Notes for

Lectures on Evaluation of Social Programs

by

Professor V. Joseph Hotz
The Irving B. Harris Graduate School of Public Policy Studies
University of Chicago
I. The Evaluation Problem and General Issues in Designing Evaluations

1. Definition of Social Program Evaluation Research

*Evaluation Research* seeks to identify and measure the relationship between interventions and their impacts on people's behavior or performance.

- Typically, the causal variables of interest are the results of systematic interventions, typically manipulated via programs or policies of governments or other organizations.

*Example:* What is the effect of a government training program on a variety of human behaviors and performances, such as labor market success?

- The Hypotheses relevant to policy analysis and evaluation research are generally provided by the very form of the intervention to which evaluation research is directed.

*Example:* Does a particular education and training program increase the cognitive achievement or employment and earnings of those affected; if so, by how much? Are these programs, cost-effective?

2. The Fundamental Hurdles Confronting Evaluation Research

2.1 Selection Bias

2.1.1 The problem of selection bias

The possibility that the participants (or construction of a program) were different (as measured by the outcomes of interest) from those not receiving a treatment, i.e., biased relative to the control group, for reasons (conscious, unconscious, deliberate or accidental) having to do with the way in which they were selected, or that they self-selected, for the study.

2.1.2 More formal characterization of the selection bias problem

Let:

\[ Y_{1t} \] denote the earnings (outcome) of the \( i^{th} \) person in calendar year \( t \) if they receive training (treatment) for training (treatment) received in year \( k \) (\( t > k \)).

\[ Y_{0t} \] denote the earnings the individual would receive in year \( t \) if the person did not receive training (treatment) in year \( k \).

- What one observes is individuals being of one of two types:
\[ d_i = \begin{cases} 
1, & \text{if the } i^{th} \text{ individual receives training in year } k, \\
0, & \text{if the } i^{th} \text{ individual does not receive training in year } k, 
\end{cases} \]

where \( d_i \) is an indicator of the \( i^{th} \) person's training or treatment status.

- Let \( Y_t \) denote the observed outcome for the \( i^{th} \) individual in year \( t \).

- The counterfactual outcome associated with the counterfactual state—the treatment (or its absence) that they don't receive—is \( Y_{0it} \), for those who actually receive the training or treatment (i.e., those for which \( d_i = 1 \)), and \( Y_{1it} \), for those who do not receive training or treatment (i.e., those for which \( d_i = 0 \)). It follows that:

\[ Y_{it} = Y_{1it} d_i + Y_{0it} (1 - d_i) = \alpha_i d_i + Y_{0it} \quad (1) \]

- One is interested in knowing

\[ \alpha_i = Y_{1it} - Y_{0it}, \]

for \( t > k \).

- The Fundamental Problem of Evaluation Research (or, more generally, Causal Inference):

  All individuals are only observed in one of the two treatment states so for the same individual we only observe \( Y_{1it} \) or \( Y_{0it} \) but not both! In general, the counterfactual outcome is inherently unobservable since individuals cannot simultaneously participate and not participate in a program.

All evaluation design strategies represent attempts to get second best ways of measuring the counterfactual state.

At best we can learn something about aspects of the distribution of \( \alpha_i \), such as its mean or conditional mean.

- The Problem with Using Observational Data:

Suppose we are willing to settle for learning about the average impact of a program on those who actually receive the training (treatment). That is, suppose we focus on trying to learn about:

\[ \alpha_i = E(\alpha_i | d_i = 1) \]
• Simple Mean-Difference Estimator:

Consider the means of the post-program outcomes for the treatment group and a comparison group who did not receive the treatment:

\[ E(\overline{Y}_T) = E(Y_{it}|d_i = 1) = E(Y_{out} + \alpha_i|d_i = 1) \]  \hspace{1cm} (2)

and

\[ E(\overline{Y}_{Nt}) = E(Y_{it}|d_i = 0) = E(Y_{out}|d_i = 0) \]  \hspace{1cm} (3)

Consider the mean of the difference of \( \overline{Y}_T \) and \( \overline{Y}_{Nt} \):

\[ E(\overline{Y}_T - \overline{Y}_{Nt}) = E(Y_{out} + \alpha_i|d_i = 1) - E(Y_{out}|d_i = 0) \]

\[ = E(\alpha_i|d_i = 1) + [E(Y_{out}|d_i = 1) - E(Y_{out}|d_i = 0)] \]  \hspace{1cm} (4)

The problem of selection bias concerns whether the term \( E(Y_{out}|d_i = 1) - E(Y_{out}|d_i = 0) \) is zero. In general, one cannot presume that it is!

2.1.3 Why might selection bias arise: Incentive Effects for program participation (or non-participation)

• Consider the case of the impact of training on earnings due to a government-sponsored training program:

  • Would one expect the earnings that trainees would have received if they had not gone through training are equal to the earnings of the comparison group? In general, the answer is no!

  • The type of individual who applies for training programs is likely to have less education, on average, than those who do not apply and/or program operators may choose to use low educational attainment as a criterion for selection in order to serve those who are more disadvantaged.

  • Those seeking training, especially in training programs which do not provide a stipend (as is the case with JTPA), might also be highly motivated to obtain a job.

  • To the extent that educational attainment and motivation affect earnings, those seeking training would not have had, on average, earnings in the absence of the program as those in the comparison group.

  • In such situations, the earnings of the comparison group do not, on average, measure the earnings of trainees in the absence of training and the estimation strategy described above does not isolate the mean impact of training on the trained.
More generally, selection bias arises when there is a direct relationship between outcomes and choice of treatments. The distribution of $d_i$ and $(Y_{0i}, Y_{1i})$ are not independent. That is:

$$f(d_i, Y_{0i}, Y_{1i}) \neq f_d(d_i)f_{Y_{0i}, Y_{1i}}(Y_{0i}, Y_{1i})$$

- Economic models of selection (e.g., the Roy Model) suggest that choices of actions may depend upon the relative gains from alternative choices.

In the training example, one might hypothesize that individuals choose whether or not to obtain training so as to maximize the present value of their income. Such enrollment decision rules give rise to the statistical problem of selection bias described above.

2.1.4 Alternative approaches to deal with the selection bias problem

2.1.4.1 Experimental Designs:

- Such designs use random assignment of treatment and control status to generate so that the resulting control group will be guaranteed to meet the condition: $E(Y_{0i}|d_i = 1) - E(Y_{0i}|d_i = 0)$.

2.1.4.2 Quasi-Experimental Designs:

- **Statistical (econometric) adjustments:**

These techniques attempt to use econometric methods to "adjust out" (or control for) the systematic differences between the non-experimental comparison group and the treatment group.

A variety of methods are used.

- **Matching techniques:**

These methods attempt to generate a non-experimental comparison group by trying to find individuals who appear to be "the same as" the members of the treatment group.

2.2 The Problems of Making Type I and II Errors

- In the design of testing of propositions, we typically establish a Null Hypothesis ($H_0$) and an Alternative Hypothesis ($H_a$). We know there are two types of "mistakes" or "errors" that can be made:

  **Type I Errors:** incorrectly concluding that $H_a$ is true (incorrectly rejecting $H_0$)

  **Type II Errors:** failing to detect an effect when there was one (failing to reject $H_0$)
when, in fact, $H_A$ is true).

In classical statistical hypothesis testing, we usually try to minimize Type I Error by the choice of $H_0$ and $H_A$ and by rigging things so that we require strong evidence against $H_0$ before we reject it (setting of the significance level of the test.)

- In order to minimize the problems of lack of statistical power necessary to avoid Type II Error, one can:
  
  - Design evaluations with adequate sample sizes in order to have a chance to obtain minimal effects. (Issues of minimum detectable effects).
  
  - Design evaluations with treatments to enable separation between treatment effects, i.e., make treatment distinctive.

As Cook and Campbell (1979) argue in their book, when making causal inferences such as the impact of a program, a necessary condition is variation in the treatment. In part, this means that one wants differences between what the treatment group experiences versus what the control group experiences, i.e., the treatment is different enough from $H_0$ to make its impact on behavior, if there is any, detectable.

**Example:** If one expects that a particular training program will have minimal effects if only administered for a short-time, then may want to try bolder treatments. Unfortunately, not always possible. (Alteration of the program.)

2.3 The Contamination Bias Problem

- Want to avoid factors entering an experiment that affects the treatment or control (comparison) group in ways which distort the comparison we seek. Can come in many forms:

2.3.1 (Classic) Contamination Bias (or Cross-over) Problem

- The control group members actually receive the treatment. Solutions may involve controlling the disbursement of the treatment, but not always possible.

**Example:** Gary Income Maintenance Experiment.

2.3.2 No-Show Problem

- The treatment group does not receive the treatment

**Example:** No-shows in a training program. Selected for the program but do not show up to receive the treatment.
Example: More subtle. The NIT program of those who are eligible but never receive any payments. Whom do we compare? Those who receive treatment versus the members of the control group? Potential for form of selection bias if choice element in receiving payments. Everyone selected to be eligible for the treatment versus the control group? Issue here is related to whether this is the outcome of interest.

2.3.3 Attrition Bias

- Like No-Shows but cases where receiving the treatment but stop or cannot follow people after receipt of the program. Often case that drop outs may be selective.

Example: Attrition problems in the NIT.

2.4 The “Program (and Program Administrators) Don’t Play Dead” Problem

- Need to get administrators to agree to study and cooperate in conducting the study. Typically, programs have procedures, administrators, etc., which constrain the ability to do evaluative research.

Example: Turn downs in the National JTPA Study.

2.5 The Quality and Consistency of the Treatment and Program Problem

- From perspective of policy, one generally doesn’t want to alter the program. At times it is convenient to do so for purposes of conducting an evaluation.

Example: Altering the pool of applicants in JTPA.

- The “Repairman’s Dilemma”: Should researchers see to it that the quality of a program and its procedures are maintained at a high quality? Should one want to evaluate the program as is, warts and all?

- Does the evaluation, per se, create an artificial program, which in the end, is not of much interest to policy makers.

2.6 The Problems of Gathering Data

- Necessary part of an evaluation but it can have its problems.

- Differential Reporting Incentives: Treatments may have a strong incentive to report, but may be less so for controls?

Example: NIT experiments.
• Minimizing intrusiveness of data gathering. Use of survey interviews versus other ways of monitoring, through administrative data. Differential reporting.

• Anticipating gathering the right data. Problem of unexpected consequences of a program

Example: SIME/DIME marital instability). How to make sure you gather the right data.

• Adequate Baselines in “before and after” studies.

• Attrition Problem Again: How do you make sure you can find people?

3. Internal Versus External Validity of An Evaluation

• The above “hurdles” all represent threats the validity of the study in terms of ability to make inferences concerning the impact of a program on behavior based on our evaluation. (Inferences about Causality). Following the terminology of Campbell and Stanley (1963), we worry about two forms of validity: internal validity and external validity of our evaluation.

3.1 Internal Validity

• The approximate validity with which we infer that a relationship between treatments and outcomes is causal or that the absence of a relationship implies the absence of cause. Particular attention here is on the sample, program, and program participants we studied. For that group, can we reliably draw a conclusion as to what the treatment did to behavior. Most of the hurdles noted above threaten this form of valid inference and represent what we seek, in designing our evaluation study, to avoid or minimize.

3.2 External Validity

• The approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternative measures of the cause and effect and across different types of programs, participants and environmental conditions (e.g., states of the economy, types of program administration, etc.) Here the concern is the representativeness of our findings—however internally valid they are—to other circumstances. Issues of selection of sites, alterations of the program, etc. may matter crucially in the generalizations we can make for a particular evaluation study.

4. Contexts for Evaluation Research

• In his book on the role of social science research in government, Richard Nathan (1988) distinguishes between two types of evaluation research: Demonstration Research and Evaluation Research.
4.1 Demonstration Research

- Demonstration research is designed to test new programs and policy innovations implemented through a limited number of pilot or demonstration projects.

  **Examples:** The NIT experiments, the NSW Demonstration, and the other social experiments are examples. Such evaluations involve the design and testing of a new program.

- A key feature of demonstration projects is that they provide an easier rationale for the use of random assignment with its denial of services to members of a control group than is the case in other social contexts. In general, random assignment raises ethical and legal problems because of the potential that the denial of treatments may harm people or deny them something to which they are entitled, either legally or effectively. Demonstration projects minimize the problems associated with the denial because the treatment represents a service to which the population is not entitled. That is, the null treatment for controls in a demonstration project is the status quo.

- Finally, demonstration research differs from evaluations of on-going programs in terms of the goals of the research. Demonstration research typically has the more limited goal of determining whether a program *might* work. Such research focuses on questions of feasibility and likely direction of impacts. Given this focus, demonstration research generally is not expected to provide results that generalize to all potential program participants and to all possible states of nature in which the program might operate if it were adopted.

4.2 Evaluation of On-Going Programs

- Evaluate the impacts an existing program. There are at least three problems that arise (or are more difficult) in evaluating on-going programs than in conducting demonstration research.

4.2.1 Lack of Control

- The first problem is the inherent lack of control over the design of the program. The “treatments” are dictated by the program and frequently they are not neatly categorized as they can be in demonstration projects. The selection processes in an existing program may not be based on easily quantifiable criteria. They may differ across program units or program administrators. Such diversity complicates the analysis of the program's impact. More importantly, unlike demonstration studies, researchers are generally not free to change the way an existing program operates. This is true because in evaluating an on-going program, interest centers on how the program operates "as is." Typically, those who commission evaluation research are interested in the impact of the program(s) which currently exist.
4.2.2 Establishing Reliable Information on the Counterfactual State

- Information on what behavior would be like if the program did not exist or if it had not provided services to a program participant is a much more difficult to obtain. This may be so because the use of random assignment is generally difficult to implement. Program operators or public officials are likely to object to the denial of services to individuals who apply to a program, objecting that it is inappropriate to use individuals as human “guinea pigs.” This reluctance is heightened when such evaluations involve substantial intrusions into the program such as implementing an experimental design. As Nathan (1988) notes, this lack of cooperation stems from the inherent differences in objectives between those running an on-going program and those trying to evaluate it. Program administrators are interested in providing services to individuals; they do not view their role as helping to facilitate evaluation of their program. Such administrators “may not want research to be conducted because they fear it would show a policy they favor to be ineffective or, if it works, to have results that fall short of what had been promised.”

4.2.3 The Differences in What You Want to Learn—Not Altering the Existing Program

- The third problem is that the question being addressed in evaluations of on-going programs are more difficult to answer relative to those for demonstration projects. As noted above, demonstration research seeks to address the question of what might happen if a new policy was to be implemented. Such evaluations are “feasibility studies,” determining whether something might work. In contrast, in evaluations of existing programs the central question is: does it work? This question is inherently more demanding because it is important that the results of such evaluations be representative of the program and populations it serves.

5. The Types Evaluation Research

- This is determined ultimately by what one’s answers are to the following question: What questions are to be addressed? What are the policy issues? What are the outcomes one wishes to study?

  - Process Analysis:

    How does the program work?

  - Impact Analysis:

    Does the program have a work? What effect does it have on behavior? How “big” is the impact?

    Main focus on my remarks.
• Cost-Benefit Analysis:

Is the program cost-effective? Judged by some criteria, do the benefits of the program, outweigh the costs?

♦ Each alternative analysis will entail different designs.
References


II. Research Questions of Interest for Evaluation Research Concerning the Impact of Programs

1. Some Preliminary Definitions and Notation

♦ **Treatment Regimes:**

Let $S$ denote *Treatment Regimes*, where $S=s$ denotes the particular treatment regime an individual (or group) selects or to which is assigned or is otherwise exposed.

Examples of possible treatment regimes would be:

- being accepted in or assigned to a training program or not,
- having access to a new drug
- being eligible for a particular subsidy or not,
- living in a region which has access to a particular set of services or is subject to a particular set of laws or regulations

For sake of illustration and simplicity, suppose that there are only 2 treatment regimes, which are denoted by:

$S=T$ denotes having access to the treatment regime (and its incumbent services)

$S=C$ denotes *not having access* to the services in treatment regime $T$.

♦ **Treatment Choices:**

Let $D(S=s) = D(s) = k$ denote the *Treatment Choice Decision* an individual makes, *conditional* on being in a particular treatment regime $s$, where $k (= 0, 1, ..., K)$ denotes the particular treatment chosen.

Examples of treatments which an individual might choose would be:

- A person actually receiving the training having been assigned to a training program.
- A patient actually taking the full dosage of a drug
- Someone actually exercising their option of claiming a subsidy (such as welfare) in a state
Again, for sake of illustration and simplicity, let there be only 2 treatment choices, given by:

\[ D(s) = 0 \] denotes the **null treatment choice** in which the individual, assigned to treatment regime \( s \), chooses to *not* use any of the services available in the \( S=T \) regime.

\[ D(s) = 1 \] denotes the treatment choice in which the individual, assigned to treatment regime \( s \), chooses to *fully comply* and actually receive the services available in the \( S=T \) regime.

For example:

- A person assigned to a training program who actually receives the complete course of the prescribed training would be recorded as: \( D(T) = 1 \)

- Someone who was accepted into the training program but did not choose to participate would be recorded as: \( D(T) = 0 \)

- The individual in an experimental control group who was not given access to the treatment drug being studied would be recorded as: \( D(C) = 0 \)

♦ **Outcomes:**

Conceptually, at least, one can characterize the outcomes that an individual would realize if they were under different treatment regimes and/or if the individual were to choose different treatments, regardless of the treatment regimes and/or treatment choices the individual actually makes.

\[ Y_s = Y | S = s \] denote the outcome an individual would realize if they had been in Treatment Regime \( s \), for \( s = T \) or \( C \).

\[ Y_k = Y | D(\cdot) = k \] denote the outcome an individual would realize if they had experienced Treatment Choice \( k \), for \( k = 0 \) or \( 1 \).

♦ Conditional on \( X \), we treat the above variables as random. Our research interest is in designing (and conducting) evaluations which enable us to estimate aspects of their distribution.

Let \( f(\cdot; \ldots) \) denote a density function for its arguments and \( P(\cdot) = Pr(\cdot) \) denote the probability function for a discrete event.
2. Alternative Questions of Interest Concerning the Impacts of Programs in Evaluation Research

- There are a number of alternative questions one might ask in characterizing the impact of a program and its treatments on outcomes. The questions differ with regard to what population is of interest, what aspect of the program structure is considered and what summary of the distribution of outcomes one is interested in using (e.g., the mean, the median, etc.). Each answer potentially different policy-relevant questions and differ in the difficulty in designing evaluations to answer them. Below, we focus on a subset of possible questions and limit most of our attention to expected values (means) of impacts. (See Heckman (1992) and Manski (1992) for discussions of identifying aspects of distributions other than the mean.)

Q1: What is the effect on outcomes of receipt of a particular treatment for those who chose that treatment?

\[ \alpha = E(Y_1 - Y_0 | D(T) = 1, S = T) \]

For the running training example, this is the effect of training on those who actually receive training.

Q2a: What is the effect on the outcomes of individuals who have access to a particular treatment regime?

\[ \tilde{\alpha} = E(Y_T - Y_C | S = T) \]

This effect measures the impact of having access to a regime, relative to not having it. Here, one is not concerned whether one chooses to take a particular treatment or not. The relevant notion here is the "insurance value" of having access to a training program and the possible effect it might have on behavior and outcomes. In the training example, it would represent the average effect on earnings of having access to a training program.

Q2b: What is the likelihood of an individual selecting a particular treatment regime, given the feasible regimes and that one has some discretion what regime they face?

\[ P(S = s | \text{feasible set for } S) \]

Here the interest is on whether an individual chooses a particular treatment regime. In many contexts of interest for program evaluation, this may not be feasible. For example, an individual may not be able to determine whether a state government has a training program or not; the presence of such programs are exogenous to the individual. But that individual may be able to choose what governmental services they have by choosing where to live. It is the latter type of decision that is at issue in Q2b.
Q2c: *What is the likelihood of an individual choosing a particular treatment, given access to a particular treatment regime?*

\[ P(D(T) = 1 | S = T) \]

In the training example, one may be interested in whether an individual, who is eligible for a training program, elects their option and chooses it.

Q3: *What would be the effect of a particular treatment on the outcome of a randomly selected member of the population?*

\[ \alpha^* = E(Y_1 - Y_0) \]

The question would be relevant if one is considering the likely consequences of making a program treatment mandatory. For example, one might be interested in knowing what the effect of a mandatory drug testing program would have on the productivity of the average worker.

♦ Q3 is inherently the most difficult question about which to make inferences and Q1 is the easiest, although it may not be that easy.

♦ In general, observational data does not allow one to make unbiased inferences for Q1, Q2a or Q3.

- Observational data gives information on \( f(Y \mid D(s)=k, S=s) \), i.e., on the distribution of outcomes for the choices that individuals make. Generally, observational data does not provide any information on the distribution of *counterfactual outcomes*.

- As a consequence—as noted in Lecture I—inferences drawn from observational data may be subject to selection bias.

- The question arises as to the use of alternative designs, noted in Lecture I, for drawing inferences about the Q1, Q2a, and Q3.

♦ Note that one may be able to draw inferences about Q2b or Q2c.

- The feasibility of drawing such inferences hinges crucially on the nature of observed variation in treatment regimes.
References


III. The Logic of Experimental Evaluation Designs and Their Applicability in Social Contexts

1. The Ideal Experiment: Its Underlying Assumptions and Advantages

1.1 A Simple Experiment with Random Assignment of Treatment Regime, 2 Treatment Choices and a Perfectly Embargoed Control Group

♦ Maintained Assumptions in this Case:

- **Assumption A1**: Treatment Regimes, $S=T$ and $S=C$, are randomly assigned to members of a sample.

- **Assumption A2**: No "Hawthorne Effects"

  Let $T^*$ and $C^*$ denote the treatment statuses that would exist in the absence of an experiment. (What would exist in the "real world.") Let $T$ and $C$ denote the treatment statuses that exist in the context of the experiment.

  We assume that: $T^* = T$ and $C^* = C$.

- **Assumption A3**: Perfectly Embargoed Control Group:

  Assume that the design on the experiment is such that the following condition holds for all members of the control group:

  $$P(D(C) = 1) = 0$$

  That is, no members of the control group are able to choose treatment status 1.

- **Assumption A4**: There are only two treatment choices associated with the $S=T$ Treatment Regime: $D(T) = 1$ and $D(T) = 0$.

♦ Inferences about Q2a:

- It follows from the above assumptions that:

  $$E(Y_T|S = T) - E(Y_C|S = C) = \bar{\alpha} + [E(Y_C|S = T) - E(Y_C|S = C)]$$

  $$= \bar{\alpha}$$

  Since $E(Y_C|S=T) = E(Y_C|S=C) = E(Y_C)$.

- Thus, the use of random assignment in this case ensures that the simple mean difference between the outcomes of the treatment and control groups is an unbiased estima-
Inferences about Q1:

- While $S$ is randomly assigned, the above design does not guarantee that Treatment Choice is random. In general, it is not! The existence of the potential for experimental subjects to exercise choice is an example of non-compliance in experiments.

- Note that one can assume that the following condition holds:

Assumption A5:

$$P(D(T) = 0) = 1$$

in which case $E(Y_{iT}|S=T) = E(Y_{i0}|D(T)=1)$ and, thus, $\alpha_1 = \bar{\alpha}$. Then the mean difference in outcomes between experimental and control groups yields an unbiased estimator of $\alpha_1$.

- Even if one does not make Assumption A5, it turns out that one can make unbiased inferences about $\alpha$ (This result is due to Bloom (1984). See Hotz and Sanders (1994) for citation.)

The above Assumptions imply that the following result:

$$\frac{E(Y_{iT}|S=T) - E(Y_{i0}|S=C)}{P(D(T) = 1)} = E(Y_{i0}|D(T) = 1) - E(Y_{i0}|D(T) = 1) = \alpha.$$  \hspace{1cm} (2)

The derivation of (2) is straightforward. First, note that $E(Y_{iT}|S=T)$ can always be written as the following weighted average:

$$E(Y_{iT}|S=T) = P(D(T) = 1)E(Y_{i0}|D(T) = 1) + [1 - P(D(T) = 1)]E(Y_{i0}|D(T) = 0).$$  \hspace{1cm} (3)

Moreover, $E(Y_{iC}|S=C)$ can also be expressed as a weighted average of the mean of $Y_0$ for the two latent types, $D^L(C) = 0$ and $D^L(C) = 1$, where the weights are the proportions of the control group that are these latent types. That is:

$$E(Y_{iC}|S=C) = P(D^L(C) = 1)E(Y_{i0}|D^L(C) = 1) + [1 - P(D^L(C) = 1)]E(Y_{i0}|D^L(C) = 0)$$

$$= P(D(T) = 1)E(Y_{i0}|D(T) = 1) + [1 - P(D(T) = 1)]E(Y_{i0}|D(T) = 0)$$

where the second expression follows from the assumption that the control group is perfectly embargoed from treatment choice and from the no Hawthorne effect assumption. The result in (2) follows since the difference between the mean outcomes for the treatment and control groups is proportional to $[E(Y_{iT}|D(T)=1) - E(Y_{i0}|D(T)=1)]$, where the factor of proportionality is the inverse of the probability of choosing Treatment 1 [$P(D(T)=1)$].
Thus, under this special case, one can identify the effect of a treatment for those who receive it, even though the choice process governing the decision to take the treatment is not random.

- Inferences about Q3:
  - In general, the above experimental design does not provide data with which to make unbiased inferences about Q3.
  - The exception is if one assumes: Assumption A5 and
    Assumption A6: The difference, $Y_1 - Y_0$ is the same for all individuals (the constant treatment effect assumption).

    Then it follows that $\alpha_1 = \alpha^*$.  

2. Conducting Experiments in Social and Program Contexts: The Less than Ideal Case

- While the experimental design has many desirable properties and an ideal, the use of such designs in social contexts often entail actual designs which do not meet the conditions noted in Section 1. We discuss, in turn:
  - some of the violations that are likely to arise in social contexts
  - their consequences for inferences drawn from simple random assignment designs
  - potential adjustments for such problems

2.1 Noncompliance with “Intended” Treatment Protocols

- The Problem of “No-Shows” for the S=T Treatment Regime
  - Individuals assigned to S=T treatment regime end up choosing the null treatment, $D(T)=1$.

    Individuals who are accepted into a training program do not show up for the program.

    Individuals in a drug clinical trial who are assigned a new treatment do not take it.

  - Problem discussed in Section 1. Simple mean differences in outcomes of experimental and controls does not provide unbiased estimates for Q1 (i.e., $\alpha_1$).

    We noted a solution, due to Howard Bloom, for “correcting” for no-shows in case where controls are perfectly embargoed from treatment choice [Assumption 3] and
there are only two treatment choices, $D(T)=1$ and $D(T)=0$, for the $T$ treatment regime [Assumption 4].

As noted in Hotz and Sanders (1994), this “correction” for identifying $\alpha_i$ does not hold if either of these two assumptions are violated.

There may be more than 2 treatment choices—i.e., $D(T) = k$, for $k = 0, \ldots, K$, $K > 2$.

- **The Problem of “Cross-Overs” from the $S=C$ Treatment Regime**

- Individuals assigned to $S=C$ treatment regime end up choosing the null treatment, $D(T)=1$.

  Individuals applying for a particular training program who are randomly assigned to the control group, end up going to getting training from another source.

  Individuals in a drug clinical trial assigned to the placebo group end up getting the drug via “drug sharing.”

- Again, in this case, the simple mean differences in outcomes of experimental and controls does not provide unbiased estimates for the more general form of Q1 in which interest focuses on:

  $$\alpha_k = E(Y_k - Y_0 | D(T) = k, S = T)$$

  for $k = 1, \ldots, K$. That is, we want to know the effects of various types of treatment choices—such as partial compliance with a treatment protocol—relative to the null treatment case.

- Note that both problems are inherent when dealing with human subjects. In this sense, experimental evaluations with human subjects are different than those in agriculture, etc.

- **Possible “Solution” to Problems of Non-Compliance: Use of Experimental Data to Identify a Bound on $\alpha_k$.**

  Two cases to consider:

  - **Estimating Effects for “Partial Compliance” where there is Perfect Embargoing of the Control Group:**

    Situations in which experimental subjects can exercise choice over the “treatment” they actually receive, but control subjects are perfectly embargoed from treatment choice [i.e., cross-overs, in the strict sense of this term, are not allowed].

  - **Estimating Effects for “Partial Compliance” when Controls are Not Perfectly Embargoed from Choice:**
Situations in which control subjects are not perfectly embargoed from treatment choice [i.e., cross-overs can occur].

Here, I only describe the bounding strategy for the first case. [See Hotz and Sanders (1994) for bounds applicable to second case and details.]

Given perfect embargoing, expression (4) generalizes to:

\[
E(Y_0|S = C) = \sum_{j=0}^{K} P(D^j(C) = j)E(Y_0|D^j(C) = j) = P(D^k(C) = k)E(Y_0|D^k(C) = k) + \sum_{j=k+1}^{K} P(D^j(C) = j)E(Y_0|D^j(C) = j), \quad (5)
\]

\[
= P(D^k(T) = k)E(Y_0|D^k(T) = k) + [1 - P(D^k(C) = k)]E(Y_0|D^k(C) = \neg k),
\]

where \(D^k(C) = \neg k\) denotes the fact that the latent treatment choice is not \(k\).\(^1\)

The fact that \(E(Y_C|Z=C)\) "contains" \(E(Y_0|D(T)=k)\) provides scope for identifying bounds on \(\alpha_k\).

To see this, note that from (5), one can solve for \(E(Y_0|D(T)=k)\) to obtain the more general form of an expression for (2):

\[
\alpha_k = \left(\frac{1}{P(D(T) = k)}\right)\left[c_1 - c_{2k} + (1 - P(D(T) = k))E(Y_0|D^k(C) = \neg k)\right], \quad (6)
\]

where

\[
c_1 = E(Y_T|S = T) - E(Y_C|S = C),
\]

\[
c_{2k} = [1 - P(D(T) = k)]E(Y_{0\neg}|D(T) = \neg k),
\]

for \(k = 1, \ldots, K\). Since \(E(Y_j|D(T)=j), P(D(T)=j), j = 1, \ldots, K, E(Y_T|S=T)\) and \(E(Y_0|S=C)\) are identified from experimental data, \(c_1\) and \(c_{2k}\) are identified.

Because \(E(Y_0|D^k(C)=\neg k)\) is not identified by experimental data, one cannot achieve point identification of \(\alpha_k\).

However, one can bound \(\alpha_k\) by placing upper and lower bounds on the latter conditional expectation. In particular, given that \(c_1\), \(c_{2k}\), and \(P(D(T)=k)\) are identified from experimental data, it follows that deriving bounds on \(\alpha_k\) hinge on obtaining bounds on \(E(Y_0|D^k(C)=\neg k)\), the mean outcome in the control group for latent treatment groups other than \(k\).

\(^1\)Note that \(E(Y_T|S=T)\) can be written in a similar fashion.
Several alternative sets of bounds on $\alpha_k$ can be formed. These are what Hotz and Sanders call the \textit{Horowitz-Manski Bounds on $\alpha_k$}.\footnote{The derivation of these bounds, and their properties, follows from results in Horowitz and Manski (1993) on forming bounds with contaminated samples.}

$E(Y_C | S=C)$ represents a \textit{contaminated} measure of the object of interest, $E(Y_0 | D(T)=k)$ it follows that $E(Y_C | S=C)$ "contains" $E(Y_0 | D^*(C)=k)$. Given the Perfect Embargo Assumption [Assumption 4], the fraction $[1-P(D(T)=k)]$ of the control group has the latent treatment status $D^*(C)=k$ and the remaining $P(D(T)=k)$ proportion has latent status $D^*(C)=k$.

While we do not know which of the control group members have the non-Treatment-$k$ latent statuses, we can form lower and upper bounds on $E(Y_0 | D^*(C)=k)$ in the following way.

Suppose we assume that all of the observations for which $D^*(C)=k$ have values of $Y_0$ which lie below the $P(D(T)=k)$-quantile and above the $[1-P(D(T)=k)]$-quantile, respectively, in the distribution of $Y_0$ for the control group. That is, to get the lower bound on $E(Y_0 | D^*(C)=k)$, we presume that all of the observations on $Y_0$ with $D^*(C)=k$ lie in the "lower tail" of $Y_0$ and in the "upper tail" for its upper bound. Assuming $N_C$ is the number of subjects in the overall control group, let

\[ Y_{0,1-P(D(T)=k)N_C} \text{ and } Y_{0,P(D(T)=k)N_C} \]

denote, respectively, the $[1-P(D(T)=k)]N_C$ and $P(D(T)=k)N_C$ order statistics of $Y_0$ for the control group. Then the lower bound on $E(Y_0 | D^*(C)=k)$ is given by:

\[ E(Y_0 | S=C, Y_0 \geq Y_{0,1-P(D(T)=k)N_C} ), \]

the \textit{truncated mean} of $Y_0 | D(T)=k$ such that $Y_0 \leq Y_{0,1-P(D(T)=k)N_C}$, and the upper bound is

\[ E(Y_0 | S=C, Y_0 \leq Y_{0,P(D(T)=k)N_C} ). \]

Substituting the corresponding truncated means for $E(Y_0 | D^*(C)=k)$ in (5), one obtains a new set of upper and lower bounds on $\alpha_k$. Denoted by $[B_{Lk}^2, B_{Uk}^2]$, these bounds are defined as:

\[ B_{Lk}^2 = c_1 - c_{2k} + \left( \frac{1-P(D(T)=k)}{P(D(T)=k)} \right) E(Y_0 | S=C, Y_0 \leq Y_{0,1-P(D(T)=k)N_C} ) \] (7a)

and

\[ B_{Uk}^2 = c_1 - c_{2k} + \left( \frac{1-P(D(T)=k)}{P(D(T)=k)} \right) E(Y_0 | S=C, Y_0 \leq Y_{0,P(D(T)=k)N_C} ). \] (7b)
• The Horowitz-Manski bounds are robust and impose no further restrictions or constraints on experimental data than those implied by the above Assumptions.

• They “tightest” bounds that can be formed without invoking further assumptions.

2.2 Accounting for Macro Effects Associated with Social Programs

♦ The implementation of a permanent program may have several effects on the macro environment which may represent part of the “impact” associated with a program. The following are some of the macro effects which would accompany the implementation of a new permanent program:

• Market-equilibrium Effects

  Implementation of large-scale jobs creation training program may affect the equilibrium in the labor market which is affected.

• Information Diffusion Effects

  Information about a new set of social services (job counseling for the poor) may reach different populations after information has been transmitted through a community than would be the case in its initial form.

• Social Interaction Effects

  Changes in the attitudes of a society concerning discrimination after the adoption of “open-housing” legislation may result in a different impact of such legislation on the home-buying behavior of minorities than prior to the changes in these social norms or interactions.

♦ All of these effects might not be measurable with a Demonstration Project. As a consequence, designing a demonstration project using a micro experiment would not be able to measure these effects.

♦ Observational Data and use of non-experimental evaluation methods, in which outcomes are measured over time and across geographically separated regions or neighborhoods may be better suited to deal with macro effects.

♦ Use of an experimental design in which treatment regimes are randomly assigned across regions might be a partial solution to dealing with macro effects.

2.3 Problems Accounting for Entry Effects in Experimental Designs

♦ In typical experimental design of a program—such as a training or welfare program—individuals who have applied to and/or are subject to the program are those at risk of being assigned to a Treatment Regime.
• Applicants to a new training program are either randomly assigned to have access-to-the-training \((S=T)\) or are denied access \((S=C)\). This set of subjects are then followed and their subsequent outcomes (e.g., earnings, labor force participation, etc.) is measured and compared.

• But, over time (as the program matures) and/or in different economic conditions, the applicant pool may change, i.e., those who wish to enter the new program may change. As a consequence, the results of the typical experiment need not apply to these new entrants.

• **Possible Solutions:**

  • If feasible—and this is a big if—one may be able to randomly assign an entitlement-to-treatment before an individual even applies.

    For example, the Military Draft Lottery in the U.S. during the Vietnam War Era is an example of such a random assignment.

    In such cases, one may be able to then monitor how entry is affected by a change in the regime.

  • For example, regions or neighborhoods might be randomly assigned a particular treatment—such as expedited access to a set of services—and other regions would not receive such services. Then one could monitor the impact of the new treatment (expedited services) on the differential rate of utilizing social services.

3. **The Use of Experimental Designs to Identify “Structural” Models of Behavior**

• In an earlier era of program evaluation, advocates of experiments argued for the use of random assignment to generate exogenous variation with which to identify structural models of behavior.

  • The early designers of the Negative Income Tax (NIT) Experiments in the U.S. [see Cain and Watts (1973)] advocated evaluations in which welfare guarantees and benefit-reduction rates were randomly assigned to poor populations in order to obtain better estimates of the income and substitution effects for models of labor supply equations.

  • The designers of the Residential Electricity Time-of-Use Pricing Experiments [see Aigner (1985)] used data from experiments in which different time-of-day pricing schemes were randomly assigned to residences as a way of estimating price elasticities for electricity demand equations.
More recently, the actual applications of experimental designs to program evaluation have had a black box orientation.

- The focus has been on the identification of the net impact of one treatment regime to a null treatment regime.
- Simple mean differences between experimentals and controls have been the focus of such analyses.
- Such results often tell little about how individuals would respond to different treatments that one might envision but that are not the same as those considered in the experiment itself.

In my view (and in the view of others), this is an unfortunate development.

- Heckman (1992) has argued for more attention to designing experiments with an eye to identifying parameters characterizing structural models.
- Greenberg, Meyer and Wiseman (1993) also have argued for designs of experiments in welfare-to-work initiatives in the U.S. to identify “production function” for producing work-related “skills” among the poor.

In attempting design experiments with such a goal in mind, several issues need to be considered:

- A larger number of treatments should be used in the design to maximize the information about the production function “response surface.”
- Attempts should be made to maximize the distinctness of the treatments.
- Other issues.
References


IV. Designing Experimental Evaluations of Social Programs: 
The Case of the National JTPA Study

1. Introduction

1.1 The Problem: Can we obtain "reliable" estimates of the impact of social programs such as the manpower training programs of the Job Training Partnership Act (JTPA)?

♦ Virtually all of these evaluations use nonexperimentally-based statistical methods for estimating the impact of the programs.

♦ Problem confronted in such analyses is selection bias.

♦ In the recent literature on program evaluation, several authors have argued that alternative nonexperimental estimators of program impact produce a disconcertingly wide range of estimates even when applied to the same data.

See Table 1.1.

"...estimates of program effects that are based on nonexperimental comparisons can be subject to substantial misspecification uncertainty" (Burtless and Orr, 1986, p. 613)

and that

"...randomized clinical trials are necessary to determine program effects" (Ashenfelter and Card, 1985, p. 648).

Barnow (1987) argues that

"...experiments appear to be the only method available at this time to overcome the limitations of nonexperimental evaluations" (p. 190).

♦ LaLonde and Maynard (1987) compare the experimental the experimental estimates of the National Supported Work Demonstration impact with estimates obtained using nonexperimental procedures and find that:

"the nonexperimental procedures may not accurately estimate the true program impacts. In particular, there does not appear to be any formula [using nonexperimental methods] that researchers can confidently use to replicate the experimental results of the Supported Work Program. In addition, these studies suggest that recently developed methods for constructing comparison groups are no more likely (and arguably less likely) than the econometric procedures to replicate the experimental estimates of the impact of training."

They conclude that these
"findings are further evidence that the current skepticism surrounding the results of nonexperimental evaluations is justified." (LaLonde and Maynard, 1987).

1.2 The Disadvantages of Using Non-Experimental Methods for Evaluating Manpower Training Programs

- Evidence of Wide Range of Estimates Using this Method
- "Model Misspecification Uncertainty"
  
  Controversy over what is the "correct" Selection Correction Method
- Inherent Difficulty in Conveying Results to Policy Makers

1.3 Designing The National JTPA Study: A Two-Pronged Strategy

- In light of these findings the Job Training Longitudinal Survey Research Advisory Panel recommended that to evaluate the impact of the Job Training Partnership Act (JTPA), DOL should:
  
  "perform a selected set of classical experiments over the next several years that involve random assignment of program-eligible individuals to the treatment (experimental) group and to the non-treatment (control) group...[with the intent] to use these experiments to evaluate the net impact of JTPA for selected target/treatment groups in a set of SDAs that volunteer to participate"

  and

  "Further, it is intended to use these experimental results and the understanding of the selection process gained thereby to improve the effectiveness of quasi-experimental designs as a strategy for program evaluation."

2. Evaluating Demonstration Projects versus Existing On-Going Programs: Key Differences and their Consequences for Designing Evaluation Studies

- Three Problems in Evaluating On-Going Programs:

2.1 The "treatments are dictated by the program and frequently are not neatly categorized as they can be in demonstration projects.

2.2 Establishing the "Counterfactual" State

- Information on what behavior would be like if the program did not exist or if it had not provided services to a program participation is much more difficulty to obtain with on-going programs.
2.3 *Question being addressed in evaluating on-going programs is much more difficult to answer*

- Demonstrations address the question of what *might* happen if a program is implemented.
- For on-going programs the question is: *does it work?*

3. **The JTPA System: Key Features and their Challenges To Evaluation**

- The Decentralization and Diversity of the JTPA System
- The Multi-faceted and Complex Governing Structure of the JTPA System
- Who is Served and the Role of Performance Standards in the JTPA System
- Implications of JTPA Program Features for Evaluation

4. **The Design of the Experimental Component of the National JTPA Study**

   *See Table 1.3.*

4.1 *How should the sites (SDAs) in which to conduct the study be selected and how could their participation be gained?*

- **Ideal:** Would like to use a random (or stratified) sampling scheme to select sites in order to obtain nationally representative results.
  
  20 with 30,000 clients normally served with SDAs chosen randomly

- **Reality in JTPA Study:** Take virtually any sites which would agree to participate.
  
  16 SDAs with approximately 23,000 clients in those SDAs which "cooperated"

   *See Tables 5.2, 2.1*

4.2 *How could the intrusion on the operations of the SDAs be minimized while conducting the experiments?*

- **Ideal:** Would like to minimize intrusion of program in order to assess programs as they normally operate.

- **Reality in JTPA Study:** Modifications in Performance Standards and Allocations had to be done in order to gain cooperation of local programs.

IV-3
4.3 What groups should be studied and how should they be disaggregated?

- Adult Women
- Adult Men
- Out-of-School White Youth
- Out-of-School Minority Youth

Separate groups given differences across groups in previous findings and differences in labor market conditions facing these groups.

4.4 What should be the definition of “treatments” in the Study and, thus, what type of impact estimates would be provided?

- Originally: to be:
  - On-the-Job Training (OJT)
  - Classroom Training and Occupational Skills Training (CT-OS)
  - Job Search Assistance (JSA)

- In the End: See Table 4.2

4.5 At what stage in the program’s processing of program applicants should random assignment be conducted?

See Figures 1 and 4.1

4.6 What should be the allocation of participants between treatment groups and control status?

- In the End: 1 in 3 will be randomly allocated to Control Group Status

4.7 How long should controls be “embargoed,” i.e., denied access to JTPA services?

4.8 How Will the Participants be Allocated Across Types of Training?

See Table 2.

1 in 3 will be randomly allocated to Control Group Status

4.9 How large should the treatment and control groups be to obtain estimates with statistical power?

See Table 5.9.
4.10 What Kinds of Analyses can one do given the (Experimental) Design?

♦ Simple Mean Experimental vs. Control Comparisons for each Treatment (OJT, CT-OS, and Other Activities (OA)) separately by Target Groups

♦ “Corrections” for No-Shows and Cross-Overs

♦ Benefit-Cost Analyses

5. Designing Experimental Evaluations of On-Going Programs: Tentative Conclusions

♦ Difficulty in Conducting Experimental Evaluations which have external validity.

♦ Because of intrusion into operations of program in order to conduct experiment, also potential problems with obtaining internal validity.

♦ Issue of sample sizes and statistical power for conducting within-site analysis.
References


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1975-76 Overall</td>
<td>$300^*</td>
<td>$129^*</td>
<td>$586^*</td>
<td>$740 to 778^*</td>
<td>705^* to 762^*</td>
<td>$840^* to 949^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Women</td>
<td>500^*</td>
<td>400^*</td>
<td>534^*</td>
<td>17 to 136</td>
<td>157 to 691^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Men</td>
<td>200^*</td>
<td>(4)</td>
<td>500^*</td>
<td>677^* to 671^*</td>
<td>779^* to 810^*</td>
<td>659^* to 703^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Women</td>
<td>600^*</td>
<td>336^*</td>
<td>782^*</td>
<td>426 to 671^*</td>
<td>779^* to 810^*</td>
<td>659^* to 703^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Men</td>
<td>200 (104)</td>
<td>658^*</td>
<td>117 to 211</td>
<td>(273)^ to 69</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Classroom Training</td>
<td>350^*</td>
<td>287^*</td>
<td>740^*</td>
<td>--</td>
<td>--</td>
<td>800^* to 1,300^*</td>
<td>200 (690^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Women</td>
<td>550^*</td>
<td>257^*</td>
<td>740^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Men</td>
<td>400^*</td>
<td>--</td>
<td>--</td>
<td>63 to 205</td>
<td>295 to 354^*</td>
<td>315 to 451^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Women</td>
<td>500^*</td>
<td>--</td>
<td>--</td>
<td>(543)^ to (457)</td>
<td>(440) to (120)</td>
<td>300</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Men</td>
<td>200^*</td>
<td>--</td>
<td>--</td>
<td>426 to 633^*</td>
<td>245 to 301</td>
<td>206 to 369^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Women</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>582^ to 773</td>
<td>102^ to 185</td>
<td>(571)^ to (99)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Men</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>On-the-Job Training</td>
<td>850^*</td>
<td>531^*</td>
<td>1,091^*</td>
<td>80 to 382</td>
<td>701^ to 724^*</td>
<td>190 to 318</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Women</td>
<td>550^*</td>
<td>--</td>
<td>--</td>
<td>80 to 382</td>
<td>701^ to 724^*</td>
<td>190 to 318</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>White Men</td>
<td>750^*</td>
<td>--</td>
<td>--</td>
<td>616^ to 756^*</td>
<td>995 to 1,231^*</td>
<td>(200)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Women</td>
<td>1,200^*</td>
<td>--</td>
<td>--</td>
<td>1,360^ to 1,549^*</td>
<td>223^ to 244</td>
<td>564^ to 587^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Minority Men</td>
<td>1,150^*</td>
<td>--</td>
<td>--</td>
<td>2,053^ to 2,857^*</td>
<td>722^ to 812^*</td>
<td>454^ to 750^*</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Women</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Men</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

**Source:** Barmun, 1987. Sources for estimates are listed in the references at the end of this report.

**Notes:**

- Estimates are for all adult participants except as otherwise indicated.
- All estimates are in post-program year dollars except for Bloom & Mclaughlin estimates, which are in 1980 dollars.
- Missing entries indicate that impact estimates were not calculated.
- Numbers in parentheses are negative impact estimates.
- *Denotes statistical significance at the 5 percent level.
- Estimates are for all adult participants except as otherwise indicated.
TABLE 1.3
SUMMARY OF RFP AND FINAL RESEARCH DESIGNS

<table>
<thead>
<tr>
<th>Feature</th>
<th>RFP Plan</th>
<th>Final Research Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>Up to 20, chosen to statistically represent the JTPA system</td>
<td>16, chosen to illustrate the diversity of the JTPA system</td>
</tr>
<tr>
<td>Sample</td>
<td>Up to 30,000 adults and youth eligible for Title IIA</td>
<td>20,606 adults and out-of-school youth eligible for Title IIA</td>
</tr>
<tr>
<td>Evaluation of JTPA &quot;As Is&quot; with Little Change in the Program</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Evaluation of JTPA as a Whole</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Evaluation of Specific Treatments</td>
<td>Focus was on specific activities, such as OJT, classroom occupational training, and job search assistance</td>
<td>Focus on the types of combinations and sequences actually provided in JTPA, including categories of activities anchored on OJT and classroom occupational training</td>
</tr>
<tr>
<td>Services for Which the Control Group Is Eligible</td>
<td>Services in the community not funded by JTPA</td>
<td>Services in the community not funded by JTPA</td>
</tr>
<tr>
<td>Phase of Site Selection Process</td>
<td>Total SDAs Contacted</td>
<td>SDAs Participating</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Date of Initial Contact:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 1: Initial Design(^b)/Probabilistic(^c) Selection (Before 1/6/87)</td>
<td>83(^d)</td>
<td>5</td>
</tr>
<tr>
<td>Phase 2: Initial Design/Expanded Recruitment(^e) (Between 1/6/87 and 4/30/87)</td>
<td>61</td>
<td>4</td>
</tr>
<tr>
<td>Phase 3: Final Design(^b)/Expanded Recruitment (After 4/30/87)</td>
<td>85</td>
<td>7</td>
</tr>
<tr>
<td><strong>Date of Final Decision:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 1: Initial Design/Probabilistic Selection (Before 1/6/87)</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>Phase 2: Initial Design/Expanded Recruitment (Between 1/6/87 and 4/30/87)</td>
<td>61</td>
<td>3</td>
</tr>
<tr>
<td>Phase 3: Final Design/Expanded Recruitment (After 4/30/87)</td>
<td>120</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>229</td>
<td>16</td>
</tr>
</tbody>
</table>

NOTES: 
\(^a\)SDAs were dropped because they were in the midst of administrative reorganization; they were facing a state takeover because of performance problems; their program configuration could not be accommodated within the research design; or they were too geographically dispersed or served too few people. 
\(^b\)Summarized in Table 4.4 
\(^c\)The term is used because every SDA in a given category would have an equal probability of being selected into the sample. 
\(^d\)Seventy-three of these SDAs were contacted by MDRC under the probabilistic selection process. The other ten were not identified as priority SDAs under the probabilistic selection process, so no substantive discussions were held during Phase 1. In Phase 2, when the recruitment procedures were changed, MDRC recontacted most of these SDAs. 
\(^e\)Under expanded recruitment MDRC was allowed to recruit sites under any given category without regard to the probabilistic selection process.
<table>
<thead>
<tr>
<th>Concern</th>
<th>Percent of SDAs Citing the Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethical and Public Relations Implications of:</td>
<td>61.8</td>
</tr>
<tr>
<td>Random Assignment in Social Programs</td>
<td></td>
</tr>
<tr>
<td>Denial of Services to Controls</td>
<td>54.4</td>
</tr>
<tr>
<td>Potential Negative Effect of Creation of a Control Group on Achievement of Client Recruitment Goals</td>
<td>47.8</td>
</tr>
<tr>
<td>Potential Negative Impact on Performance Standards</td>
<td>25.4</td>
</tr>
<tr>
<td>Implementation of the Study When Service Providers Do Intake</td>
<td>21.1</td>
</tr>
<tr>
<td>Objections of Service Providers to the Study</td>
<td>17.5</td>
</tr>
<tr>
<td>Potential Staff Administrative Burden</td>
<td>16.2</td>
</tr>
<tr>
<td>Possible Lack of Support by Elected Officials</td>
<td>15.8</td>
</tr>
<tr>
<td>Legality of Random Assignment and Possible Grievances</td>
<td>14.5</td>
</tr>
<tr>
<td>Procedures for Providing Controls with Referrals to Other Services</td>
<td>14.0</td>
</tr>
<tr>
<td>Special Recruitment Problems for Out-of-School Youth</td>
<td>10.5</td>
</tr>
<tr>
<td>Sample Size</td>
<td>228</td>
</tr>
</tbody>
</table>

**SOURCE:** Based on responses of 228 SDAs contacted about possible participation in the National JTPA Study.

**NOTES:** Concerns noted by fewer than 5 percent of SDAs are not listed. Percentages may add to more than 100.0 because SDAs could raise more than one concern.
RANDOM ASSIGNMENT MODEL FOR THE NATIONAL JTPA STUDY

Recruitment → Eligibility Determination → Assessment → Recommended For:

Other Services Activity Category → Random Assignment

Classroom Teaching Activity Category → Random Assignment

Out Activity Category → Random Assignment

Control Group → Services

Control Group → Services

Control Group → Services

-19-
FIGURE 4.1

RANDOM ASSIGNMENT TO TREATMENTS
OR TO A CONTROL GROUP

Recruitment

Eligibility
Determination

Random Assignment

Service #1

Service #2

Service #3

Control Group
No JTPA Services
TABLE 4.2

ACTIVITIES AVAILABLE TO APPLICANTS ASSIGNED TO EACH TREATMENT CATEGORY

<table>
<thead>
<tr>
<th>Assigned Treatment Category</th>
<th>On-the-Job Training (OJT)</th>
<th>Classroom Training in Occupational Skills (CT-OS)</th>
<th>Combined OJT and CT-OS (Customized Training)</th>
<th>Basic Education</th>
<th>Job Search Assistance/Job Placement</th>
<th>Work Experience</th>
<th>Other Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-the-Job Training</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Classroom Training in</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Occupational Skills</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Other Activities*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

NOTE: *This category is intended for applicants recommended for activities other than OJT or classroom training or applicants recommended for both OJT and classroom training. There is a ceiling on the proportion of applicants eligible for this category; the limit is negotiated individually with each SDA.
<table>
<thead>
<tr>
<th>Target Group</th>
<th>OJT</th>
<th>CT-OS</th>
<th>Other Services</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult Males</td>
<td>3,190</td>
<td>1,592</td>
<td>2,078</td>
<td>6,860</td>
</tr>
<tr>
<td>Adult Females</td>
<td>2,672</td>
<td>3,417</td>
<td>1,980</td>
<td>8,069</td>
</tr>
<tr>
<td>Out-of-School Youth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1,571</td>
<td>2,097</td>
<td>2,009</td>
<td>5,677</td>
</tr>
<tr>
<td>Minority</td>
<td>1,027</td>
<td>981</td>
<td>873</td>
<td>2,881</td>
</tr>
<tr>
<td>Total</td>
<td>7,433</td>
<td>7,106</td>
<td>6,067</td>
<td>20,606</td>
</tr>
</tbody>
</table>

NOTE: Sample includes 1,364 persons who were randomly assigned to the treatment or control group at a ratio of 3:1 or 6:1.
TABLE 5.9
RESEARCH SAMPLE FOR THE NATIONAL JTPA STUDY, BY SDA

<table>
<thead>
<tr>
<th>SDA</th>
<th>Target Sample</th>
<th>Actual Sample</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital Area, MS (Jackson)</td>
<td>1,220</td>
<td>1,478</td>
<td>+258</td>
</tr>
<tr>
<td>Concentrated Employment Program, MT (Butte)</td>
<td>825</td>
<td>683</td>
<td>-142</td>
</tr>
<tr>
<td>Coosa Valley, GA (Rome)</td>
<td>1,800</td>
<td>1,840⁺</td>
<td>+40</td>
</tr>
<tr>
<td>Corpus Christi/Nueces County, TX</td>
<td>1,500</td>
<td>1,609⁺</td>
<td>+109</td>
</tr>
<tr>
<td>Crawford/Hancock/ Marion/ Wyandot Counties, OH</td>
<td>1,150</td>
<td>1,154</td>
<td>+4</td>
</tr>
<tr>
<td>East Central Iowa (Cedar Rapids)</td>
<td>2,963</td>
<td>498</td>
<td>-2,465</td>
</tr>
<tr>
<td>Greater Omaha, NE</td>
<td>1,600</td>
<td>1,362</td>
<td>-238</td>
</tr>
<tr>
<td>Heartland, FL (Lakeland)</td>
<td>4,850</td>
<td>597</td>
<td>-4,253</td>
</tr>
<tr>
<td>Jersey City, NJ</td>
<td>1,600</td>
<td>1,686⁺</td>
<td>+86</td>
</tr>
<tr>
<td>Larimer County, CO (Fort Collins)</td>
<td>1,200</td>
<td>1,027</td>
<td>-173</td>
</tr>
<tr>
<td>Macon/De Witt Counties, IL (Decatur)</td>
<td>750</td>
<td>471</td>
<td>-279</td>
</tr>
<tr>
<td>Northeast, IN (Fort Wayne)</td>
<td>3,600</td>
<td>3,608</td>
<td>+8</td>
</tr>
<tr>
<td>Northwest, MN (Crookston and Thief River Falls)</td>
<td>550</td>
<td>560</td>
<td>+10</td>
</tr>
<tr>
<td>Oakland, CA</td>
<td>1,065</td>
<td>1,072⁺</td>
<td>+7</td>
</tr>
<tr>
<td>Providence/Cranston, RI</td>
<td>1,750</td>
<td>1,759⁺</td>
<td>+9</td>
</tr>
<tr>
<td>Springfield, MO</td>
<td>2,000</td>
<td>1,202</td>
<td>-798</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20,423</strong></td>
<td><strong>20,606</strong></td>
<td><strong>-7,817</strong></td>
</tr>
</tbody>
</table>

*Some persons at this site were randomly assigned to the treatment or control group at a ratio higher than 2:1.*
Table A: Distribution of Specific JTPA Services Received by Those in (Randomly Assigned) Treatment Groups

<table>
<thead>
<tr>
<th>Specific Program Services Received</th>
<th>Percentage of Treatment/Control Group that Received Particular Services: ( ^a )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classroom Training</td>
</tr>
<tr>
<td></td>
<td>Treatment Group</td>
</tr>
<tr>
<td></td>
<td>Treatment Group</td>
</tr>
<tr>
<td>Never enrolled</td>
<td>27.60%</td>
</tr>
<tr>
<td>Classroom training in occupational skills</td>
<td>56.20%</td>
</tr>
<tr>
<td>Basic education ( ^b )</td>
<td>12.90</td>
</tr>
<tr>
<td>On-the-job training</td>
<td>3.80</td>
</tr>
<tr>
<td>Job search assistance</td>
<td>19.50</td>
</tr>
<tr>
<td>Work experience</td>
<td>4.00</td>
</tr>
<tr>
<td>Miscellaneous ( ^c )</td>
<td>9.90</td>
</tr>
<tr>
<td>Sample Size</td>
<td>4119</td>
</tr>
<tr>
<td>Other Services</td>
<td>Treatment Group</td>
</tr>
<tr>
<td></td>
<td>3064</td>
</tr>
<tr>
<td></td>
<td>4.70</td>
</tr>
<tr>
<td></td>
<td>19.70</td>
</tr>
</tbody>
</table>


a. Entries in Table are the percentages of treatment group who received that service. Note that members of a treatment group may have received more than one program service, so percentages do not sum to 100.0%.

b. "Basic Education" includes Adult Basic Education (ABE), high school or General Educational Development (GED) preparation, and English as a Second Language (ESL).

c. "Miscellaneous" included assessment, job-readiness training, customized training, vocational exploration, job shadowing, and tryout employment, among other services.
Professor V. Joseph Hotz

V. Designing Non-Experimental Evaluations of Social Programs: Alternative Methods of Estimation and the Associated Data Requirements

1. Expressing Model in Regression Format

Following most of the evaluation literature, \( Y_{on} \) is characterized as a function of a vector of observed characteristics, \( X_n \) (including a constant), and of unobserved characteristics, \( U_n \), which takes the following form,

\[
Y_{on} = X_n \beta + U_n
\]

(1.1)

where \( \beta \) is a vector of parameters to be estimated. For now, we assume that the training effect is invariant across individuals but not across time, so that \( \alpha_n = \alpha \). Let \( X_t = (X_{t1}, \ldots, X_{tT}) \), where \( T \) is the number of periods of data on \( X \) available for each observation. Then the observed outcomes may be written as

\[
Y_{it} = X_{it} \beta + d_i \alpha + U_{it}, \text{ for } t = 0, \ldots, T
\]

(1.2)

Finally, we assume that the vector \( X_i \) is independent of the unobservables, \( U_n \), in the sense that \( E(U_{it}|X_i) = 0 \) for all \( i,t \).

In this formulation of outcomes, the existence of selection bias (that \( E(Y_{on}|d_i=1) - E(Y_{on}|d_i=0) \neq 0 \)) is characterized by the existence of (statistical) dependence between \( d_i \) and \( U_n \). In particular:

\[
E(U_{it}|X_{it}, d_i) \neq 0
\]

(1.3)

1.1 Origins of Selection Bias

To illustrate how this stochastic relationship between \( d_i \) and \( U_n \) arises, suppose that the process governing selection into a training program (i.e., the decisions of individuals eligible for the program to apply and/or of program administrators to enroll applicants) can be characterized
in terms of an index function framework. Let the index, $IN_i$, be a function of both observed ($Z_i$) and unobserved ($V_i$) variables. (The vector $Z_i$ may include all of the variables in $X_i$.) For simplicity, we follow assume that this function is linear in a vector of observables, $Z_i$, and unobserved factors, $V_i$, i.e.,

$$IN_i = Z_i \gamma + V_i$$

(1.4)

where $\gamma$ is a vector of parameters. Then the $i^{th}$ individual’s training status is

$$d_i = \begin{cases} 
1 & \text{if and only if } IN_i > 0, \\
0 & \text{otherwise.} 
\end{cases}$$

(1.5)

$V_i$ is assumed to be independently and identically distributed across persons, where the distribution function of $V_i$ is denoted as $F(v_i) = Pr(V_i < v_i)$. Assuming that $V_i$ is distributed independently of $Z_i$, we may write:

$$Pr(d_i = 1|Z_i) = 1 - F(-Z_i \gamma) = p(Z_i)$$

which Rosenbaum and Rubin (1983) call the “propensity score.”

To the extent that either $Z_i$ and/or $V_i$ are correlated with $U_i$, selection bias will be present in nonexperimental settings, i.e., $E(U_i|X_i,d_i)$ will not be zero. As discussed in Heckman and Robb (1985), such correlation generally arises in models in which program participation decisions depend, in part, on the likely impact of the program on an individual’s subsequent outcomes. Its presence implies that

$$E(Y_i|d_i,Z_i) = X_i \beta + d_i \alpha_i + E(U_i|d_i,X_i) \neq X_i \beta + d_i \alpha_i$$

(1.6)

so that an ordinary (or nonlinear) least squares (OLS) regression of $Y_i$ on $X_i$ and $d_i$ will not yield consistent estimates of $\alpha_i$ (or $b$). (Why not?)

In order to characterize alternative non-experimental (quasi-experimental) design strategies, we present a series of alternative methods for estimating program impacts (primarily
\( E(\alpha_u|d_i=1) = \alpha' \). These methods differ in the types of data required for their implementation. They also differ in the types of assumptions they require to be true in order to eliminate selection bias. By considering those assumptions, one can get a sense of what estimators might be most appropriate in what situations.

****Add discussion of alternative data sources here. Include a table with "cross-section," "repeated cross-section," and "panel" data along vertical and "Participants" and "Non-participants" along horizontal.

2. Control Function Estimators for Use with Post-Program Data on Participants and Members of a Comparison Group

2.1 General Considerations

Consider the case where we have post-program data \((t > k)\) on both trainees (the treated) and a comparison group of individuals who did not receive training (the treatment). One class of methods to estimate the training (treatment) effect are control function estimators account for selection bias by adding a particular type of function of observable variables—denoted by \(h(d, X, Z, \pi)\), where \(\pi\) is a vector of parameters—to \(X\beta + d\alpha\), which adjusts for the dependence between \(d\) and \(U\) in order to control for this dependence and enable one to obtain an unbiased estimate of \(\alpha\) (or, more generally, \(E(\alpha_u|d_i=1)\)). In particular, the appropriate control function, \(h(d, X, Z, \pi)\), is just \(E(U|d, X, Z)\). For a particular choice of \(h(d, X, Z, \pi)\), the augmented outcome equation becomes

\[
Y' = X\beta + d\alpha + h(d, X, Z, \pi) + U'
\]  

(2.1)

where, when \(h(d, X, Z, \pi) = E(U|d, X, Z)\), \(E(U|d, X, Z) = 0\). The inclusion of the \(X\beta\) term, which include contemporaneous \(X\)'s, is to make sure one isolates the treatment effect for other factors which are changing in the post-program period and which may account for some of the variation in \(Y\). The inclusion of the term \(h(d, X, Z, \pi)\) is, as stated above, to control for the differ-
ences in selection into the treatment between the trainees (treated) and the members of the non-experimentally-based comparison group. Under the assumptions given above, least squares estimation of (2.1) using post-training \((t > k)\) data on program participants and comparison-group members will yield unbiased estimates of the training impact, \(\alpha_t\).

Alternative control functions have been proposed in the evaluation and econometric literature; different functional forms for \(h(d_i, X_i, Z_i, \pi)\) reflect different assumptions about the distribution of the unobservables \(U_i\) and \(V_i\), and/or the types of conditioning variables in \(h(\ )\). We briefly describe several specifications which have received substantial attention in the literature.

2.2 Selection on Observables

One variant within the control function class arises when the dependence between \(U_i\) and \(d_i\) is assumed to be due to the observed variables, \(Z_i\), influencing selection into the program. Under the selection-on-observables assumption, it follows that while

\[ E(U_i | d_i, X_i) \neq 0 \quad \text{and} \quad E(U_i | d_i, X_i, Z_i) \neq 0 \]

it is the case that

\[ E(U_i | d_i, X_i, Z_i) = E(U_i | X_i, Z_i). \]

In this case, controlling for a function of \(X_i\) and \(Z_i\) (but not \(d_i\)) solves the selection bias problem. As before, augmenting (1.2) with an appropriate control function, i.e.,

\[ Y_u = X_u \beta + d_i \alpha_t + h(X_i, Z_i) + U_i^* \] (2.2)

and utilizing least squares to estimate (2.2), will yield consistent estimates of \(\alpha_t\). Thus the selection bias problem (i.e., the correlation between \(U_i\) and \(d_i\)) can be eliminated by accounting for the observable factors which influence the selection process.

Alternative functional forms for \(h(\ )\) can be used to form members of the subclass of se-
lection-on-observable models. For example, Barnow, Cain and Goldberger (1980) consider the case where \( h(X_u, Z_i) \) is a linear function of \( X_u \) and \( Z_i \), e.g.,

\[
h(X_u, Z_i) = X_u \theta_u + Z_i \theta_u \tag{2.3}
\]

where \( \theta_u \) and \( \theta_u \) are parameter vectors.

There is an important special case of the control function estimator which uses a linear control function, namely, the regression discontinuity design quasi-experimental estimator which has frequently been used in the educational evaluation literature. The basis for the appropriate use of this estimator is in situations where the assignment of the program depends entirely on the \( Z \)'s. For example, it might be the case that admission to a remedial educational program is entirely based on the performance on a (screening) test. Suppose all children who score below some level \( Z_i < Z_a \) are admitted to the program and those above are not. Then one can use the linear control function estimator described above, using those individuals who were excluded as a comparison group. The key issue here is that the selection into the program (i.e., the determination of who receives the treatment) must be solely based on \( Z_i \).

An alternative but related strategy has been proposed by Rosenbaum and Rubin (1983). They propose the use of control functions which are expressed solely as a function of the propensity score, \( p(Z_i) = 1 - F(-Z_i \gamma) \). That is, the propensity score control function is given by

\[
h(X_u, Z_i) = p(Z_i) \tag{2.4}
\]

where \( p(Z_i) \) is estimated separately (via logit or probit analysis), predicted values are formed, and these predicted propensity scores are included as regressors in the earnings (outcome) equation. The latter is then estimated using least squares methods. Under the assumptions that selection is only on observables, the estimation on the augmented outcome equations will yield a consistent
estimator of $\alpha_t$ if the assumptions motivating this particular control function is an accurate representation of the selection process governing the available data.

2.3 The Mills Ratio (or "Heckman") Procedure

One commonly used control function estimator proposed by Heckman (1976), is based on the assumption that the joint distribution of $U_u$ and $V_i$ is bivariate normal. Under this set of assumptions, $h(d_uX_uZ_u, \pi) = E(U_u|d_uX_uZ_u)$ is proportional to the Mills ratio, i.e.,

$$h(d_uX_uZ_u, \pi) = \frac{\phi(Z_i\gamma)}{1 - \Phi(Z_i\gamma)} \tag{2.5}$$

where $\phi$ and $\Phi$ are the standardized normal density and distribution functions, respectively. Under the joint normality assumption, the inclusion of the Mills Ratio in the outcome equation (1.2), i.e.,

$$Y_u = X_u\beta + d_u\alpha_t + \theta_t \left( \frac{\phi(Z_i\gamma)}{1 - \Phi(Z_i\gamma)} \right) + U_u' \tag{2.6}$$

where $\theta_t$ is a parameter to be estimated.

A consistent estimate of $\alpha_t$ when (2.6) is estimated by least squares. In practice, a two-stage procedure is used since $\gamma$ is generally unknown and must also be estimated. In the first stage, $\gamma$ is consistently estimated by estimating the selection status ($d_i = 1$ or 0) using probit estimation methods and, in the second stage, (2.6) is estimated by least squares, using the consistent estimate of $\gamma$ from the first stage. (See Heckman (1976) for details.) Note that this estimator requires that the specification of the outcome equation (1.2), the selection process, (1.4) and (1.5), and the joint normality of $U_u$ and $V_i$ be consistent with the data to insure that the training impact, $\alpha_t$, is consistently estimated.
3. **Longitudinal Data Estimators**

3.1 **General Considerations**

In this section we consider cases in which data is available both before and after the occurrence of the training (treatment). We consider several estimators which characterize alternative quasi-experimental designs.

We begin with an estimator which only requires longitudinal data on the trainees (treated). We then consider cases where longitudinal data (or, in some cases, repeated cross-sections of data) are available on both program participants and comparison-group members which include observations on pre-program outcomes \( t < k \) and for the post-program period \( t > k \). In this case there exist a class of estimators for use in nonexperimental evaluation designs which invoke alternative assumptions about the nature of the correlation between \( U_i \) and \( V_i \), the unobserved factors entering the outcome function and selection rule, respectively. Based on these assumptions, the resulting longitudinal data estimators entail transformations of the outcome equation in (1.2) so that \( d_i \) is not correlated with the disturbance term in the resulting transformed equations.

Within this class of longitudinal estimators, three particular types have been widely used in the training evaluation literature. We consider each in turn. In all of the cases which follow, the method does not require complete specification of the distribution of \( U_i \) or \( V_i \) at the same time, assumptions on the intertemporal covariance structure of the disturbances must be invoked.

3.2 **Before and After Estimators**

Probably the most common quasi-experimental design is to compare the outcomes of trainees (the treated group) *after* training (treatment) with their outcomes *before* receipt of treatment and attribute the difference to the effect of the treatment. Note that in this case, data is only
needed for the treated ($E$); no comparison group is needed to use this method. More formally, the estimator can be characterized as follows. Let

$$d_{it}^* = \begin{cases} 1 & \text{if has experienced training in period } t' \geq t, \\ 0 & \text{otherwise}. \end{cases}$$

Furthermore, assume that:

$$\beta_i = \beta; \alpha_{it} = \alpha;$$

that there are no $X$'s in the model, only an intercept, and that only data on trainees ($d_{it} = 1$) is used. Then the model of outcomes is given by:

$$Y_{it} = \beta_0 + \alpha d_{it}^* + U_{it} \quad (3.1)$$

In this situation, then $\alpha$ is an estimator of $E(\alpha|d_{it}=1) = \alpha^*$. In fact, it is equivalent to $\bar{Y}_{post} - \bar{Y}_{pre}$.

Given the assumptions made, this estimator is a consistent of $E(\alpha|d_{it}=1)$. The issue is the appropriateness of the assumptions. These assumptions will be violated, and thus the appropriateness of this estimator for measuring $E(\alpha|d_{it}=1)$ is in question, if there are either (time) varying changes in the environment, maturation of the individuals and/or heterogeneity in the impact of the treatment. In such cases, the lack of a control group seriously inhibits the ability to measure $E(\alpha|d_{it}=1)$ with this estimator.

3.3 The Fixed Effect (or First Difference) Estimator

This estimator has been used in a number of nonexperimental evaluations of the impacts of training on earnings. (In fact, its application requires that the outcome variable be continuous; application to discrete or categorical outcome variables, such as labor force participation and/or welfare recipiency is not direct.) At a minimum, this estimator requires the presence of observations on outcomes prior to and after the receipt of training for trainees and comparison group
members. This specification is motivated by a model in which $U_u$ is assumed to have the form

$$U_u = \phi_u + \nu_u,$$  \hspace{1cm} (3.2)

where $\phi_u$ is a zero mean, person-specific component or "fixed effect," $\nu_u$ is a zero mean random component independent of all other values of $(t \neq t')$ and of $\phi_u$. In this specification, individuals are assumed to have permanent differences in the levels of their outcomes and these permanent components, the $\phi_i's$, are assumed to directly influence an individual's program participation decision. Holding constant this person-specific component, it follows that there should be no further dependence between $V_i$ (and thus $d_i$) and $U_u$. Thus, the conditional expectation of the difference in a pre- and post-training set of $U_u$'s does not depend on $d_u$, i.e.,

$$E(U_u - U_{u'}|d_u, Z_i) = 0, \text{ for all } t, t', t > k > t'.$$  \hspace{1cm} (3.3)

Consistent estimates of the impact of training can be obtained by regressing the difference between $Y_u$ and $Y_{u'}$ on $d_i$ and $X_{u'}$, i.e., transforming (1.2) by first differencing. The transformed equation is given by

$$Y_u - Y_{u'} = d_i \alpha_i + X_u \beta_i - X_{u'} \beta_i + (\nu_u - \nu_{u'}), \text{ for } t > k > t'.$$  \hspace{1cm} (3.4)

Under the assumptions of the fixed effect model, estimating (3.4) by least squares yields a consistent estimator of $\alpha_i$.

### 3.4 The Random Growth Estimator

The random growth estimator is a generalization of the fixed effect estimator. It is motivated by a special case of a more general class of models in which $U_u$ is assumed to have a factor structure.\footnote{See Pudney (1982) for discussion of this more general class.} In particular, suppose $U_u$ is of the following form,

$$U_u = \phi_u + t \phi_{2u} + \nu_u$$  \hspace{1cm} (3.5)
where \( \phi_{1i} \) is as before and \( \phi_{2i} \) is a person-specific growth rate for the outcome variable \( Y_{it} \). One also assumes that \( (\phi_{1i}, \phi_{2i}) \) have zero means and finite variances and are independent of \( \nu_{it} \) for all \( i \) and \( t \). In this model, individual outcomes are allowed to differ both in levels and in rates of growth, and the individual’s participation in a program is assumed to depend on both components, thus inducing dependence between \( U_{it} \) and \( d_i \).

Given the form of \( U_{it} \), one can transform the outcome equation to eliminate \( \phi_{1i} \) and \( \phi_{2i} \) from the outcome equation. Values of \( Y_{it} \) in two consecutive pre-training periods to proxy these components, one obtains the following transformed outcome equation,

\[
[Y_{it} - Y_{it-1}] - (t - t')[Y_{it'} - Y_{it'-1}] = d_i \alpha_i + [X_{it} \beta_{1i} - X_{it} \beta_{1i'}] - (t - t')[X_{it'} \beta_{1i'} - X_{it'-1} \beta_{1i'-1}] + [\nu_{it} - \nu_{it'}] - (t - t')[\nu_{it'} - \nu_{it'-1}] \tag{3.6}
\]

where \( t > k > t' \). Again, the transformation of the outcome equation results in \( d_i \) being uncorrelated with the resulting composite disturbance in (3.6). Under standard conditions, estimation of (3.6) by least squares will yield consistent \( \alpha_i \).

3.5 The Autoregressive Disturbance Estimator

Another commonly used longitudinal estimator is based on the assumption that the outcome disturbances, \( U_{it} \), have an autoregressive structure. In the case of a first order autoregressive structure,

\[
U_{it} = \rho U_{it-1} + \nu_{it}, \tag{3.7}
\]

where \( \rho \) is a parameter (assumed to not equal \( \pm 1 \)) and \( \nu_{it} \) is a mean zero independently distributed random disturbance. One further assumes that the correlation between \( d_i \) (and thus \( V_t \) and \( U_{it} \) can be accounted for by the serial correlation in the \( U_{it} \)'s. (Such a correlation might arise in situations in which eligible individuals make their decisions about entering a training program based on their
outcome experiences—and anticipated future outcomes if they do not receive training—and the disturbances which influence these outcomes are have an autoregressive structure.) Once again a transformation of the outcome equation can be used to eliminate the selection bias problem. In the case of a first order autoregressive disturbance, using pre-program data from period \( t' (t' < k) \) and post-program data from period \( t (t > k) \) the autoregressive structure of \( U_n \) (and ) leads to the following (standard) transformation:

\[
Y_n = \rho^{t'-t} Y_{n'} + X_n \beta_t - \rho^{t'-t} X_n \beta_{t'} + (1 - \rho^{t'-t}) \alpha_t d_t + \left[ \sum_{j=0}^{t-t'+1} \rho^j v_{t,t'-j} \right]
\]

(3.7)

(Transformations for higher order autoregressive disturbance structures are similar to that in (3.7).) Under the assumptions for the autoregressive model, nonlinear least squares methods applied to (3.7) will yield consistent estimates of \( \alpha_t \).

4. Instrumental Variables (IV) Estimators

Instrumental Variables (IV) estimators presume the existence of variables (elements of \( Z_n \), for example) which are independent of the outcome equation disturbance, \( U_n \), but are correlated with the training status of individuals, \( d_t \). Under conditions given in Heckman and Robb (1985), such instrumental variables can be used to construct a consistent estimator of \( \alpha_t \), using relatively standard IV estimation procedures. Note that with IV estimators, no explicit distributional assumptions about \( U_n \) or \( V_n \) need to be made. Generically, the IV procedure can be characterized as follows. Let \( W_i \) represent an element of \( Z_i \) which has the following properties:

\[
\text{Cov}(W_i, d_i) \neq 0 \text{ but } E(W_i, U_n) = 0
\]

Under standard conditions, there exists a population projection relationship,

\[
d_i = \pi_0 + \pi_i W_i + e_i \tag{4.1}
\]

where \( E(e_i) = 0 \) and \( E(e_i, W_i) = 0 \). (In practice, this projection can be estimated with a linear re-
gression of \( d_i \) on \( W_i \).) An IV estimator of \( \alpha \) can be obtained by forming a predicted value of \( d_i \) by:

\[
\hat{d}_i = \pi_0 + \pi_1 W_i \tag{4.2}
\]

and including this predicted value in the outcome equation as a separate regressor:

\[
Y_i = X_i \beta + \hat{d}_i \alpha + U_i', \tag{4.3}
\]

where \( U_i' = U_i - \epsilon_i \). It can be shown that for large samples, \( U_i' \) is uncorrelated with \( \hat{d}_i \). Thus equation (4.3) can be estimated with standard regression procedures to obtain consistent estimates of \( \alpha \). The intuition for why IV estimators eliminate the selection-bias problem is that by forming a prediction equation for \( d_i \) as a function of \( W_i \) enables one to create a new variable, \( \hat{d}_i \), whose variation in identifying the effect of training (the treatment) is not confounded with the determinants of \( Y_{oi} \). That is, using this predicted value, rather than \( d_i \), works—given an instrument, \( W_i \), that meets the conditions noted above—because the variation contained in \( \hat{d}_i \) has purged \( d_i \) of that part which was correlated with \( U_i \).

Another type of estimator, closely akin to the IV procedure, can be formed by substituting the propensity score, \( p(Z_i) \), for the training dummy variable, \( d_i \). Unlike the IV estimation procedure just described, this latter estimator requires not only that \( Z_i \) is distributed independently of \( V_i \), but also knowledge of the form of the distribution function for \( V_i \). Then instead of projecting \( d_i \) on an instrument (\( W_i \)), the first stage of estimation entails estimating the coefficients on \( Z_i \) (i.e., \( \gamma \)) in the selection rule, using conventional discrete choice estimation methods such as probit or logit analysis and forming the predicted probability of \( d_i = 1 \), i.e., \( \hat{p}(Z_i) = 1 - F(-Z_i \hat{\gamma}) \). This latter function can then be used as an instrument for \( d_i \) in the outcome equation (1.2), i.e.,

\[
Y_i = X_i \beta + \alpha \hat{p}(Z_i) + U_i' \tag{4.4}
\]

and least squares can be applied to (4.4) to consistently estimate \( \alpha \) (see Heckman (1978)). Note
that as an alternative to a two-step estimation procedure, nonlinear regression strategies can be used to estimate (1.2) and the \( F(-Z_i \gamma) \) jointly.

5. Statistical Matching Procedures and Non-Parametric Methods

Statistical matching procedures for estimating program impacts in nonexperimental designs construct a matched sample for the program participants using data from a comparison group of nonparticipants and use the differences in post-program outcomes between participants and their comparison group match to estimate the program impact. The potential advantage of matching procedures is that they do not require specification of functional forms for outcome equations to estimate impacts. At the same time, for matching procedures to produce unbiased impact estimates, it must be the case that the observable variables used to construct the matched pairs adequately account for the differences in latent outcomes, \( Y_{0n} \), between participants and nonparticipants. More precisely, let \( Z_i \) be the characteristics along which participants are matched with observations from a comparison group. Matching methods are predicated assumption that conditioning on these characteristics, the difference between the outcomes of participants and comparison group members is, on average, zero. That is,

\[
E(Y_{0i} - Y_{ti} | Z_i, Z_j) = E(\alpha_n | \delta_i = 1)
\]  

(5.1)

for \( i > k \) where observation \( i \) is from the participant sample (\( E \)) and observation \( j \) is a member of a nonparticipant comparison group (\( N \)). Thus, matching procedures—as is the case for selection-on-observables control function procedures—require that analysts can select (and measure) the appropriate variables upon which to base matching.

A number of alternative variants of matching techniques can be found in the evaluation lit-
erature;\(^3\) some of the estimates contained in the table of estimates of CETA programs taken from Barnow (1985), which was in an earlier handout.\(^3\) In what follows, we outline the elements of these procedures.

Generically, the creation of matched-pair samples all start with a set of characteristics \((Z_i)\) and find a member of the comparison-group who is similar to each program participant according to a particular metric. To determine similarity, a distance function is used to measure how “close”—or “similar”—two observations are with respect to this set of characteristics. Let \(m(Z_i,Z_j)\) denote such a distance function for observations \(i\) and \(j\). Different distance functions have been used in the literature for determining matches, e.g., the discriminant function and the Mahalanobis metric.\(^4\) The matched-pair for program participant \(i\) is formed by choosing that comparison-group member \(j^*\) which minimizes \(m(Z_i,Z_j)\) for all \(j \in N\). Denote the resulting matched-pair sample by \((\{i,j^*(i)\}, \text{for all } i \in E)\). Using this sample, an estimate of the average training impact can be formed either by a simple mean of the pairs, i.e.,

\[
\hat{\alpha}_i = \frac{1}{N_g} \sum_{i=1}^{N_g} (Y_{ui} = Y_{i^*(i)})
\]

Alternatively, some analysts (e.g., Rubin (1979)) propose using the matched-pair sample to estimate a linear (in \(X\)) version of the outcome equation in (1.2) by ordinary least squares to estimate

\(^2\)In particular, see the work of Rubin (1973a, 1973b, 1979, 1980) and Rosenbaum and Rubin (1983, 1984, 1985) for discussions of matching procedures. See also the work of Barros (1986) for a discussion of the relationship between matching and nonparametric estimation procedures.

\(^3\)See the studies by Dickinson, et al. (1894), Westat (1984) and Fraker and Maynard (1984).

\(^4\)Using the discriminant function,

\[
m(Z_i,Z_j) = (Z_i - Z_j)D D(Z_i - Z_j)',
\]

where \(D = (\bar{Z}_T - \bar{Z}_N)S^{-1}\) with \(\bar{Z}_T\) being the mean \(Z\) vector for the program participant (trainee) sample, \(\bar{Z}_N\) being the mean \(Z\) vector for the nonparticipant comparison group, and \(S\) being the pooled within-sample covariance matrix for \(Z\). Using the Mahalanobis metric, the distance function is given by

\[
m(Z_i,Z_j) = (Z_i - Z_j) S^{-1}(Z_i - Z_j)',
\]
the program impact, net of $X$ and $Z$. Under the assumptions outlined above—namely those equivalent to the selection-on-observables conditions—the above procedures represent consistent estimators of the impact of the training program on outcome $Y_u$ for the trained.

As noted above, the essential issue for matching procedures is the characteristics used to select matches. Rosenbaum and Rubin (1983, 1984, 1985) note that a particularly good choice for a characteristic upon which to match is the *propensity score*, i.e., predicted probability of the program participation, $\hat{p}(Z_i) = 1 - F(-Z_i \hat{\gamma})$. While its use still requires that the appropriate set of variables be included in $\hat{p}(Z_i)$ in order for condition (5.1) to hold, the propensity score provides a sufficient statistic for these variables when constructing matched samples and reduces to one the dimensionality of the characteristics upon which matching takes place. To implement this strategy, one first estimates $\gamma$ using discrete choice methods like probit or logit and forms $\hat{p}(Z_i)$ for each observation in the participant and comparison-group samples, and then constructs matched-pairs using the propensity score as the basis for matching.

Finally, we note that a variant on the matching procedure is the use of *nonparametric regression methods* for deriving program impact estimates. As noted by Barros (1986), matching techniques are similar to nonparametric regression in that no explicit assumptions need be made about the functional form of the outcome equations. The primary difference between the two procedures is that matching selects a single observation from the comparison group for each program participant, namely the one which is nearest to the participant according to the distance metric used. The nonparametric regression procedure forms a "composite match" for each participant by taking a weighted average of outcomes for more than one individual (and possibly all individuals) in the comparison group, where the weights are inversely related to the distance from the partici-
pant based on the Z vector. While both are justified under similar conditions, i.e., condition (5.1), little is known about the relative efficiency properties of the two procedures.
References


VI. Choosing Among Alternative Nonexperimental Estimators in Impact Analysis:
The Case of Evaluating Manpower Training Programs

I. FOCUS OF LECTURE

A. Typical Approaches to Evaluation of Manpower Training (Other Social) Programs

- **Experimental Evaluation:**
  
  - Compare outcomes of Trainees with those for sample of Nontrainees (Control Group) formed using random assignment.
  
  - Use simple statistical techniques (Difference in Means) to measure impact.

- **Nonexperimental Evaluation:**
  
  - Compare outcomes of Trainees with Nontrainees (Comparison Group) consisting of those judged to be "comparable" to trainees except received no training.
  
  - Use matching and/or statistical adjustment procedures to account for discrepancies in observed and unobserved characteristics between two groups which might distort outcome comparisons.
  
  - Failure to properly control for differences in characteristics leads to selection bias.

B. Recent Controversy over Use of Nonexperimental Evaluation Methods

- Recent influential studies show that alternative nonexperimental estimators of impact produce a wide range of impact estimates and differ from experimental estimates. [LaLonde (1986) & Fraker and Maynard (1984, 1987) studies of National Supported Work (NSW) data]

- Conclusions drawn:
  
  "...estimates of program effects that are based on nonexperimental comparisons can be subject to substantial misspecification uncertainty" [Burtless and Orr (1986)]

  "...the nonexperimental procedures may not accurately estimate the true program impacts. In particular, there does not appear to be any formula [using nonexperimental methods] that researchers can confidently use to replicate the experimental results ... the findings are further evidence that the current skepticism surrounding the results of nonexperimental evaluations is justified." [LaLonde and Maynard (1987)]

  "...randomized clinical trials are necessary to determine program effects" [Ashenfelter and Card (1985)]

  "...experiments appear to be the only method available at this time to overcome the limitations of nonexperimental evaluations" [Barnow (1987)]

- **Given that experiments may not be feasible for conducting all evaluations, need to use nonexperimental methods. But are they credible??**
C. Two Fallacies underlying Negative Assessments of Nonexperimental Evaluation Methods

1. The "Sensitivity to Alternative Methods" Fallacy:

Nonexperimental estimation procedures should produce approximately the same program estimate. "Good" nonexperimental estimators are those which yield robust estimates across alternative specifications.

The Fallacy: If one finds there are not differences in estimated impact across alternative nonexperimental methods, this simply indicates that there is no selection bias. Sensitivity is evidence that there is selection bias. In the presence of selection bias, different estimators are based on different assumptions, will not, in general, produce the same estimate. Only the one (or subset) invoking assumptions consistent with the underlying (nonrandom) selection process will produce the correct estimate.

2. The "No Way to Choose" Fallacy:

The assumptions made in conducting nonexperimental studies are arbitrary and there is not objective way to choose among estimators making different assumptions.

The Fallacy: There are "data sensitive" testing strategies which can be used to examine the validity of the assumptions (model specifications) associated with alternative nonexperimental estimators. In the absence of experimental data, one is limited to testing restrictions implied by a particular model. Such tests may be helpful given the frequent use of methods (models) which contain such are testable restrictions. In the presence of experimental data, additional tests of the appropriateness of a particular nonexperimental estimator are available.

This paper illustrates the use of such testing strategies to re-examine the conclusions about the use of nonexperimental estimators drawn from data on the National Supported Work Demonstration (NSW) Study.
II. THE PROBLEM OF SELECTION BIAS

Let:

- \( Y^*_it \) be earnings of individual \( i \) in period \( t \) in the absence of training,
- \( Y_{it} \) be the observed value of earnings for individual \( i \) at time \( t \),
- \( d_i = 1 \) if a person receives training and = 0 otherwise,
- \( \alpha_{it} \) be the impact of training on person \( i \) at time \( t \),

and adopt convention that training occurs in period \( k \). Then:

\[
Y_{it} = Y^*_it + d_i \alpha_{it}, \quad t > k,
\]

\[
Y_{it} = Y^*_it, \quad t < k.
\] (2.1)

We focus on estimating the mean impact of training on the trained.

\[
E(\alpha_{it}|d_i = 1) = E(Y_{it} - Y^*_it|d_i = 1),
\] (2.2)

The mean post-program earnings of trainees is

\[
E(Y_{it}|d_i = 1) = E(\alpha_{it}|d_i = 1) + E(Y^*_it|d_i = 1).
\] (2.3)

The mean post-program earnings of non-trainees is

\[
E(Y_{it}|d_i = 0) = E(Y^*_it|d_i = 0).
\] (2.4)

The difference in mean earnings between trainees and non-trainees is

\[
E(Y_{it}|d_i = 1) - E(Y_{it}|d_i = 0) = E(\alpha_{it}|d_i = 1) + \left( E(Y^*_it|d_i = 1) - E(Y^*_it|d_i = 0) \right).
\] (2.5)

Selection Bias is present if

\[
E(Y^*_it|d_i = 1) = E(Y^*_it|d_i = 0),
\] (2.6)

where comparison group members would not be trainees \((d_i = 0)\). In the case of random assignment of persons to treatment,

\[
E(Y^*_it|d_i = 1) = E(Y^*_it|d_i = 0) = E(Y^*_it),
\] (2.7)
III. ALTERNATIVE NONEXPERIMENTAL ESTIMATORS FOR MEASURING THE IMPACT OF TRAINING ON EARNINGS IN THE PRESENCE OF NONRANDOM ASSIGNMENT

ASSUMPTIONS FOR APPLICATION IN A REGRESSION FRAMEWORK:

A. The Earnings Equation:

\[ Y_{it} = X_{it}'\beta + U_{it} \]  \hspace{1cm} (3.1)

where \( X_{it} \) vector of observed characteristics, \( U_{it} \) unobserved characteristics, and \( \beta \) vector of parameters.

Thus:

\[ Y_{it} = X_{it}'\beta + d_1\alpha_t + U_{it} \text{, for } t = 0, \ldots, T. \]  \hspace{1cm} (3.2)

We assume that \( E(\alpha_{it}|d_1=1) = \alpha_t \) and that \( E(U_{it}|X_i) = 0 \) for all \( i, t \).

Selection Bias is present if

\[ E(U_{it}|d_1,X_i) \neq 0 \quad \text{so} \quad E(Y_{it}|d_1,X_i) = X_{it}'\beta + d_1\alpha_t. \]

Ordinary least squares regression of \( Y_{it} \) on \( X_{it} \) and \( d_1 \) yields inconsistent estimates of \( \alpha_t \) (or \( \beta \)).

B. Selection on Observables

Suppose:

\[ E(U_{it}|d_1,X_i) \neq 0 \quad \text{and} \quad E(U_{it}|d_1,X_i,Z_1) \neq 0 \]

but

\[ E(U_{it}|d_1,X_i,Z_1) = E(U_{it}|X_i,Z_1). \]  \hspace{1cm} (3.6)

Controlling for the observed selection variables \( (Z_1) \) solves the selection bias problem.
Therefore, using (3.6), one can form estimators by noting that:

$$E(Y_{it}|d_{it},X_{it},Z_{it}) = X_{it}\beta + d_{it}\alpha_t + E(U_{it}|X_{it},Z_{it})$$

(3.7)

assuming knowledge of the functional form of $E(U_{it}|X_{it},Z_{it})$.

- **Linear Control Function Estimators** [Barnow, Cain and Goldberger (1980)]:

  Assume $E(U_{it}|X_{it},Z_{it})$ is linear function of $X_{it}$ and $Z_{it}$.

  - **Variant I**: [Assuming $E(\alpha_{it}|d_{it}=1,X_{it},Z_{it}) = \alpha_t$]

    $$Y_{it} = C_{it}\delta_t + d_{it}\alpha_t + \tilde{U}_{it}$$

    (3.8)

    where $C_{it}$ denotes the vector of all variables included in either $X_{it}$ or $Z_{it}$, \(\tilde{U}_{it} = U_{it} - E(U_{it}|d_{it},C_{it}) = U_{it} - E(U_{it}|C_{it})\) and $\delta_t$ is a parameter vector.

    - **Variant II**: [Assuming $E(\alpha_{it}|d_{it}=1) = C_{it}\theta_t$]

      $$Y_{it} = C_{it}\delta_t + d_{it}(C_{it}\theta_t) + \tilde{U}_{it}.$$  

      (3.8')

    - **Consistent** estimators of $\alpha_t$ can be obtained by ordinary least squares.

**C. Selection on Unobservables**

- Suppose dependence between $d_{it}$ and $U_{it}$ not eliminated even after controlling for $Z_{it}$.

$$E(U_{it}|d_{it},X_{it}) \neq 0 \text{ and } E(U_{it}|d_{it},X_{it},Z_{it}) \neq E(U_{it}|X_{it},Z_{it}).$$

(3.9)

- **Fixed Effect (or First Difference) Estimator**:

  - Suppose

  $$U_{it} = \phi_{it} + v_{it}$$

  where $\phi_{it}$ is a zero mean, person-specific component and $v_{it}$ is random component. Selection assumed to depend on permanent component $\phi_{it}$.

  **Consistent** estimates of $\alpha_t$ are obtained by regressing $Y_{it} - Y_{it}'$ on $d_{it}$ and
\[ X_{it} - X_{1t}. \]

- **Variant I:**
  \[ Y_{it} - Y_{1t} = d_1 \alpha_t + (X_{it} - X_{1t}) \beta + (v_{it} - v_{1t}), \text{ for } t > k > t'. \quad (3.11) \]

- **Variant II:**
  \[ Y_{it} - Y_{1t} = d_1 (C_{1t}) + (X_{it} - X_{1t}) \beta + (v_{it} - v_{1t}), \text{ for } t > k > t'. \quad (3.11') \]

**Random Growth Estimator:**

- Suppose
  \[ U_{it} = \phi_{11} + t \phi_{21} + v_{1t}. \quad (3.12) \]

where \( \phi_{21} \) a person-specific growth rate component. Dependence between \( U_{it} \) and \( d_1 \) assumed due to dependence between \( d_1 \) and \( (\phi_{11}, \phi_{21}) \).

OLS Regression on Transformation of (3.12), where pre-training earnings proxy for \( (\phi_{11}, \phi_{21}) \), will yield consistent estimates. [Pudney (1982)]

- **Variant I:**
  \[ (Y_{it} - Y_{1t}) - (t-t')(Y_{1t'} - Y_{1,t'-1}) = d_1 \alpha_t + [(X_{it} - X_{1t'}) - (t-t')(X_{1t'} - X_{1,t'-1})] \beta + [(v_{it} - v_{1t'}) - (t-t')(v_{1t} - v_{1,t'-1})], \quad (3.13) \]
  for \( t > k > t' \).

- **Variant II:**
  \[ (Y_{it} - Y_{1t}) - (t-t')(Y_{1t'} - Y_{1,t'-1}) = d_1 (C_{1t}) + [(X_{it} - X_{1t'}) - (t-t')(X_{1t'} - X_{1,t'-1})] \beta + [(v_{it} - v_{1t'}) - (t-t')(v_{1t} - v_{1,t'-1})], \quad (3.13') \]
  for \( t > k > t' \).
IV. TESTING ALTERNATIVE SPECIFICATIONS

A. The Pre-Program Tests

- **Data Requirements:** Pre-program earnings and regressor variables for future program participants (trainees and, when available, controls) and comparison group members.

- **Testing Principle:** Apply Candidate selection correction procedure pre-program data. If procedure appropriate, should make the adjusted earnings equation of future trainees and comparison group members alike. If not alike after adjustment, reject candidate correction procedure.

- Test whether $\alpha_t = 0$ in Linear Control Function, Fixed Effect, and Random Growth Estimators, using pre-program earnings data, where $d_i = 1$ if $i$th observation is a future participant (trainee or control) and $d_i = 0$ if member of comparison group.

B. Tests of Model Restrictions

- **Data Requirements:** Same as above or less for some estimators.

- **Testing Principle:** Many selection estimators invoke additional restrictions which can be subjected to test. [See Heckman and Robb (1986)] Rejection of such restrictions would cause rejection of a candidate selection correction method.

- For Linear Control Function Estimator, no testable model restrictions without strong beliefs about the functional form of the earnings equation (3.2) and the appropriate regressor variables.

- For Fixed Effect and Random Growth Estimators, values of pre-training earnings, from periods other than those entering equations (3.8), (3.8'), (3.13) and (3.13'), should not appear in these equations. Test that the coefficients on these extraneous Y values are equal to zero.

C. The Post-Program Tests

- **Data Requirements:** Require access to experimental data, i.e., controls who do not receive training. Controls are like trainees except they do not receive training.

- **Testing Principle:** If a selection correction procedure is valid, it should make the adjusted earnings equations for controls and comparison group members alike.

- Test in Linear Control Function, Fixed Effect, and Random Growth Estimators is whether $\alpha_t$ (or $\theta_t$) = 0.

- Value of this test comes in evaluating an estimator that might be suitable for nonexperimental evaluation of the same program when experimental data are not available or in picking an estimator for a similar program.
V. A RE-ANALYSIS OF THE NATIONAL SUPPORTED WORK DATA

A. The Data Used

- Data from the National Supported Work (NSW) experiment previously analyzed by LaLonde (1986), Fraker and Maynard (1984, 1987), and LaLonde and Maynard (1987).

- Use NSW participants (both trainees and controls) who were either AFDC Recipients (Females) or High School Dropouts (Youth) that were enrolled in the program in either 1976 or 1977.


- Only have grouped data, i.e., mean values of earnings for cells of individuals for both the NSW participants and those in the CPS were provided by the Social Security Administration. Such data precludes the use of many non-linear nonexperimental estimators. Restrict our investigation to linear nonexperimental estimators.

- Variables defined in Table 1 and Mean Values found in Table 2.

B. Estimates of the Impact of Training

- Estimates of Program Impact: High School Dropouts in Table 3 and AFDC women in Table 4.

- Mean Differences Using Experimental Data:
  - AFDC Women: $440 in 1978 (significant) and $267 (only marginally significant)

- Mean Differences Using Nonexperimental Comparison Group Data:
  - High School Dropouts: $-1910 in 1978 and $-1917 in 1979 (both significant)
  - AFDC Women: $157 in 1978 and $79 in 1979 (both insignificant)

- Same "Sensitivity" across estimators found in previous studies.

- These results suggest selection bias is an empirically important problem using the above nonexperimental data to evaluate impact of training on earnings.
Table 1. Definition of Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earnings variables</td>
<td></td>
</tr>
<tr>
<td>SSEARN72</td>
<td>SSA earnings in 1972 (in 1976 dollars)</td>
</tr>
<tr>
<td>SSEARN73</td>
<td>SSA earnings in 1973 (in 1976 dollars)</td>
</tr>
<tr>
<td>SSEARN74</td>
<td>SSA earnings in 1974 (in 1976 dollars)</td>
</tr>
<tr>
<td>SSEARN75</td>
<td>SSA earnings in 1975 (in 1976 dollars)</td>
</tr>
<tr>
<td>SSEARN78</td>
<td>SSA earnings in 1978 (in 1976 dollars)</td>
</tr>
<tr>
<td>SSEARN79</td>
<td>SSA earnings in 1979 (in 1976 dollars)</td>
</tr>
<tr>
<td>Background variables in B1</td>
<td></td>
</tr>
<tr>
<td>BLK4S</td>
<td>1 if black or Hispanic and 0 otherwise</td>
</tr>
<tr>
<td>SEX</td>
<td>1 for men and 0 for females</td>
</tr>
<tr>
<td>MARRIAGE</td>
<td>1 if married at enrollment for NSW participants or at March interview for CPS respondents and 0 otherwise</td>
</tr>
<tr>
<td>AGE</td>
<td>Age in years at enrollment for NSW participants or at March interview for CPS respondents</td>
</tr>
<tr>
<td>EDUC</td>
<td>Years of schooling completed at enrollment for NSW participants or at March interview for CPS respondents</td>
</tr>
<tr>
<td>URBAN</td>
<td>1 if in central-city standard metropolitan statistical area and 0 otherwise</td>
</tr>
<tr>
<td>7677ENR</td>
<td>1 if enrolled in 1977 for NSW participants or if interviewed in March 1977 for CPS respondents and 0 otherwise</td>
</tr>
<tr>
<td>Background variables in B2</td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td>1 if black and 0 otherwise</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>1 if Hispanic and 0 otherwise</td>
</tr>
<tr>
<td>AGESQ</td>
<td>AGE squared</td>
</tr>
<tr>
<td>HOUSESIZE</td>
<td>Number of household members at enrollment for NSW participants or at March interview for CPS respondents</td>
</tr>
<tr>
<td>DEPEND</td>
<td>Number of dependents at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]</td>
</tr>
<tr>
<td>AGEKID</td>
<td>Age of youngest dependent at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]</td>
</tr>
<tr>
<td>Work history variables in W1</td>
<td></td>
</tr>
<tr>
<td>SSEARNL1</td>
<td>Annual earnings (from SSA data) one year prior to enrollment for NSW participants or one year prior to interview of CPS respondents</td>
</tr>
<tr>
<td>SSEARNL2</td>
<td>Annual earnings (from SSA data) two years prior to enrollment for NSW participants or two years prior to the interview for CPS respondents</td>
</tr>
<tr>
<td>WORKWKS</td>
<td>Number of weeks worked in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents</td>
</tr>
<tr>
<td>UEWKS</td>
<td>Number of weeks unemployed in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents</td>
</tr>
<tr>
<td>AVEHRS</td>
<td>Average hours per week (when worked) in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents</td>
</tr>
<tr>
<td>WELFARE</td>
<td>Per capita benefit the household received from welfare and earnings of the sample member in the month prior to enrollment for NSW participants or the interview for CPS respondents</td>
</tr>
<tr>
<td>Work history variables in W2</td>
<td></td>
</tr>
<tr>
<td>SSEARNL3</td>
<td>Annual earnings (from SSA data) three years prior to enrollment for NSW participants or three years prior to interview for CPS respondents</td>
</tr>
<tr>
<td>SSEARNL4</td>
<td>Annual earnings (from SSA data) four years prior to enrollment for NSW participants or four years prior to interview for CPS respondents</td>
</tr>
<tr>
<td>CLERSALE</td>
<td>1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was a clerical or sales occupation and 0 otherwise</td>
</tr>
<tr>
<td>SERVICE</td>
<td>1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was in service sector and 0 otherwise</td>
</tr>
<tr>
<td>PROFESSION</td>
<td>1 if job prior to enrollment for NSW participants or if current/most recent job for CPS respondents was a professional occupation and 0 otherwise</td>
</tr>
<tr>
<td>AFDC</td>
<td>1 if AFDC received in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents and 0 otherwise (used only for high-school dropout (youth) results)</td>
</tr>
</tbody>
</table>
Table 2. Sample Means

<table>
<thead>
<tr>
<th>Variable</th>
<th>High-school dropouts (youths)</th>
<th>AFDC recipients (women)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSW samples</td>
<td>CPS sample</td>
</tr>
<tr>
<td>Earnings variables</td>
<td>Trainees</td>
<td>Controls</td>
</tr>
<tr>
<td>SSEARN72</td>
<td>192.7</td>
<td>228.9</td>
</tr>
<tr>
<td>SSEARN73</td>
<td>329.9</td>
<td>401.1</td>
</tr>
<tr>
<td>SSEARN74</td>
<td>561.1</td>
<td>630.6</td>
</tr>
<tr>
<td>SSEARN75</td>
<td>532.4</td>
<td>504.9</td>
</tr>
<tr>
<td>SSEARN76</td>
<td>1,704.0</td>
<td>1,751.5</td>
</tr>
<tr>
<td>SSEARN79</td>
<td>1,838.2</td>
<td>1,825.5</td>
</tr>
<tr>
<td>Background variables in B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLKHS</td>
<td>.918</td>
<td>.909</td>
</tr>
<tr>
<td>SEX</td>
<td>.083</td>
<td>.084</td>
</tr>
<tr>
<td>MARRIAGE</td>
<td>.044</td>
<td>.033</td>
</tr>
<tr>
<td>AGE</td>
<td>18.200</td>
<td>18.347</td>
</tr>
<tr>
<td>URBAN</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>7677ENR</td>
<td>.645</td>
<td>.651</td>
</tr>
<tr>
<td>Background variables in B2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td>.736</td>
<td>.706</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>.182</td>
<td>.203</td>
</tr>
<tr>
<td>HOUSESIZE</td>
<td>4.704</td>
<td>4.746</td>
</tr>
<tr>
<td>DEPEND</td>
<td>2.167</td>
<td>2.292</td>
</tr>
<tr>
<td>AGEKID</td>
<td>9.341</td>
<td>9.215</td>
</tr>
<tr>
<td>AGESQ</td>
<td>1,189.06</td>
<td>1,181.43</td>
</tr>
<tr>
<td>Work history variables in W1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSEARNL1</td>
<td>559.0</td>
<td>539.3</td>
</tr>
<tr>
<td>SSEARNL2</td>
<td>436.4</td>
<td>447.4</td>
</tr>
<tr>
<td>WORKWKS</td>
<td>9.3</td>
<td>9.3</td>
</tr>
<tr>
<td>UEWKS</td>
<td>10.3</td>
<td>11.2</td>
</tr>
<tr>
<td>HOURS</td>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>WELFARE</td>
<td>33.8</td>
<td>33.7</td>
</tr>
<tr>
<td>Work history variables in W2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSEARNL3</td>
<td>357.1</td>
<td>400.6</td>
</tr>
<tr>
<td>SSEARNL4</td>
<td>180.6</td>
<td>180.1</td>
</tr>
<tr>
<td>CLERALSE</td>
<td>.101</td>
<td>.108</td>
</tr>
<tr>
<td>SERVICE</td>
<td>.256</td>
<td>.212</td>
</tr>
<tr>
<td>PROFESSION</td>
<td>.057</td>
<td>.040</td>
</tr>
<tr>
<td>AFDC</td>
<td>.045</td>
<td>.043</td>
</tr>
<tr>
<td>Number of observations</td>
<td>555</td>
<td>676</td>
</tr>
<tr>
<td>Number of cells</td>
<td>69</td>
<td>67</td>
</tr>
</tbody>
</table>
Table 3. Estimates of Training Effects for High-School Dropouts (youths)

<table>
<thead>
<tr>
<th>Model and control variables sets</th>
<th>1978 earnings (a.)</th>
<th>1979 earnings (C0.)</th>
<th>1978 earnings (a.)</th>
<th>1979 earnings (C0.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No control variables</td>
<td>-1,910 (243)</td>
<td>-1,917 (243)</td>
<td>-1,910 (243)</td>
<td>-1,917 (243)</td>
</tr>
<tr>
<td>B1</td>
<td>-1,884 (247)</td>
<td>-2,119 (247)</td>
<td>-1,827 (328)</td>
<td>-2,092 (328)</td>
</tr>
<tr>
<td>B1 + B2</td>
<td>-1,279 (273)</td>
<td>-1,569 (273)</td>
<td>-1,079 (273)</td>
<td>-1,498 (273)</td>
</tr>
<tr>
<td>B1 + W1</td>
<td>-1,117 (246)</td>
<td>-1,539 (246)</td>
<td>-1,146 (246)</td>
<td>-1,444 (246)</td>
</tr>
<tr>
<td>B1 + B2 + W1 + W2</td>
<td>-889 (328)</td>
<td>-986 (328)</td>
<td>-889 (328)</td>
<td>-1,331 (328)</td>
</tr>
</tbody>
</table>

Fixed-effect estimates constructed with t' = 1972 pretraining earnings

| No control variables            | -1,904 (223)      | -1,910 (223)      | -1,831 (223)      | -2,070 (223)      |
| B1                              | -1,886 (242)      | -2,172 (242)      | -1,831 (242)      | -2,070 (242)      |
| B1 + B2                         | -1,360 (270)      | -1,544 (270)      | -1,227 (270)      | -1,647 (270)      |

Fixed-effect estimates constructed with t' = 1974 pretraining earnings

| No control variables            | -1,456 (203)      | -1,482 (203)      | -1,370 (203)      | -1,636 (203)      |
| B1                              | -1,411 (227)      | -1,883 (227)      | -1,370 (227)      | -1,636 (227)      |
| B1 + B2                         | -1,035 (255)      | -1,330 (255)      | -964 (255)        | -1,383 (255)      |

Random-growth estimates constructed with t' = 1973 and t' - 1 = 1972 pretraining earnings

| No control variables            | -649 (336)        | -446 (336)        | -231 (336)        | -236 (336)        |
| B1                              | -231 (414)        | -241 (414)        | -231 (414)        | -241 (414)        |
| B1 + B2                         | -23 (515)         | -85 (515)         | -23 (515)         | -125 (515)        |

Weighted average of estimates

| -24 (165)                      | -154 (165)        | -24 (165)         | -154 (165)        |

Random-growth estimates constructed with t' = 1973 and t' - 1 = 1973 pretraining earnings

| No control variables            | -499 (328)        | -287 (328)        | -589 (328)        | -721 (328)        |
| B1                              | -614 (431)        | -659 (431)        | -589 (431)        | -701 (431)        |
| B1 + B2                         | -824 (497)        | -850 (497)        | -777 (497)        | -806 (497)        |

Weighted average of estimates

| -616 (426)                     | -724 (426)        | -616 (426)        | -724 (426)        |

Experimental estimates

| -48 (144)                     | 9 (144)           | -48 (144)         | 9 (144)           |

NOTE: Standard errors are in parentheses.

Experimental controls from the experiment (d = 1) and comparison-group members (d = 0). Again, tests of the vector hypothesis $\beta_0 = 0$ are consistent with the test based on $C_0$, and are not reported here.

Under the heading “Model-restriction tests,” we report $P$ values for the hypotheses $\alpha_i = 0$ and $C_0 = 0$ ($t > k$), for earnings models fit on a pooled sample of ex-

Table 4. Estimates of Training Effects for AFDC Recipients (women)

<table>
<thead>
<tr>
<th>Model and control variables sets</th>
<th>1978 earnings (a.)</th>
<th>1979 earnings (C0.)</th>
<th>1978 earnings (a.)</th>
<th>1979 earnings (C0.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No control variables</td>
<td>157 (154)</td>
<td>79 (154)</td>
<td>157 (154)</td>
<td>79 (154)</td>
</tr>
<tr>
<td>B1</td>
<td>666 (154)</td>
<td>494 (154)</td>
<td>666 (154)</td>
<td>494 (154)</td>
</tr>
<tr>
<td>B1 + B2</td>
<td>231 (154)</td>
<td>-195 (154)</td>
<td>231 (154)</td>
<td>-195 (154)</td>
</tr>
<tr>
<td>B1 + W1</td>
<td>853 (154)</td>
<td>468 (154)</td>
<td>853 (154)</td>
<td>468 (154)</td>
</tr>
<tr>
<td>B1 + B2 + W1 + W2</td>
<td>937 (154)</td>
<td>441 (154)</td>
<td>937 (154)</td>
<td>441 (154)</td>
</tr>
</tbody>
</table>

Weighted average of estimates

| 374 (164)                      | 238 (164)         | 374 (164)         | 238 (164)         |

Fixed-effect estimates constructed with t' = 1972 pretraining earnings

| No control variables            | 231 (152)         | 153 (152)         | 231 (152)         | 153 (152)         |
| B1                              | 899 (152)         | 508 (152)         | 899 (152)         | 508 (152)         |
| B1 + B2                         | 938 (152)         | 1,032 (152)       | 938 (152)         | 1,032 (152)       |

Fixed-effect estimates constructed with t' = 1974 pretraining earnings

| No control variables            | 475 (135)         | 397 (135)         | 475 (135)         | 397 (135)         |
| B1                              | 713 (135)         | 522 (135)         | 713 (135)         | 522 (135)         |
| B1 + B2                         | 948 (135)         | 570 (135)         | 948 (135)         | 570 (135)         |

Random-growth estimates constructed with t' = 1973 and t' - 1 = 1972 pretraining earnings

| No control variables            | 494 (367)         | 460 (367)         | 494 (367)         | 460 (367)         |
| B1                              | 78 (367)          | -217 (367)        | 78 (367)          | -217 (367)        |
| B1 + B2                         | 486 (433)         | -1,372 (433)      | 486 (433)         | -1,372 (433)      |

Random-growth estimates constructed with t' = 1974 and t' - 1 = 1973 pretraining earnings

| No control variables            | 1,276 (356)       | 1,396 (356)       | 1,276 (356)       | 1,396 (356)       |
| B1                              | 1,183 (356)       | 1,109 (356)       | 1,183 (356)       | 1,109 (356)       |
| B1 + B2                         | 1,278 (356)       | 850 (356)         | 1,278 (356)       | 850 (356)         |

Weighted average of estimates

| 440 (142)                      | 267 (142)         | 440 (142)         | 267 (142)         |

NOTE: Standard errors are in parentheses.
C. Results of Model Selection Tests for High School Dropouts (Youth)

- See Table 5

- The pre- and post-program tests and model restriction tests generally produce consistent findings.
  - Linear control function and fixed effect models are decisively rejected.
  - The random growth model is not.

- Tests of model restrictions for random growth model are mixed.
  - Using 1973 and 1974 earnings to proxy the unobserved components, \( \phi_{11} \) and \( \phi_{21} \), the model is not rejected, but using 1972 and 1973 earnings, it is.
  
  - Based on further analysis of a less restricted form of the random growth model (reported on in paper), it appears that the true specification for Youth is close to the random growth specification in (3.13) or (3.13') and that the rejection for 1972 and 1973 is due to the use of extremely long lags in pre-training earnings.

- "Weighted Average of Estimates" in Table 3.
  - Random Growth Model produces the same inference about program impact as the experiment.
  
  - The former estimates are more negative than the latter, but the standard errors of the former are also bigger.

- Based on the consistency of the Pre-Program and Model Restriction Tests with the Post-Program Tests, one would have chosen the same nonexperimental estimator (Random Growth) with or without experimental data.

- Summary: We reach different conclusion than past studies on this data (LaLonde and Fraker and Maynard) concerning the ability to discriminate between alternative nonexperimental estimators.
  
  - The nonexperimental models we reject are the source of discrepancy previously reported in the literature.
### Table 5. Specification Tests of Nonexperimental Estimators for High-School Dropouts (youths)

<table>
<thead>
<tr>
<th>Control variable set</th>
<th>Preprogram tests</th>
<th>1976 earnings</th>
<th>1979 earnings</th>
<th>Postprogram tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α₁ = 0</td>
<td>Cβ₀ = 0</td>
<td>α₁ = 0</td>
<td>Cβ₀ = 0</td>
</tr>
<tr>
<td>1975 earnings as dependent variable</td>
<td>1978 or 1979 earnings as dependent variable</td>
<td>1978 or 1979 earnings as dependent variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear control function estimators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No control variables</td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₁</td>
<td>.000</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₁ + B₂</td>
<td>.000</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₁ + W₁</td>
<td>.000</td>
<td>.208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₁ + B₂ + W₁</td>
<td>.016</td>
<td>.336</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Fixed-effect estimators       |                  |               |               |                   |                   |                   |
| No control variables          | .000            |               |               |                   |                   |                   |
| B₁                         | .000            | .000          |               |                   |                   |                   |
| B₁ + B₂                    | .000            | .019          |               |                   |                   |                   |

| Random-growth estimators      |                  |               |               |                   |                   |                   |
| No control variables          | .000            |               |               |                   |                   |                   |
| B₁                         | .375            | .173          | .316          | .080             |                   |                   |
| B₁ + B₂                    | .558            | .128          | .614          | .042             |                   |                   |

The rejected version of the model uses the longest lags in preprogram earnings of any of the fitted models to eliminate the permanent and random-growth components in \( U_{t'} \) (\( t' = 6 \) and \( t' = 7 \) for 1978 and 1979, respectively). Earnings functions are well known to be concave in age or experience. The linear growth specification (3.14) may become a progressively poorer approximation as the lag length increases between the dependent variable and the proxy variables. A better model might augment (3.14) to include a third component, \( \phi_{2t} \), multiplied by \((t - t')^2\). To find sufficient proxy variables for this model requires a third year of preprogram earnings data, which is not available to us. Models with autoregressive specifications for \( U_t \) failed specification tests.

A slight extension of (3.14) produces a model that passes specification tests and produces estimates of program impact very close to those obtained from the random-growth model reported in Table 3. In place of (3.14), we write

\[
U_{t'} = \phi_{1t} + b_t \phi_{2t} + v_{t'}, \quad b_t \neq b_{t'}, \quad t \neq t'.
\]
D. Results of Model Selection Tests for AFDC Recipients (Women)

- See Table 8.

- Pre- and Post-Program Tests not decisive. The Tests of Model Restrictions has more bite.
  - For both the pre-program and post-program versions of these tests, the fixed effect and random growth models are rejected.
  - By default, do not reject the linear control function estimators.

- "Weighted Average of Estimates" in Table 4.
  - Linear Control Function Estimator leads to the same inference as found from the experiment.
  - Again, the source of the difference in our results and previous studies is the rejected models.

VI. CONCLUSIONS

- We have considered two fallacies concerning nonexperimental evaluations.

- Presented an illustration of how to choose among alternative nonexperimental estimators.
  - Shown that simple model selection strategy based on easily implemented specification tests can eliminate nonexperimental evaluation models that do not produce estimated program impacts close to the experimental results.
  - We can eliminate the most unreliable and misleading estimators which give rise to the "sensitivity" problem found in the literature.

- Our results, while far from definitive, are encouraging. They suggest a feasible a strategy for the reliable (and credible) use of nonexperimental methods to evaluate social programs.
estimators may not be well-founded and that a systematic procedure exists to identify estimators that replicate the inferences drawn from experimental methods.

5.3 Results of Model-Selection Tests for AFDC Recipients (women)

Table 8 reports the results of specification tests applied to alternative earnings equations for AFDC women. The format of this table is the same as that of Table 5. Neither the preprogram tests nor the postprogram model-specification tests are decisive in rejecting any of the models. (The coefficient estimates associated with these tests range from -686 to 476 for the linear control function models, -449 to 750 for the fixed-effect models, and -1,961 to 1,071 for the random-growth models.) The tests of model restrictions have much more bite. For both the preprogram and postprogram versions of these tests, the fixed-effect and random-growth models are decisively rejected. By
REFERENCES


