable restriction implied by assumption (12). To see this, suppose that restriction (14) is true and define

$$F_c^*(y|d=0) = \frac{F_c(y) - kF_t(y|d=1)}{1-k}. \tag{16}$$

When restriction (14) is true, we can never reject $F_c = F_c^*$ (that is, we could never find evidence that the data were not generated by F_c^*). Furthermore, $F_c = F_c^*$ implies assumption (12). Therefore it is impossible to find a testable restriction of assumption (12) that is violated when restriction (14) is not. The appendix describes a testing strategy for hypothesis (14). We next discuss identification in the presence of partial treatment.

V. Role of Partial Treatment When Identifying Assumption (6) Fails

A. *Partial Treatment*

For certain types of programs, assumptions (6) and (12) are plausible. For example, in drug trials, if dropouts leave a program before receiving any dose of the drug, it is unlikely that any treatment effect exists for them. However, if the dropouts receive some of the drug before dropping out of the program, then their mean treatment effect is likely to be nonzero and assumptions (6) and (12) will be violated.

It is clarifying to extend the previous framework and introduce three latent random outcome variables for each person. These correspond to the potential outcome variables central to the Roy (1951) model, the switching regression model of Quandt (1988), or to models of discrete choice (see, e.g., McFadden (1981)). In this new notation, Y_p is the outcome a person receives if he or she receives full treatment; Y_d is the outcome a person receives if he or she is randomized into the treatment group but then drops out before receiving full treatment; and Y_c is the outcome a person receives if he or she is randomized out of the program. The econometrician observes only one of these three random variables for each person: Y_c if the person is randomized out; Y_p if the person receives full treatment; and Y_d if the person is randomized into the treatment group but drops out.

Suppose that receipt of full treatment alters the base state outcome—what the person receives if he or she is randomized into the control group—by an amount α, whereas partial treatment alters the base state outcome by an amount $\nu\alpha$, where ν is an adjustment factor, and where both α and ν may be random variables.[9] Then we may write

$$Y_d = Y_c + \nu\alpha \tag{17}$$

$$Y_p = Y_c + \alpha \tag{18}$$

$$Y_t = dY_d + (1-d)Y_p. \tag{19}$$

We do not assume that participants drop out of the program at random; d may be arbitrarily correlated with Y_c, α, and ν. In this notation,

$$E(Y_t) = kE(Y_d|d=1) + (1-k)E(Y_p|d=0) \tag{20}$$

and

$$\Delta_p = E(\alpha|d=0). \tag{21}$$

The impact of full treatment on the fully treated is just the expected value of the treatment effect α for persons who do *not* drop out. Substituting into equation (10) and simplifying, we obtain

$$\text{plim}\,(\overline{\Delta}_p) = \Delta_p + \frac{k}{1-k}E(\alpha\nu|d=1). \tag{22}$$

If k is positive, then the estimator $\overline{\Delta}_p$ is consistent for the parameter Δ_p if and only if $E(\alpha\nu|d=1) = 0$. Thus if there is no partial effect of treatment for the dropouts ($\nu \equiv 0$ for almost everyone), $\overline{\Delta}_p$ is consistent for Δ_p. Even in the presence of some partial treatment it is possible that the equality $E(\alpha\nu|d=1) = 0$ could occur by coincidence. However, it would be fortuitous if this equality held and $\nu \neq 0$ for almost everyone. Note that if $\nu \equiv 0$, assumption (12) is valid.

B. *Parameter of Interest in the Presence of Partial Treatment*

While much of the literature on program evaluation focuses exclusively on Δ_p as the parameter of interest, it is not always true that it is the most economically interesting summary measure of the effect of a program on its participants.[10] The parameter Δ_p measures the impact of the program only for those who do not drop out, yet programs often affect the dropouts as well as those who receive the full program treatment. When dropouts are also affected by the program, Δ_p may not be the primary parameter of interest, and the commonly used estimator $\overline{\Delta}_p$ is an inconsistent estimator of Δ_p. We now consider two examples that demonstrate these points.

[8] This problem is common in hypothesis testing. Acceptance of an implication of a model does not imply acceptance of the full model.

[9] One can think of ν as being bounded between 0 and 1, but this is not necessary.

[10] Heckman and Robb (1985), Moffitt (1992), Heckman (1992), Heckman and Smith (1993, 1995, 1998), and Heckman et al. (1997) discuss a variety of alternative parameters of interest in evaluating social programs.

Example 1: In our first example we suppose that α, the effect of the program, is a pure stigma effect. Anyone randomized into the treatment group and therefore associated with the program, even the dropouts, gets an outcome drawn from a common distribution that reflects this stigma. Thus $Y_p = Y_d$, so that $\nu = 1$. In this case, assumption (6) is violated. If the program stigmatizes its participants, then presumably $E(Y_d|d = 1) < E(Y_c|d = 1)$ and $E(Y_p|d = 0) < E(Y_c|d = 0)$, from which it follows that both $\Delta_p < 0$ and $\Delta < 0$.

The stigmatization experienced by the nondropouts may be more or less than the stigmatization experienced by the dropouts. There is no particular reason why the expected stigma effect among the nondropouts, given by Δ_p, should be of greater economic interest than the expected effect of stigma among the dropouts, given by

$$\Delta_d = E(Y_d|d = 1) - E(Y_c|d = 1)$$

or than the expected effect of stigma on anyone randomized into the treatment group, given by Δ.

In the stigma case, $\overline{\Delta}_p$ is an inconsistent estimator of Δ_p. Under random sampling,

$$\text{plim}\,(\overline{\Delta}_p) = \Delta_p + \frac{k}{1-k}[E(Y_d|d = 1) - E(Y_c|d = 1)] = \frac{\Delta}{1-k}.$$

Thus $\overline{\Delta}_p$ overstates the stigma effect in absolute value and is downward inconsistent for Δ_p if stigma has a negative effect on the dropouts.

Example 2: In this example Δ_p is an economically interesting parameter but provides only a partial description of the full impact of the program on participants. Consider a job subsidy program, and assume for simplicity that the persons to be offered the subsidy are all initially not employed. The program operates by offering participants a subsidized job with a wage $Y_p = y_p$ drawn from distribution F_p. Both the experimental controls and the experimental treatment group members have access to a common unsubsidized job market, from which they are assumed to receive a wage offer $Y_c = y_c$ drawn from distribution F_c. Treatment group members receive their unsubsidized offer after learning the value of their subsidized offer, but prior to deciding whether or not to accept it. The two wage offers Y_p and Y_c need not be statistically independent of each other.

This program confers an option value on each person in the treatment group. Persons in the treatment group who turn down the subsidized job offer (and thereby drop out of the program) because they receive a better unsubsidized offer do not exercise the option provided by the program. Persons randomized into the treatment group have expected wages of $E(\max\{Y_p, Y_c\})$, while those randomized into the control group have expected wages of $E(Y_c)$.

In this example,

$$\Delta_p = E(Y_p|Y_p > Y_c) - E(Y_c|Y_p > Y_c)$$

measures that part of the "treatment" attributable to those whose unsubsidized wage offers do not exceed their subsidized wage offer. The proportion of treatment group members for whom the unsubsidized wage exceeds the subsidized wage offered by the program is given by k, where

$$k = \Pr(Y_c > Y_p).$$

In this example the program confers an option on everyone in the treatment group. The option value is

$$\Delta = E\max\{Y_p, Y_c\} - E(Y_c).$$

Under random sampling, this parameter is consistently estimated by $\overline{\Delta}$. At the same time, $\overline{\Delta}_p$ consistently estimates Δ_p because the treatment group dropouts receive the same outcome, in expected value terms, as their analogs in the control group, namely, their unsubsidized wage offer. However, in the case described in this example, Δ_p provides only an incomplete characterization of the effect of the program. For a complete picture, both Δ and Δ_p are required.

VI. Identification in the Presence of Partial Treatment

For many econometric selection models it has been shown that the use of exclusion restrictions, or what are sometimes inappropriately termed "instruments," is sometimes helpful for the identification of certain parameters of interest (see, e.g., Heckman and Robb (1985), Heckman (1990a), Imbens and Angrist (1994), Heckman (1997), and Heckman and Smith (1996, 1998)). In this section we explore whether this approach is useful for the identification of Δ_p when assumption (6) fails to hold.

Here we focus on two parameters that are often considered in conducting evaluations. We have devoted most of our attention in this paper to Δ_p, defined in equation (5) as

$$\Delta_p \equiv E(Y_p|d = 0) - E(Y_c|d = 0).$$

Recall that in application to the control group, $d = 0$ indicates an individual who would not drop out of the program if he or she were randomized into the treatment group. Thus Δ_p is the expected effect of participating in the program among those who participate or, more precisely, the effect of full treatment on the fully treated. We contrast this with the parameter

$$\tilde{\Delta}_p \equiv E(Y_p) - E(Y_c). \tag{23}$$

This is the effect of the program on participants if nobody drops out.[11] The two parameters define different counterfactuals and answer different policy questions. Both are versions of "treatment on the treated," where treatment means "full treatment" in both cases but where the set of persons over whom the expected value is computed differs between Δ_p and $\tilde{\Delta}_p$. For Δ_p the conditioning set is the fully treated; for $\tilde{\Delta}_p$ it is the entire treatment group.

We define variable Z subject to an exclusion restriction as some random variable with support $\mathscr{Z}$ that influences the decision to drop out, but does not influence the outcomes directly. A standard definition of Z requires it to satisfy the following conditions:

$$E(Y_p|Z) = E(Y_p)$$
$$E(Y_d|Z) = E(Y_d) \tag{24}$$
$$E(Y_c|Z) = E(Y_c)$$

and

$$\Pr(d = 1|Z = z_1) \neq \Pr(d = 1|Z = z_2)$$

for some (z_1, z_2) such that $z_1 \neq z_2$ (see Heckman (1997)).

From the data commonly available in a social experiment, it is possible to identify $E(Y_p|d = 0)$, the mean outcome of the nondropouts in the treatment group, $E(Y_d|d = 1)$, the mean outcome of the treatment group dropouts, and $E(Y_c)$, the mean outcome of the controls. The missing piece of information required to identify Δ_p is $E(Y_c|d = 0)$, the mean outcome of the controls who would have been participants, had they been in the treatment group. The missing piece of information required to identify $\tilde{\Delta}_p$ is $E(Y_p|d = 1)$, the expected outcome conditional on full treatment of the treatment group dropouts.

First consider how to use assumption (24) to identify $\tilde{\Delta}_p$. Clearly $E(Y_c)$ is identified from the control group alone. Identification of $E(Y_p)$ requires the same type of conditions required to identify a selection model. (See Amemiya (1985) for numerous examples of sample selection correction procedures and Heckman (1990a) for a survey of some nonparametric estimators for selection models.)

A common practice is to use $E(Y_p|d = 0)$ as a proxy for $E(Y_p|d = 1)$ or $E(Y_p)$ in an attempt to secure the identification of $\tilde{\Delta}_p$. The general problem here is that of using a truncated sample to estimate the mean of a full distribution. As demonstrated in Heckman (1990a) or Heckman and Honoré (1990), access to Z can sometimes aid in identifying $E(Y_p|d = 1)$ in this context. Take, for example, an "index structure" representation,

$$E(Y_p|d = 0, Z) = E(Y_p|d = 0, p(Z))$$

where $p(Z) = \Pr(d = 0|Z)$. With sufficient variation in Z, one can recover $E(Y_p)$ in the limit as

$$\lim_{p(Z)\to 1} E(Y_p|p(Z)) = E(Y_p).$$

Alternatively, as noted in Heckman (1990a,b), if there is a value z^* of Z such that $\Pr(d = 1|Z = z^*) = 0$, then

$$E(Y_p|Z = z^*) = E(Y_p)$$

and the counterfactual required to identify $\tilde{\Delta}_p$ can be obtained without having to assume an index structure. A more traditional approach uses the exclusion restriction in conjunction with parametric selection bias correct methods to estimate $E(Y_p)$. An example of this approach is given later in this section.

Identifying Δ_p is more challenging because it requires breaking $E(Y_c)$ into its two components $E(Y_c|d = 1)$ and $E(Y_c|d = 0)$. Traditional selection bias methods are not informative in this case. However, a limit argument similar to the one used above to identify $\tilde{\Delta}_p$ can be used to identify Δ_p over some range of the data. If there is some limiting value of $Z = z^{**}$ such that

$$\lim_{Z\to z^{**}} \Pr(d = 0|Z) = 1$$

then we can construct $E(Y_c|d = 0,\ Z = z^{**}) = E(Y_c|Z = z^{**})$. Then for $Z = z^{**}$ we can construct

$$\Delta_p(Z = z^{**}) = E(Y_p|d = 0, Z = z^{**}) - E(Y_c|d = 0, Z = z^{**}). \tag{25}$$

Conditional on $Z = z^{**}$, Δ_p and $\tilde{\Delta}_p$ are the same. However, the expression in equation (25) cannot be identified for all values of Z. It is only identified on the set of values for which $Z = z^{**}$. We now present an example that helps to clarify this discussion, and which shows that Δ_p is not identified even in a model with strong functional form assumptions.

Example 1: Restrict the model to the following special case:

$$Y_c = \mu_c + \epsilon_c$$
$$Y_p = \mu_p + \epsilon_p$$
$$Y_d = \mu_d + \epsilon_d$$
$$d = 1(Z\gamma + \nu \geq 0)$$

where $1(\cdot)$ is the indicator function, which takes on the value 1 if its argument is true and 0 if it is false. We assume $(\epsilon_c, \epsilon_p, \epsilon_d, \nu)$ is a mean zero normal random vector that is statistically independent of the scalar random variable Z. In

[11] Recall that we continue to implicitly condition on application to and acceptance into the program.

addition, we normalize the variance of ν to 1. In this case we can write

$$E(Y_c|d = 0, Z) = E(Y_c|Z, Z\gamma + \nu < 0) = \mu_c + \sigma_{cv}\lambda(-Z\gamma) \quad (26)$$

where σ_{cv} is the covariance between ν and ϵ_c, and $\lambda(\cdot)$ is the inverse Mills ratio. This is the traditional sample selection model of Heckman (1976). We do not observe d for the control group. As a result, even with access to Z we can never hope to identify σ_{cv} without further assumptions. Thus we cannot identify

$$\Delta_p = \mu_p - \mu_c + (\sigma_{pv} - \sigma_{cv})\lambda(-Z\gamma) \quad (27)$$

where σ_{pv} is the covariance between ν and ϵ_p. However, note that we can identify $\mu_c = E(Y_c)$ and we can also identify $\mu_p = E(Y_p)$, so the parameter $\Delta_p = \mu_p - \mu_c$ is identified.

There is a conditioning assumption which delivers identification of Δ_p, although it only holds in very special cases:

$$E(Y_c|Z, d) = E(Y_c|d) \quad (28)$$

for some $z_1 \in \mathscr{Z}$ and $z_2 \in \mathscr{Z}$, such that

$$\Pr(d = 1|Z = z_1) \neq \Pr(d = 1|Z = z_2).$$

Assumption (28) requires only that the variable Z take on two distinct values, while much of the literature requires that the support of Z be the real line $\mathscr{R}_1$, or that there be some value of Z for which $\Pr(d = 0|Z) = 1$. Neither of these is required for identification if we are willing to make assumption (28).[12]

We first show that this assumption is sufficient for the identification of Δ_p. Suppose assumption (28) holds for some particular values z_1 and z_2. We define $P_1 = \Pr(d = 1|Z = z_1)$ and $P_2 \equiv \Pr(d = 1|Z = z_2)$, where $P_1 \neq P_2$. Then it is easy to show that

$$E(Y_c|d = 0) = \frac{P_2E(Y_c|Z = z_1) - P_1E(Y_c|Z = z_2)}{P_2 - P_1}. \quad (29)$$

Using equation (29) we can obtain $\Delta_p = E(Y_p|d = 0) - E(Y_c|d = 0)$.

A major interpretive issue is whether assumption (28) is plausible. Except in very special cases, it is not. This assumption requires that Y_c be dependent on Z, but only through the seemingly irrelevant event corresponding to $d = 0$. This condition is not satisfied in standard discrete choice models. We present an example in which this condition holds, and demonstrate the sensitivity of the identifying power of assumption (28) to small perturbations in the assumption set.

Example 2: Let

$$Z = \alpha_z\theta + U(Z) \quad (30)$$

$$Y_c = \alpha_y\theta + U(Y_c) \quad (31)$$

$$d = \theta \quad (32)$$

where $\theta \in \{0, 1\}$ is a binary random variable and where $(U(Z), U(Y_c))$ are mutually independent and are also independent of θ. Heuristically, persons with $\theta = 1$ are unmotivated while those with $\theta = 0$ are motivated. Motivated persons do not drop out ($d = 0$ for $\theta = 0$). Let Z be ability. Then if $\alpha_z < 0$, more motivated persons have higher ability. If $\alpha_y < 0$, then more motivated persons have higher income. This is an instance of a signaling model in which d perfectly signals θ.

In this example it is clear that

$$E(Y_c|d, Z) = E(Y_c|d) \quad (33)$$

so assumption (28) is satisfied. Dropout status is a perfect predictor of θ, which drives the correlation among all three random variables.

However, if equation (32) is modified slightly to add nondegenerate random variable $U(d)$ so that

$$d = \theta + U(d) \quad (34)$$

where $E(U(d)) = 0$, $U(d)$ is conditionally independent of θ, and where $U(Z)$, $U(Y_c)$, and $U(d)$ are mutually independent, then assumption (28) is violated. Adding a little "noise" to θ in the form of $U(d)$ in determining d makes Z a useful predictor of Y_c given d. In this setting, assumption (28) is a very fragile assumption.

VII. Analyzing the National JTPA Study

A. *JTPA and Partial Treatment*

In this section we apply our analysis to data from the National JTPA Study (NJS). This is a recent experimental evaluation of the employment and training programs financed under the JTPA. This job training program provides basic education, classroom training in occupational skills, subsidized on-the-job training at private firms, job search assistance, and other employment and training services to economically disadvantaged persons. The program includes a performance standards system in which locally managed training centers compete for incentive payments based on their success at placing enrollees in steady, high-paying jobs.

In the NJS, random assignment occurred prior to formal enrollment in the program. A substantial fraction of those randomly assigned to the experimental treatment group

[12] Note that assumption (28) is the polar opposite of the assumption traditionally made in matching, which is that $E(Y_c|Z, d) = E(Y_c|Z)$.

TABLE 1.—SAMPLE MEAN EARNINGS, k, AND MEAN DIFFERENCE IMPACT ESTIMATES (FULL ABT 18-MONTH IMPACT SAMPLE)

Target Group, Outcome	Adult Men	Adult Women[a]	Male Youth	Female Youth
Treatment group, $E(Y_t)$	13096.43	8261.81	9997.76	6163.67
	(210.78)	(132.31)	(253.98)	(163.33)
Full participants, $E(Y_p)$	13638.23	8424.97	10274.77	6114.49
	(260.96)	(155.89)	(310.36)	(196.76)
Less than full, $E(Y_d)$	12181.06	7946.47	9442.33	6256.13
	(354.80)	(244.35)	(441.13)	(290.81)
Controls, $E(Y_c)$	12530.09	7470.98	10781.72	6202.09
	(305.57)	(180.20)	(401.80)	(248.77)
Fraction dropping out	0.3718	0.3410	0.3328	0.3472
Impact estimate	566.34	790.83	−783.96	−38.42
	(371.21)	(223.56)	(475.34)	(297.60)
Sample size	4420	5724	1747	2301

Notes: Estimated standard errors are in parentheses.
[a] Adult female nonrespondents not included.

never enrolled in the program. As shown in table 1, over 37% of adult male (ages 22 and older) treatment group members did not enroll in JTPA during the 18 months following random assignment. Similar nonenrollment rates are reported for the other three target groups in the NJS: adult females (ages 22 and older), male out-of-school youth (ages 16 to 21), and female out-of-school youth (ages 16 to 21). These nonenrollment rates are used to generate estimates of Δ_p using estimator (10).

This high rate of nonenrollment results in part from the time lag that often occurs between random assignment and the initiation of training. For courses given on an academic schedule, the applicant must wait until the beginning of the next quarter or semester. During this waiting period, the applicant may find a job or lose interest, and so may fail to enroll. However, as long as the training and job-seeking activities of potential trainees randomized in are the same as those of potential trainees of the same type randomized out, the instrumental-variable estimate of Δ_p remains valid.

A potentially serious problem may arise from the mechanics of the JTPA performance standards system. Incentive payments to the training centers under the JTPA performance standards system depend only on the performance of their *enrollees*. At the same time, enrollment in JTPA is very flexible, so that training centers can often delay enrolling someone until it is clear that the person is likely to succeed in training. For trainees assigned to job search assistance or to on-the-job training, this can mean that enrollment occurs when they find a job or an employer willing to provide them with on-the-job training. Those not enrolled for this reason are counted as dropouts, even though they often receive assistance in looking for jobs, writing resumes, and presenting themselves in interviews and may acquire information about the local labor market not available to persons randomized out. These activities would likely increase their future earnings even if they do not find a job or an on-the-job training slot during the period of their contact with the JTPA program.

Table 2 shows the extent of JTPA contact following random assignment among a subset of the treatment group nonenrollees. Over half of this subset of nonenrollees received some JTPA services. This evidence suggests that assumption (6) may be inappropriate, as those in the control group who would have been dropouts had they been in the treatment group did not receive these services. Many of these JTPA dropouts are not "no shows" who receive no treatment from the program.

In this experiment it is very hard to justify why Δ_p should be *the* parameter of interest (which it is defined to be by the analysts in the NJS; see Orr et al. (1995)). Over half of the people who dropped out actually received services. Clearly the impact of the program on these people is of some

TABLE 2.—PERCENTAGE DISTRIBUTION OF POST-RANDOM-ASSIGNMENT ACTIVITY IN JTPA OF TREATMENT GROUP MEMBERS WHO DID NOT ENROLL

Activity	Nonenrollees (%)
No further contact	15
Further contact, but not eligible	1
No longer interested[a]	11
Got job on own	5
Moved	2
Health problems	1
In another program	1
Reason unknown	3
Interested, but made contact only and received no services	20
Interested and received service(s)[b]	53
Received further assessment and counseling	11
Referred to classroom training provider(s)	5
Received support service(s)	2
Referred to employer(s) for possible on-the-job training	36
Participated in job club or received job search assistance	20
Total	100
Sample size	307

Notes: Calculations are based on data for a random sample of 307 treatment group members in the 18-month study sample who did not enroll in JTPA.
Source: Kemple et al. (1993).
[a] When totaled, subcategory percentages are over 11% because nonenrollees could cite more than one reason for no longer being interested in JTPA.
[b] When totaled, subcategory percentages are over 53% because some nonenrollees received more than one service.

interest. On the other hand, it is hard to argue that Δ_p is of *no* interest. It informs us of the impact of the program on the group of people who were actually enrolled in it.

B. *Sensitivity of Impact Estimates to Violations of Assumption (6)*

We explore the sensitivity of the experimental impact estimates to departures from assumption (6) in two ways. Our first analysis considers the sensitivity of estimates of Δ_p to assumptions about the impact of partial treatment on the nonenrollees. Suppose that the difference in expected outcomes between the treatment group dropouts and their analogs in the control group is given by $\epsilon = E(Y_d|d = 1) - E(Y_c|d = 1)$. Then a simple modification of the derivation leading up to equation (10) produces an adjusted estimator of Δ_p,

$$\overline{\Delta}_p^{\epsilon} = \overline{\Delta}_p - \frac{k}{1-k}\epsilon. \tag{35}$$

Table 3 displays $\overline{\Delta}_p^{\epsilon}$ calculated with ϵ equal to \$150, \$100, \$50, −\$50, −\$100, and −\$150. As in table 1, the outcome variable Y corresponds to the sum of self-reported earnings in the 18 months after random assignment. The estimated means and sample sizes underlying the values in table 3 appear in table 1. As noted above, in the context of the JTPA evaluation, the most likely source of bias is receipt of partial treatment by some of the treatment group dropouts. Suppose that the mean impact of these partial services on earnings in the 18 months after random assignment is \$300, which is quite reasonable given the annual impact findings for job search assistance reported in Gueron and Pauly (1991). As the evidence in table 2 indicates that roughly half of the dropouts received some partial treatment, this corresponds to a value of ϵ equal to \$150. For adult males, ϵ = \$150 implies an adjusted earnings impact estimate of $\overline{\Delta}_p^{\epsilon}$ = \$990.33. The figure in square brackets indicates that this estimate is \$88.78, or around 10%, lower than the estimate given by the instrumental-variable estimator.

The second analysis examines the sensitivity of the estimates of Δ_p to assumptions about the relative magnitudes of $E(Y_c|d = 1)$ and $E(Y_c|d = 0)$. Making an assumption about the relative magnitudes of these two conditional expectations represents an alternative way to identify $E(Y_c|d = 0)$. The general form of this identifying assumption is

$$E(Y_c|d = 1) = \eta E(Y_c|d = 0) \tag{36}$$

where η is the constant of proportionality.

Substituting equation (36) into equation (4) and solving for $E(Y_c|d = 0)$ yields

$$E(Y_c|d = 0) = \frac{E(Y_c)}{1 + k(\eta - 1)}. \tag{37}$$

TABLE 3.—SENSITIVITY OF ESTIMATES OF IMPACT OF TREATMENT ON TREATED USING THE INSTRUMENTAL VARIABLE ESTIMATOR TO ASSUMPTIONS ABOUT IMPACT OF PARTIAL TREATMENT ON EARNINGS OF NONENROLLEES (FULL ABT 18-MONTH IMPACT SAMPLE)

Target Group	Adult Men	Adult Women	Male Youth	Female Youth
Bloom estimate $\overline{\Delta}_p$[a]	901.55 (590.60)	1200.00 (339.19)	−1174.96 (712.30)	−58.85 (455.92)
Δ_p^{ϵ} with ϵ = \$150	990.33 [88.78]	1277.61 [77.61]	−1100.15 [74.81]	20.92 [79.78]
Δ_p^{ϵ} with ϵ = \$100	960.74 [59.19]	1251.74 [51.74]	−1125.08 [49.87]	−5.67 [53.19]
Δ_p^{ϵ} with ϵ = \$50	931.14 [29.59]	1225.87 [25.87]	−1150.02 [24.94]	−32.26 [26.59]
Δ_p^{ϵ} with ϵ = −\$50	871.95 [−29.59]	1174.13 [−25.87]	−1199.90 [−24.94]	−85.45 [−26.59]
Δ_p^{ϵ} with ϵ = −\$100	842.36 [−59.19]	1148.26 [−51.74]	−1224.83 [−49.87]	−112.04 [−53.19]
Δ_p^{ϵ} with ϵ = −\$150	812.77 [−88.78]	1122.39 [−77.61]	−1249.77 [−74.81]	−138.63 [−79.78]

Notes: Estimated standard errors are in parentheses; bias is in brackets.
Since ϵ is a constant, standard errors for Δ_p^{ϵ} are equal to standard errors for Δ_p and are not repeated.
[a] Estimates presented here differ from those in Bloom et al. (1993) because (1) estimates are calculated using simple means without regression adjustment and (2) imputed values for adult female nonrespondents based on UI earnings data are not used.

The value on the left-hand side decreases as η increases so that the overall mean $E(Y_c)$ remains the same. Under these assumptions we can write

$$\begin{aligned}\overline{\Delta}_p^{\eta} &= \overline{Y}_t(d = 0) - \overline{Y}_c(d = 0) \\ &= \overline{Y}_t(d = 0) - \left[\frac{\overline{Y}_c(d = 0)}{1 + k(\eta - 1)}\right].\end{aligned} \tag{38}$$

We can provide information about the sensitivity of the estimates to the choice of η by calculating $\overline{\Delta}_p^{\eta}$ for different values of η. Hotz and Sanders (1994) present another way of conducting sensitivity analyses in a more structured setting.

Table 4 gives estimates of Δ_p^{η} constructed using data on self-reported earnings in the 18 months after random assignment from the NJS for various choices of η. The first row repeats the estimates of Δ_p obtained using the instrumental-variable estimator. The next five rows present estimates of Δ_p^{η} for values of η equal to 0.50, 0.75, 1.00, 1.25, and 1.50. The final row displays the value of η that equates $\overline{\Delta}_p^{\eta}$ and $\overline{\Delta}_p$ for each target group. The table shows that for $\eta \in [0.5, 1.5]$, varying the selection process into the dropout group by varying the value of η strongly influences the resulting impact estimates $\overline{\Delta}_p^{\eta}$. Setting η = 0.50, which implies that those controls who would have been dropouts have only half the mean earnings of those who would have been participants, produces strongly negative impact estimates in all cases. In contrast, for η = 1.5, which implies a lower mean outcome for those who would have received the treatment than for those who would have dropped out, very large impact estimates are produced in all cases. The estimates for η = 1.0, which corresponds to the case of

TABLE 4.—ESTIMATED IMPACTS ON EARNINGS IN THE 18 MONTHS AFTER RANDOM ASSIGNMENT ASSUMING $E(Y_c|d = 1) = \eta E(Y_c|d = 0)$ (FULL ABT 18-MONTH IMPACT SAMPLE)

Target Group	Adult Men	Adult Women	Male Youth	Female Youth
IV estimate Δ_p[a]	901.55	1200.00	−1174.96	−58.85
	(590.60)	(339.19)	(712.30)	(455.92)
Δ_p^η with $\eta = 0.50$	−1753.23	−581.53	−2658.96	−1390.44
	(457.15)	(267.38)	(573.28)	(359.63)
Δ_p^η with $\eta = 0.75$	−175.93	257.77	−1485.31	−677.10
	(426.13)	(251.21)	(537.03)	(336.04)
Δ_p^η with $\eta = 1.00$	1108.13	953.99	−506.94	−87.60
	(401.84)	(238.27)	(507.71)	(317.18)
Δ_p^η with $\eta = 1.25$	2173.79	1540.82	321.14	407.74
	(382.45)	(227.76)	(483.65)	(301.85)
Δ_p^η with $\eta = 1.50$	3072.39	2042.18	1031.09	829.81
	(366.73)	(219.10)	(463.67)	(289.22)
Values of η such that $\Delta_p^\eta = \Delta_p$	0.9564	1.0999	0.8247	1.0134

Notes: Estimated standard errors are in parentheses.

[a] Estimates presented here differ from those in Bloom et al. (1993) because (1) estimates are calculated using simple means without regression adjustment and (2) imputed values for adult female nonrespondents based on UI earnings data are not used.

random dropping out, come close to those from the Bloom estimator for three of the four target groups, with male youth the exception. The last row of table 4 provides additional evidence on this point, as the value of η that equates the two estimates lies close to 1.0 for all groups other than male youth.

C. *Testing Assumption (12) in the NJS*

The testing strategy used here begins with simple tests of differences between the outcome distributions of the control group and the dropouts from the treatment group. Under assumption (12), the control and dropout outcome distributions will be equivalent if the factors causing persons to drop out are unrelated to their outcomes. If $F_c(y) = F_d(y|d = 1)$, then restriction (14) will hold. In practice, when these two oucome distributions are not statistically distinguishable, the implied outcome distribution for the participant analogs in the control group is extremely unlikely to violate any of the restrictions tested by the more complicated tests proposed in the appendix. Thus, in general, when the initial battery of tests fails to reject the equivalence of the control and treatment group dropout outcome distributions, the additional tests proposed in this paper are superfluous. In results not reported here we do not reject the hypothesis of equality of the two distributions for any demographic group using both Kolmogorov–Smirnov and Wilcoxin tests. As a result, we do not pursue this issue further with the data from the NJS.[13]

[13] See Heckman et al. (1994) for detailed results of these tests and the tests proposed in the appendix and applied to the NJS data.

VIII. Conclusion

This paper examines several aspects of an instrumental-variable estimator commonly used to produce estimates of Δ_p, the impact of full treatment on the fully treated, in the context of experimental evaluations in which not all treatment group members receive treatment. We present and discuss the key assumption that justifies this estimator and argue that it is not likely to hold in the commonly encountered case where the dropouts receive a partial "dose" of the treatment prior to dropping out.

When the identifying assumption underlying the model fails to hold, the instrumental-variable estimator considered in this paper does not provide consistent estimates of the parameter Δ_p. For example, if dropouts receive partial treatment, the instrumental-variable estimator discussed in this paper produces inconsistent estimates of Δ_p. We develop statistical tests of the assumption underlying the estimator.

This paper makes the general point that the parameter Δ_p, the effect of full treatment on the fully treated, may not be the main parameter of economic interest in evaluating a social program, especially in situations where the assumptions justifying the instrumental-variable estimator do not hold. There are many parameters of economic interest other than Δ_p. We present examples where the unadjusted mean impact of assignment to treatment, Δ, may be of greater interest than Δ_p.

We also discuss the role of identifying assumptions based on exclusion restrictions. In general, different exclusion restrictions identify different parameters, a point that has often generated confusion in the literature on program evaluation. Identification of Δ_p based on exclusion restrictions is a delicate operation that is not robust to small perturbations in the assumptions.

Applying our statistical tests to the data from the recent NJS, we do not reject the identifying assumption underlying the widely used instrumental-variable estimator, despite substantial receipt of partial treatment by dropouts in the experimental treatment group. At the same time, sensitivity analyses conducted using the JTPA data reveal that the empirical consequences of the failure of the key identifying assumption can be quite substantial.

REFERENCES

Amemiya, Takeshi, *Advanced Econometrics* (Cambridge, MA: Harvard University Press, 1985).

Angrist, Joshua, and Guido Imbens, "Sources of Identifying Information in Evaluation Models," unpublished manuscript, Harvard University (1991).

Bloom, Howard, "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review* 8:2 (1984), 225–246.

Bloom, Howard, Larry Orr, George Cave, Steve Bell, and Fred Doolittle, *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months* (Bethesda, MD: Abt Associates, 1993).

Dubin, Jeffrey, and Douglas Rivers, "Experimental Estimates of the Impact of Wage Subsidies," *Journal of Econometrics* 56 (1993), 219–242.

Efron, Bradley, and David Feldman, "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association* 86 (1991), 9–17.

Gueron, Judith, and Mark Pauly, *From Welfare to Work* (New York: Russell Sage Foundation, 1991).

Heckman, James, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 15 (1976), 475–492.

——— "Varieties of Selection Bias," *American Economic Review* 80:2 (1990a), 313–318.

——— "Alternative Approaches to the Evaluation of Social Programs," Fifth World Congress of the Econometric Society, Barcelona Lecture (1990b).

——— "Randomization and Social Policy Evaluation," in Charles Manski and Irwin Garfinkel (eds.), *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press, 1992).

——— "Instrumental Variables: A Study of the Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32:3 (1997), 441–462.

Heckman, James, and Bo Honoré, "The Empirical Content of the Roy Model," *Econometrica* 58:5 (1990), 1121–1150.

Heckman, James, and Richard Robb, "Alternative Methods for Evaluating the Impact of Treatment on Outcomes," in James Heckman and Burton Singer (eds.), *Longitudinal Analysis of Labor Market Data* (Cambridge, UK: Cambridge University Press, 1985).

Heckman, James, and Jeffrey Smith, "Assessing the Case for Randomized Evaluation of Social Programs," in Karsten Jensen and Per Kongshoj Madsen (eds.), *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies* (Copenhagen, Denmark: Ministry of Labour, 1993).

——— "Evaluating the Welfare State," written in 1995, forthcoming in Steinar Strom (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial* (Econometric Society Monograph Series, Cambridge University Press, 1998).

——— "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9:2 (1995b), 85–110.

——— "Experimental and Non-Experimental Evaluation," in Günter Schmid, Jacqueline O'Reilly, and Klaus Schömann (eds.), *International Handbook of Labour Market Policy and Evaluation* (Cheltenham, UK: Edward Elgar, 1996).

Heckman, James, Jeffrey Smith, and Nancy Clements, "Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts," *Review of Economic Studies* 218 (Oct. 1997).

Heckman, James, Jeffrey Smith, and Christopher Taber, "Accounting for Dropouts in Evaluations of Social Experiments," Technical Working Paper 166, NBER (1994).

Hotz, V. Joseph, and Seth Sanders, "Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice," Working Paper 94-2, Population Research Center, University of Chicago (1994).

Imbens, Guido, and Joshua Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62:2 (1994), 467–476.

Kemple, James, Fred Doolittle, and John Wallace, "The National JTPA Study: Site Characteristics and Participation Patterns," (New York: Manpower Demonstration Research Corporation, 1993).

Mallar, Charles, Stuart Kerachsky, and Craig Thorton, "The Short-Term Economic Impact of the Job Corps Program," in Ernst Stromsdorfer and G. Farkas (eds.), *Evaluation Studies Review Annual,* vol. 5 (Beverly Hills, CA: Sage Publications, 1980).

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in Charles Manski and Daniel McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: MIT Press, 1981), 198–272.

Moffitt, Robert, "Evaluation Methods for Program Entry Effects," in Charles Manski and Irwin Garfinkel (eds.), *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press, 1992).

Orr, Larry, Steve Bell, Winston Lin, George Cave, and Fred Doolittle, *The National JTPA Study: Impacts, Benefits and Costs of Title IIA* (Bethesda, MD: Abt Associates, 1995).

Perlman, Michael, "One-Sided Testing Problems in Multivariate Analysis," *Annals of Mathematical Statistics* 40 (1969), 549–567.

Quandt, Richard, *The Econometrics of Disequilibrium* (New York: Basil Blackwell, 1988).

Roy, Andrew, "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers* 3 (1951), 135–146.

Smith, David, Jane Kulik, and Ernst Stromsdorfer, "The Economic Impact of the Downriver Community Conference Economic Readjustment Activity Program: Choosing between Retraining and Job Search Placement Strategies," in Kevin Hollenbeck, Frank Pratzner, and Howard Rosen (eds.), *Displaced Workers: Implication for Educational and Training Institutions* (Columbus, OH: National Center for Research in Vocational Education, 1984).

Wolak, Frank, "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models," *Econometrica* 59:4 (1991), 981–995.

APPENDIX

Testing Strategy

The intuition for our proposed test is obtained by rewriting restriction (14) in the text in a different but equivalent form,

$$\frac{\Pr(Y_c \in b) - k\Pr(Y_d \in b)}{1-k} \geq 0 \qquad \text{for all } b \subset \mathscr{B} \tag{14'}$$

where $\mathscr{B}$ is the class of subsets of the union of the supports of Y_c and Y_d. Conditioning on $d = 1$ is kept implicit in defining $\Pr(Y_d \in B)$. Unless $\mathscr{B}$ is finite, testing restriction (14′) for every element of $\mathscr{B}$ poses a difficult statistical problem. However, it is straightforward to test restriction (14′) using a finite number of subsets of $\mathscr{B}$. In the case where $\mathscr{B}$ is finite, this procedure completely exhausts the implications of restriction (14). In the case where $\mathscr{B}$ is infinite, we could in principle develop a more powerful test. However, the tests developed here are easy to implement in practice and have known asymptotic properties.

The null hypothesis we consider is

$$H_0: \quad \frac{1}{1-k}[\Pr(Y_c \in b_j) - k\Pr(Y_d \in b_j)] \geq 0$$
$$\text{for all } j = 1, \ldots, J.$$

For each j we can estimate $[1/(1-k)][\Pr(Y_c \in b_j) - k\Pr(Y_d \in b_j)]$ by using sample probabilities to estimate population probabilities. Define $\hat{P}$ to be the vector of these estimates and P the vector of true values. Letting n be the sample size, where we assume for simplicity that the treatment and control samples are of the same size, it is easy to show that under random sampling,

$$\sqrt{n}(\hat{P} - P) \xrightarrow{d} N(0, \Sigma) \tag{A.1}$$

where each element of the matrix Σ is of the form

$$\sigma_{ij} = \frac{\Pr(Y_c \in b_i \cup b_j) - \Pr(Y_c \in b_i)\Pr(Y_c \in b_j)}{(1-k)^2} + \frac{k^2[\Pr(Y_d \in b_i \cup b_j) - \Pr(Y_d \in b_i)\Pr(Y_d \in b_j)]}{(1-k)^2}. \tag{A.2}$$

For simplicity we assume that k is known. Modifying the covariance matrix to account for the estimation of k is straightforward (see Heckman et al. (1994)).[14]

[14] If $n_c/n_d = \varphi \neq 1$, then equation (A.1) becomes $\sqrt{n_c}(\hat{P} - P) \xrightarrow{d} N(0, \Sigma)$ and equation (A.2) becomes $\sigma_{ij} = \{[\Pr(Y_c \in b_i \cup b_j) - \Pr(Y_c \in b_i)\Pr(Y_c \in b_j)] + \varphi k^2[\Pr(Y_d \in b_i \cup b_j) - \Pr(Y_d \in b_i)\Pr(Y_d \in b_j)]\}/(1-k^2)$.

Testing H_0 requires that we define a test statistic $t(\hat{P})$ and derive the asymptotic distribution of this test statistic under the null hypothesis. The composite nature of the null hypothesis complicates the procedure. There is no similar test region in this case. That is, the size of the test will vary

As noted previously, the distribution of the test statistic varies across values of the parameters consistent with the null hypothesis. Using the analysis of Perlman (1969), it is straightforward to show that the probability of accepting the null hypothesis can be written as follows:[15]

$$\lim_{n\to\infty} \Pr(t > c) = \begin{cases} 1.00, & P_p^*(1) > 0, \quad P_p^*(2) > 0, \quad P_p^*(3) > 0 \\ \Phi(-c(1)), & P_p^*(1) = 0, \quad P_p^*(2) > 0, \quad P_p^*(3) > 0 \\ BVN(-c(1), -c(2); \rho_{12}), & P_p^*(1) = 0, \quad P_p^*(2) = 0, \quad P_p^*(3) > 0 \end{cases}$$

across alternative values of $(k, F_c, F_t(\cdot|d = 1))$ consistent with the null hypothesis. For any critical region we calculate the size of the test based on the least favorable distribution, which is the distribution consistent with the null hypothesis for which the probability of rejection is greatest. One test statistic is based on the Wald test. We can define

$$t(\hat{P}) = \inf_{M \geq 0} [(\hat{P} - M)'\Sigma(\hat{P} - M)].$$

Drawing on the analysis of Perlman (1969), Wolak (1991) derives the least favorable distribution for this testing problem and provides a partial characterization of it. He shows that under the least favorable distribution at least two of the restrictions given in restriction (14′) will bind, so that at least two of the test regions will be at the boundary of the parameter space.

Another possible test statistic is

$$t(\hat{P}) = \begin{bmatrix} \dfrac{\hat{P}_1}{\sqrt{\hat{\sigma}_{11}}} \\ \vdots \\ \dfrac{\hat{P}_J}{\sqrt{\hat{\sigma}_{JJ}}} \end{bmatrix}$$

where $\hat{P}_j = [1/(1-k)][\Pr(\widehat{Y_c} \in b_j) - k\Pr(\widehat{Y_d} \in b_j)]$ so that $t(\hat{P})$ is just the vector of t-statistics for each of the individual cells. We choose the critical region $c = (c(1), \ldots, c(J))$ so that we reject when any element of $t(\hat{P})$ is less than the corresponding element of c. In other words, we perform J t-tests and reject the null hypothesis whenever we reject the null hypothesis for any of the individual t-tests. Using this test statistic, it is easy to derive the least favorable distribution.

We derive the least favorable distribution for the simple case of a model with three cells which form a partition of the support of $Y_c \cup Y_d$. The extension to the case of J cells simply generalizes the same line of reasoning. Let $P_c(i)$ denote the probability that Y_c lies in cell b_i, let $P_d(i)$ denote the probability that Y_d lies in cell b_i, and let $\hat{P}_c(i)$ and $\hat{P}_d(i)$ be their sample analogues. In what follows we take k, the fraction of the treatment group that drops out, as fixed and known. Define

$$P_p^*(i) = \frac{P_c(i) - kP_d(i)}{1-k}$$

and

$$\hat{P}_p^*(i) = \frac{\hat{P}_c(i) - k\hat{P}_d(i)}{1-k}.$$

Now consider the following null hypothesis:

$$H_0: \quad P_p^*(1) \geq 0$$
$$P_p^*(2) \geq 0$$
$$P_p^*(3) \geq 0.$$

where Φ denotes the cdf of a univariate standard normal random variable, $BVN(a, b; c)$ denotes the cdf of a standardized bivariate normal distribution with upper limits a and b and correlation c, and ρ_{12} denotes the asymptotic correlation between $\hat{P}(1)$ and $\hat{P}(2)$.

The least favorable distribution dictates the size of the test. It corresponds to those parameter values for which the limit of $\Pr(t > c)$ is minimal. In the case considered in this paper, the least favorable distribution will occur when the constraint binds in two cells. The feasible value of ρ_{12} that minimizes $BVN(-c(1), -c(2); \rho_{12})$ turns out to be $\rho_{12} = -k$, as we now demonstrate.

For any critical region, the size of the test is

$$\alpha = 1 - BVN(-c(1), -c(2); -k).$$

A least favorable distribution can be derived by minimizing the correlation with respect to $(P_c(1), P_c(2), P_d(1), P_d(2))$ subject to the constraints

$$P_c(1) = kP_d(1)$$
$$P_c(2) = kP_d(2)$$
$$1 - P_d(1) - P_d(2) \geq 0.$$

The solution to this problem is as follows:

$$P_c = \begin{bmatrix} \dfrac{k}{2} \\ \dfrac{k}{2} \\ 1-k \end{bmatrix}, \qquad P_d = \begin{bmatrix} \dfrac{1}{2} \\ \dfrac{1}{2} \\ 0 \end{bmatrix}.$$

At this solution $\rho_{12} = -k$.

In practice, the size of the test will not be very sensitive to ρ_{12} as long as ρ_{12} is negative. Suppose that $c = c(1) = c(2)$ and $\Phi(-c) = 1 - \alpha$. Consider the following two extreme cases:

$$\Phi(-c, -c; -1) = \Pr(|t| > c) = 1 - 2\alpha$$

$$\Phi(-c, -c; 0) = \Pr(t(1) > c)\Pr(t(2) > c) = (1-\alpha)^2.$$

However, since $(1-\alpha)^2 - (1-2\alpha) = \alpha^2$ and α is small even at the two extremes, the difference in the size of the test will be very small. For example, when $\alpha = 0.025$, $1 - 2\alpha = 0.95$ and $(1-\alpha)^2 = 0.9506$.

[15] This problem is completely symmetric with respect to the cells, so without loss of generality we ignore the case where $P_p^*(3) = 0$.

TABLE A.1.—MONTE CARLO EVIDENCE ON LEAST FAVORABLE DISTRIBUTION
PROBABILITY OF ACCEPTANCE

True Distribution	Critical Region $c(1) = -1.96$, $c(2) = -1.96$, $c(3) = -1.96$	Critical Region $c(1) = -1.80$, $c(2) = -2.20$, $c(3) = -1.80$	Critical Region $c(1) = -1.65$, $c(2) = -3.30$, $c(3) = -1.65$
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.00, 1.00)$	0.950	0.948	0.948
$P_c = (0.20, 0.20, 0.60)$ $P_d = (0.40, 0.40, 0.20)$ $P_p = (0.00, 0.00, 1.00)$	0.946	0.946	0.947
$P_c = (0.30, 0.10, 0.60)$ $P_d = (0.60, 0.20, 0.20)$ $P_p = (0.00, 0.00, 1.00)$	0.947	0.945	0.949
$P_c = (0.25, 0.26, 0.49)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.02, 0.98)$	0.971	0.960	0.945
$P_c = (0.25, 0.30, 0.45)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.10, 0.90)$	0.976	0.965	0.951
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.49, 0.01)$ $P_p = (0.00, 0.01, 0.99)$	0.964	0.957	0.948
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.25, 0.25)$ $P_p = (0.00, 0.25, 0.75)$	0.973	0.961	0.947
$P_c = (0.26, 0.26, 0.48)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.02, 0.02, 0.96)$	0.994	0.995	0.993
$P_c = (0.30, 0.30, 0.40)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.10, 0.10, 0.80)$	1.000	1.000	1.000

Note: Results were obtained using 10,000 Monte Carlo runs, with sample size 2000 and $k = 0.5$.

We perform a Monte Carlo study to examine the sensitivity of the test size to various distributions of the data consistent with the null hypothesis. We use 10,000 Monte Carlo draws each with a sample size of 2000 and $k = \frac{1}{2}$. Table A.1 presents the fraction of Monte Carlo draws for which the null hypothesis was accepted. The theoretical size is supported by Monte Carlo analysis.

All three critical regions were chosen so that asymptotically the size of the test should be close to 0.05. Since the constraint does not bind in the third cell for any of the true distributions considered, asymptotically only $c(1)$ and $c(2)$ are relevant. The relative sizes of $c(1)$ and $c(2)$ are chosen to be the same in the first column. We predict that the probability of rejecting is 0.95 when the first two constraints bind, 0.975 when only one binds, and 1.0 when none bind. The second critical region is not symmetric. Here the probability of rejecting based on $c(1)$ is 0.964, the probability of rejecting based on $c(2)$ is 0.986, and the joint probability of rejecting is approximately 0.95 when both constraints bind. The third critical region is chosen so that the probability of rejecting based on $c(2)$ is very small. Then the probability of rejecting based on $c(1)$ is 0.95 and the joint probability is 0.95. These theoretical predictions are very close to Monte Carlo predictions. It appears that 2000 observations are sufficient to justify the application of asymptotic theory.

We also perform some Monte Carlo runs to gauge the power of the tests. These results appear in table A.2. Note that it is the probability of accepting the null hypothesis (1 − power) that is reported rather than the power itself. The test has low power against moderate violations of the null hypothesis that are concentrated in a particular cell. For small departures, the test is inconsistent—power is less than size.

Rather than testing the individual cells, we may want to test whether the cdf remains bounded between 0 and 1. Using three cells as before, we can

TABLE A.2.—MONTE CARLO EVIDENCE ON POWER OF TEST
PROBABILITY OF ACCEPTANCE

True Distribution	Critical Region $c(1) = -1.96$, $c(2) = -1.96$, $c(3) = -1.96$	Critical Region $c(1) = -1.80$, $c(2) = -2.20$, $c(3) = -1.80$	Critical Region $c(1) = -1.65$, $c(2) = -3.30$, $c(3) = -1.65$
$P_c = (0.24, 0.25, 0.51)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (-0.02, 0.00, 1.02)$	0.822	0.793	0.764
$P_c = (0.24, 0.38, 0.38)$ $P_d = (0.50, 0.25, 0.25)$ $P_p = (-0.02, 0.51, 0.51)$	0.851	0.805	0.765
$P_c = (0.24, 0.24, 0.52)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (-0.02, -0.02, 1.04)$	0.851	0.805	0.765
$P_c = (0.33, 0.33, 0.34)$ $P_d = (0.67, 0.17, 0.16)$ $P_p = (-0.01, 0.49, 0.52)$	0.932	0.909	0.882
$P_c = (0.30, 0.35, 0.35)$ $P_d = (0.67, 0.17, 0.16)$ $P_p = (-0.07, 0.53, 0.54)$	0.144	0.112	0.087

Note: Results were obtained using 10,000 Monte Carlo runs, with sample size 2000 and $k = 0.5$.

TABLE A.3.—MONTE CARLO EVIDENCE ON LEAST FAVORABLE DISTRIBUTION, CDF TEST
PROBABILITY OF ACCEPTANCE

True Distribution	Critical Region $c(1) = -1.96$, $c(2) = -1.96$, $c(3) = -1.96$, $c(4) = -1.96$	Critical Region $c(1) = -1.80$, $c(2) = -2.20$, $c(3) = -1.80$, $c(4) = -1.80$	Critical Region $c(1) = -1.65$, $c(2) = -3.30$, $c(3) = -1.65$, $c(4) = -1.65$
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.00, 1.00)$	0.969	0.961	0.949
$P_c = (0.20, 0.20, 0.60)$ $P_d = (0.40, 0.40, 0.20)$ $P_p = (0.00, 0.00, 1.00)$	0.963	0.957	0.948
$P_c = (0.30, 0.10, 0.60)$ $P_d = (0.60, 0.20, 0.20)$ $P_p = (0.00, 0.00, 1.00)$	0.965	0.959	0.950
$P_c = (0.25, 0.26, 0.49)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.02, 0.98)$	0.973	0.960	0.945
$P_c = (0.25, 0.30, 0.45)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.00, 0.10, 0.90)$	0.976	0.965	0.951
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.49, 0.01)$ $P_p = (0.00, 0.01, 0.99)$	0.970	0.961	0.948
$P_c = (0.25, 0.25, 0.50)$ $P_d = (0.50, 0.25, 0.25)$ $P_p = (0.00, 0.25, 0.75)$	0.973	0.961	0.947
$P_c = (0.26, 0.26, 0.48)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.02, 0.02, 0.96)$	0.998	0.997	0.994
$P_c = (0.30, 0.30, 0.40)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (0.10, 0.10, 0.80)$	1.000	1.000	1.000

Note: Results were obtained using 10,000 Monte Carlo runs, with sample size 2000 and $k = 0.5$.

write this restriction in terms of the probabilities defined above:

$$H_0: \quad P_p^*(1) \geq 0$$
$$P_p^*(1) + P_p^*(2) \geq 0$$
$$P_p^*(2) + P_p^*(3) \geq 0$$
$$P_p^*(3) \geq 0.$$

Note that the condition

$$P_p^*(1) + P_p^*(2) + P_p^*(3) = 1$$

is imposed automatically. We can proceed exactly as before. We obtain the vector of test statistics by forming the sample analogs of the four conditions above and dividing each by its standard error. The test is analogous to the one performed above.

Note that the conditions imposed here are weaker than before. It is possible for $P_p^*(2)$ to be negative but still satisfy this null hypothesis. However, if the previous conditions hold, then these conditions must hold as well.

As before, the constraint can bind in at most only two of the cells, so we can appeal to our earlier argument. The only difference is that the correlations of the t-statistics will be somewhat different than before. The least favorable distribution will occur when the constraint binds in two of the cells. Notice that if the two constraints that bind are $P_p^*(1) = 0$ and $P_p^*(3) = 0$, then the correlation will be negative. For any other two constraints, the correlation will be positive. We know that the correlation is minimized at the least favorable distribution, so a least favorable distribution must occur when $P_p^*(1) = 0$ and $P_p^*(3) = 0$. In this case the least favorable distribution will be exactly the same as the one derived above.

We present Monte Carlo results for these test statistics in tables A.3 and A.4. Note that since the case $P_p^*(1) = 0$ and $P_p^*(3) = 0$ is analogous to the case $P_p^*(1) = 0$ and $P_p^*(2) = 0$ reported previously, we do not report it here.

As in our previous results, the predicted probabilities are very close to their Monte Carlo analogues. For all three test statistics the least favorable distribution will occur when $P_p^*(1) = 0$ and $P_p^*(3) = 0$, where the probability of acceptance is 0.95. The advantage of the cdf test is that it should have more power against some alternatives. This result is borne out in table A.4, where the cdf test has more power against violations of the null hypothesis in both the first and the second cells.

We now return to the original test but allow an arbitrary number of cells. In this case it is trivial to show that the constraint will bind in $K - 1$ of the cells for at least one least favorable distribution.

TABLE A.4.—MONTE CARLO EVIDENCE ON POWER OF CDF TEST PROBABILITY OF ACCEPTANCE

	Critical Region		
True Distribution	$c(1) = -1.96$ $c(2) = -1.96$ $c(3) = -1.96$ $c(4) = -1.96$	$c(1) = -1.80$ $c(2) = -2.20$ $c(3) = -1.80$ $c(4) = -1.80$	$c(1) = -1.65$ $c(2) = -3.30$ $c(3) = -1.65$ $c(4) = -1.65$
$P_c = (0.24, 0.25, 0.51)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (-0.02, 0.00, 1.02)$	0.814	0.794	0.770
$P_c = (0.24, 0.38, 0.38)$ $P_d = (0.50, 0.25, 0.25)$ $P_p = (-0.02, 0.51, 0.51)$	0.850	0.811	0.766
$P_c = (0.24, 0.24, 0.52)$ $P_d = (0.50, 0.50, 0.00)$ $P_p = (-0.02, -0.02, 1.04)$	0.662	0.700	0.765
$P_c = (0.33, 0.33, 0.34)$ $P_d = (0.67, 0.17, 0.16)$ $P_p = (-0.01, 0.49, 0.52)$	0.936	0.914	0.890
$P_c = (0.30, 0.35, 0.35)$ $P_d = (0.67, 0.17, 0.16)$ $P_p = (-0.07, 0.53, 0.54)$	0.145	0.111	0.085

Note: Results were obtained using 10,000 Monte Carlo runs, with sample size 2000 and $k = 0.5$.

Suppose this were not the case. Suppose that the constraint binds in only $K^* < K - 1$ of the cells. Asymptotically the probability of accepting the null hypothesis depends only on the K^* cells for which the constraint binds. Since the cell probabilities in the other $K - K^*$ are irrelevant, we can change them arbitrarily without affecting the probability of accepting the null hypothesis. Therefore we can redefine a new least favorable distribution by taking all of the mass from the $K - K^*$ cells for which the constraint does not bind and putting it into a single one of those cells. The probability of accepting remains unchanged, but the constraint now binds in $K - 1$ of the cells. This new null hypothesis may be uninteresting, but it does demonstrate that in searching for the least favorable null hypothesis we can restrict ourselves to those in which the constraint binds in $K - 1$ of the cells.