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
and
Population Research Center
University of Chicago and NORC

Delivered at

The World Bank
Washington, DC

July 20-21, 1994

Latest Revision: April 25, 2007

The following provides an outline of the topics to be considered during the two days of lectures I will be giving on evaluation design and methods as they apply to various types of social programs and/or organized interventions into communities and targeted populations. My lectures are intended to provide an introduction to issues that one needs to confront in designing an evaluation of some program or intervention and the range of options available to address these issues.

I will focus on three issues: Experimental Evaluation Designs, Non-Experimental Evaluation Designs and Process Analysis. Even though the prospects of using an experimental design may be remote, I will treat experimental designs because it provides a useful contrast between ideal evaluation designs and the practical problems in social contexts which the feasibility of experimental evaluations problematic. Note that I will not cover cost-benefit analysis in my lectures, even though it is often a focus of evaluations. I have omitted it because I suspect the Bank's Research Division already has expertise in this area and/or can find experts who are more versed in this area than I.

Below, I provide a preliminary reading list of background material for these lectures. I have "*" those readings which you might want to look at in advance of the lectures. I will attempt to structure most of my lectures so that they do not depend upon your reading specific pieces. Rather, I will try to cover topics in a relatively self-contained way and then suggest further readings which you might want to consult for further details and/or more technical treatments.

Reading List

I. The Evaluation Problem and General Issues in Designing Evaluations

Topics: Introduction to issues in designing social evaluations
Definition of the central problem in evaluations: self selection
Alternative types of evaluation designs
General issues the appropriateness of alternative designs

Readings:

Levitan, S. (1992), *The Evaluation of Federal Social Programs: An Uncertain Impact*, George Washington University, mimeo.

Nathan, R. (1988), *Social Science in Government: Uses and Misuses*, New York: Basic Books, Chapter 1.

Manski, C. and I. Garfinkel (1992), "Introduction," in C. Manski and I. Garfinkel, eds. *Evaluating Welfare and Training Programs*. Harvard University Press.

Campbell, D. and J. Stanley (1963), *Experimental and Quasi-Experimental Design for Research*, Chicago: Rand-McNally, 1966.

Heckman, J. (1989), "Causal Inference and Self-Selection," *Journal of Educational Statistics*, Summer 1989.

Holland, P. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, Vol. 81, 1986, 945-970.

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

Topics: The experimental design and the advantages of random assignment
Ideal experiments versus actual implementations

- The feasibility of implementing random assignment designs in social and program contexts
- What can be estimated with data from alternative experimental designs?
  1. The "black box" nature of experimental results versus capturing the structure of behavior
  2. Dealing with macro-level effects of programs
  3. Dealing with the entry effects of programs
  4. Non-compliance with Treatment Designs: Problems of "No-Shows" and "Cross-overs"

Readings:

Burtless, G. (1988), "The Social and Scientific Value of Controlled Experimentation," *Proceedings of the Fortieth Annual Meeting*, Industrial Relations Research Association.

Burtless, G. and L. Orr (1986), "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources*, 21, 1986, pp. 606-639.

Barnow, B. (1988), "The Uses and Limits of Social Experiments," *Proceedings of the Fortieth Annual Meeting*, Industrial Relations Research Association.

Heckman, J. (1992), "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Hausman, J. and D. Wise (1985), "Technical Problems in Social Experimentation: Cost versus Ease of Analysis," in *Social Experimentation*, ed. by Jerry Hausman and David Wise, Chicago: University of Chicago Press. [See also discussion by John Conlisk].

Hotz, V. J. (1992), "Designing an Evaluation of the Job Training Partnership Act," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Greenberg, D., R. Meyer, and M. Wiseman (1993), "Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs," Institute for Research on Poverty Discussion Paper # 999-93.

Manski, C., I. Garfinkel, and C. Michalopoulos (1992), "Micro Experiments and Macro Effects," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Moffitt, R. (1992), "Evaluation Methods for Program Entry Effects," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Hotz, V. J. and S. Sanders (1994), "Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice," Unpublished Manuscript, University of Chicago March 1994.

III. Non-Experimental (or Quasi-Experimental) Designs

Topics: An overview of the non-experimental methodology and the nature of the selection problem
A taxonomy of statistical methods for use in non-experimental evaluations
Issues arising in designing non-experimental evaluations of social programs

Readings:

Moffitt, R. (1991), "Program Evaluation with Nonexperimental Data," *Evaluation Review*, Vol. 15, No. 3, June 1991, 291-314.

Heckman, J. and R. Robb (1986), "Alternative Methods for Evaluating the Impact of Interventions: An Overview," *Journal of Econometrics*, Vol. 30, 1986, pp. 238- 269.

Angrist, J. and G. Imbens (1991), "Sources of Identifying Information in Evaluation Models," Unpublished manuscript, Harvard University, August 1991.

Mohr, L. (1988), *Impact Analysis for Program Evaluation*, Beverly Hills: SAGE Publications, Chapters 5-10.

Barnow, B., G. Cain, and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies Review Annual*, ed. by E. Stromsdorfer and G. Farkas, Vol. 5, 42-59.

Cain, G., S. Bell, L. Orr, and W. Lin (1993), "Using Data on Applicants to Training Programs to Measure the Program's Effects on Earnings," Institute for Research on Poverty Discussion Paper # 1015-93.

Cook, T. and D. Campbell (1979), *Quasi Experimentation*, New York: Houghton-Mifflin, Chapters 1-5, 8.

LaLonde, R. (1987), "Evaluating the Econometric Evaluations of Employment and Training Programs," *American Economic Review*, Vol. 76, #4, 1987, pp. 604-620.

Fraker, T. and R. Maynard (1987), "Evaluating Comparison Group Designs with Employment-Related Programs," *Journal of Human Resources*, 1987, pp. 194-227.

Heckman, J. and V. J. Hotz (1989), "On the Use of Nonexperimental Methods for Estimating the Impact of Manpower Training Programs; On Reevaluating the Evaluations," *Journal of The American Statistical Association*, Vol. 84, December, 1989, 862-880.

IV. Process Analysis and its Role in Evaluations

Topics: The aims of process analysis
The advantages and disadvantages of process analysis in evaluating social programs
An example of a process analysis

Readings:

Rossi, P. and H. Freeman (1993), "*Evaluation: A Systematic Approach*," 5th Ed., Newbury Park, CA: Sage Publications, Chapter 4.

Nathan, R. (1988), *Social Science in Government: Uses and Misuses*, New York: Basic Books, Chapter 6.

Riccio, J. and D. Friedlander (1992), *GAIN: Program Strategies, Participation Patterns, and First-Year Impacts in Six Counties*. New York: Manpower Research Development Corp.

Professor V. Joseph Hotz

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_{1it} 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_{0it} 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_{it} 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:

$$Y_{it} = Y_{1it} d_i + Y_{0it}(1 - d_i) = \alpha_{it} d_i + Y_{0it} \tag{1}$$

- One is interested in knowing

$$\alpha_{it} = Y_{1it} - Y_{0it},$$

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 participant 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 α_{it}, 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:

$$\alpha_t = E(\alpha_{it} \mid d_i = 1)$$

- 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:

$$E(\bar{Y}_{Tt}) = E(Y_{it} \mid d_i = 1) = E(Y_{0it} + \alpha_{it} \mid d_i = 1) \qquad (2)$$

  and

$$E(\bar{Y}_{Nt}) = E(Y_{it} \mid d_i = 0) = E(Y_{0it} \mid d_i = 0) \qquad (3)$$

  Consider the mean of the difference of $\bar{Y}_{Tt}$ and $\bar{Y}_{Nt}$:

$$\begin{aligned} E(\bar{Y}_{Tt} - \bar{Y}_{Nt}) &= E(Y_{0it} + \alpha_{it} \mid d_i = 1) - E(Y_{0it} \mid d_i = 0) \\ &= E(\alpha_{it} \mid d_i = 1) + [E(Y_{0it} \mid d_i = 1) - E(Y_{0it} \mid d_i = 0)] \end{aligned} \qquad (4)$$

  The problem of selection bias concerns whether the term $E(Y_{0it} \mid d_i = 1)$ - $E(Y_{0it} \mid d_i = 0)$ 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_{0it},Y_{1it}) are not independent. That is:

$$f(d_i, Y_{0it}, Y_{1it}) \neq f_d(d_i) f_Y(Y_{0it}, Y_{1it})$$

  - 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: $E(Y_{0it} \mid d_i = 1)$ - $E(Y_{0it} \mid d_i = 0)$.

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 that 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 ***n***ot 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? There is the potential for *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 can***n***ot 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) that 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 individu-

als; 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

Campbell, D. and J. Stanley (1963), *Experimental and Quasi-Experimental Design for Research*, Chicago: Rand-McNally, 1966.

Cook, T. and D. Campbell (1979), *Quasi Experimentation*, New York: Houghton-Mifflin.

Nathan, R. (1988), *Social Science in Government: Uses and Misuses*, New York: Basic Books.

Professor V. Joseph Hotz

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 $D(S=s) \equiv D(s) = k$ 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 treat-

ment 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.

$Y_s \equiv Y \mid S=s$ denote the outcome an individual would realize if they had been in Treatment Regime s, for $s = T$ or C.

$Y_k \equiv Y \mid D(\cdot)=k$ 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 that enable us to estimate aspects of their distribution.

Let $f(\cdot,\cdot,...)$ denote a density function for its arguments and $P(\cdot) \equiv Pr(\cdot)$ 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?*

$$\alpha \equiv E(Y_1 - Y_0 \,|\, D(T) = 1, S = T)$$

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?*

$$\tilde{\alpha} \equiv E(Y_T - Y_C \,|\, S = T)$$

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?*

$$P(D(T)=1 \mid S=T)$$

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?*

$$\alpha^* \equiv E(Y_1 - Y_0)$$

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 $f(Y_k \mid D(s)=k, S=s)$, 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

Heckman, J. (1992), "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Manski, C. (1993), “What do the Outcomes of Homogenous Treatments Reveal about Outcomes when Treatments Vary?: The Mixing Problem,” Social Systems Research Institute Discussion Paper # 9313R, University of Wisconsin, October 1993.

Professor V. Joseph Hotz

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:

$$\begin{aligned} E(Y_T \mid S=T) - E(Y_C \mid S=C) &= \tilde{\alpha} + [E(Y_C \mid S=T) - E(Y_C \mid S=C)] \\ &= \tilde{\alpha} \end{aligned} \tag{1}$$

  Since $E(Y_C \mid S=T) = E(Y_C \mid S=C) = E(Y_C)$.

- 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 estimator for Q1.

♦ *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 $E(Y_T|S=T) = E(Y_1|D(T)=1)$ and, thus, $\alpha_1 = \tilde{\alpha}$. Then the mean difference in outcomes between experimental and control groups yields an unbiased estimator of α_1.

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

  The above Assumptions imply that the following result:

  $$\frac{E(Y_T|S=T)-E(Y_C|S=C)}{P(D(T)=1)} = E(Y_1|D(T)=1)-E(Y_0|D(T)=1)=\alpha_1. \qquad (2)$$

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

  $$E(Y_T|S=T)=P(D(T)=1)E(Y_1|D(T)=1)+[1-P(D(T)=1)]E(Y_0|D(T)=0)\,. \qquad (3)$$

  Moreover, $E(Y_C|S=C)$ 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:

  $$\begin{aligned} E(Y_C|S=C) &= P(D^L(C)=1)E(Y_0|D^L(C)=1)+[1-P(D^L(C)=1)]E(Y_0|D^L(C)=0) \\ &= P(D(T)=1)E(Y_0|D(T)=1)+[1-P(D(T)=1)]E(Y_0|D(T)=0) \end{aligned}, \qquad (4)$$

  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 $[E(Y_1|D(T)=1) - E(Y_0|D(T)=1)]$, 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 experimentals and controls does not provide unbiased estimates for Q1 (i.e., α_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 α_1 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 $k = 0,...,K$, $K > 2$.

♦ *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 experimentals 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 $k = 1,...,K$. 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 α_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:

$$\begin{aligned} E(Y_0 \mid S=C) &= \sum_{j=0}^{K} P(D^L(C)=j)E(Y_0 \mid D^L(C)=j) \\ &= P(D^L(C)=k)E(Y_0 \mid D^L(C)=k) + \sum_{j=0, j\neq k}^{K} P(D^L(C)=j)E(Y_0 \mid D^L(C)=j) \\ &= P(D^L(C)=k)E(Y_0 \mid D^L(C)=k) + [1-P(D^L(C)=k)]E(Y_0 \mid D^L(C)=\sim k) \\ &= P(D(T)=k)E(Y_0 \mid D(T)=k) + [1-P(D(T)=k)]E(Y_0 \mid D(T)=\sim k) \end{aligned}, \quad (5)$$

where $D^L(C) = \sim k$ denotes the fact that the latent treatment choice is *not* k.[1]

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

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

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

where

$$c_1 = E(Y_T \mid S=T) - E(Y_C \mid S=C),$$

$$c_{2k} = [1-P(D(T)=k)]E(Y_{\sim k} \mid D(T)=\sim k),$$

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

Because $E(Y_0 \mid D^L(C)=\sim k)$ is not identified by experimental data, one cannot achieve point identification of α_k.

However, one can bound α_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 α_k hinge on obtaining bounds on $E(Y_0 \mid D^L(C)=\sim k)$, the mean outcome in the control group for latent treatment groups other than k.

Several alternative sets of bounds on α_k can be formed. These are what Hotz and Sanders

[1]Note that $E(Y_T \mid S=T)$ can be written in a similar fashion.

call the *Horowitz-Manski Bounds on* α_k.[2]

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

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 $E(Y_0|D^L(C)=\sim k)$ in the following way.

Suppose we assume that all of the observations for which $D^L(C)=\sim k$ 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 $E(Y_0|D^L(C)=\sim k)$, we presume that all of the observations on $Y_0|D^L(C)=\sim k$ 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 $[1-P(D(T)=k)]N_C^{th}$ and $P(D(T)=k)N_C^{th}$ order statistics of Y_0 for the control group. Then the lower bound on $E(Y_0|D^L(C)=\sim k)$ is given by:

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

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

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

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

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

and

$$B^2_{Uk} = c_1 - c_{2k} + \left(\frac{1-P(D(T)=k)}{P(D(T)=k)}\right) E(Y_0|S=C, Y_0 \leq Y_{0,P(D(T)=k)N_C}). \quad (7b)$$

- The Horowitz-Manski bounds are robust and impose no further restrictions or con-

[2]The derivation of these bounds, and their properties, follows from results in Horowitz and Manski (1993) on forming bounds with contaminated samples.

straints 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 affect.

- *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

Aigner, D. (1985), "The Residential Electricity Time-of-Use Pricing Experiments: What Have We Learned?" in J. Hausman and D. Wise, eds. *Social Experimentation*, Chicago: University of Chicago Press, pp. 11-53.

Cain, G. and H. Watts (1973), *Income Maintenance and Labor Supply*, New York: Academic Press.

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

Greenberg, D., R. Meyer, and M. Wiseman (1993), "Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs," Institute for Research on Poverty Discussion Paper # 999-93.

Heckman, J. (1992), "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs in the 1990's*, ed. by Irwin Garfinkel and Charles Manski, Cambridge, MA: Harvard University Press.

Hotz, V. J. and S. Sanders (1994), "Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice," Unpublished Manuscript, University of Chicago, March 1994.

Professor V. Joseph Hotz

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 therby to improve the effectivenss 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 particpation 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.

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

Ashenfelter, O. and Card, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, 67, 648-660.

Barnow, B. (1987), "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources,* XXII, 157-193.

Burtless, G. and Orr, L. (1986), "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources*, 21, 606-639.

LaLonde, R. and Maynard, R. (1987), "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Studies*, 11, 428-51.

Stromsdorfer, E., *et al*. (1985), "Recommendations of the Job Training Longitudinal Survey Research Advisory Panel," Report to the Employment and Training Administration, U.S. Department of Labor.

TABLE 1.1

ESTIMATES OF THE IMPACT OF CETA SERVICES
ON ADULT PARTICIPANTS' ANNUAL EARNINGS

	Westat (1981)	Westat (1984)	Westat (1984)	Bassi (1983)	Bassi et al.(1984) Nonwelfare Disadvantaged Adults	Bassi et al. (1984) Welfare	Bloom & McLaughlin (1982)	Dickinson et al. (1984) Adults	Geraci (1984)
Year of Program Participation	1975-76	1975-76	1976-77	1975-76	1976-77	1976-77	1975-76	1976	1975-76
Overall	$300*	$129*	$596*	--	--	--	--	--	--
White Women	500*	408*	534*	$740 to 778*	705* to 762*	$840* to 949*	--	--	--
White Men	200	(4)	500*	--	17 to 136	578 to 691*	--	--	--
Minority Women	600*	336*	762*	426 to 671*	779* to 810*	659* to 703*	--	--	--
Minority Men	200	(104)	658*	117 to 211	116 to 369	(273) to 69	--	--	--
Women	--	--	--	--	--	--	--	--	--
Men	--	--	--	--	--	--	800* to 1,300*	13	--
							200	(690)*	--
Classroom Training	350*	267*	740*	--	--	--	--	--	--
White Women	550*	--	--	63 to 205	295 to 354*	315 to 451*	1,300*	--	--
White Men	400	--	--	--	(543)* to (457)	(440) to (120)	300	--	--
Minority Women	500*	--	--	426 to 633*	245 to 301	206 to 369*	1,000*	--	--
Minority Men	200	--	--	582 to 773	102 to 185	(571) to (99)	300	--	--
Women	--	--	--	--	--	--	800* to 1,400*	0	1,201*
Men	--	--	--	--	--	--	300	(343)	372
On-the-Job Training	850*	531*	1,091*	--	--	--	--	--	--
White Women	550*	--	--	80 to 382	701* to 724*	190 to 318	1,200*	--	--
White Men	750*	--	--	--	616* to 756*	995 to 1,231*	(200)	--	--
Minority Women	1,200*	--	--	1,368* to 1,549*	223 to 244	564 to 587	800*	--	--
Minority Men	1,150*	--	--	2,053* to 2,057*	722 to 812*	454 to 750	1,500*	--	--
Women	--	--	--	--	--	--	700* to 1,100*	35	882*
Men	--	--	--	--	--	--	300	(363)	612*

SOURCE: Barnow, 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

Feature	RFP Plan	Final Research Design
Sites	Up to 20, chosen to statistically represent the JTPA system	16, chosen to illustrate the diversity of the JTPA system
Sample	Up to 30,000 adults and youth eligible for Title IIA	20,606 adults and out-of-school youth eligible for Title IIA
Evaluation of JTPA "As Is" with Little Change in the Program	Yes	Yes
Evaluation of JTPA as a Whole	Yes	Yes
Evaluation of Specific Treatments	Focus was on specific activities, such as OJT, classroom occupational training, and job search assistance	Focus on the types of combinations and sequences actually provided in JTPA, including categories of activities anchored on OJT and classroom occupational training
Services for Which the Control Group Is Eligible	Services in the community not funded by JTPA	Services in the community not funded by JTPA

TABLE 5.2

SDAs CONTACTED AND PARTICIPATION RATE,
BY PHASE OF SELECTION PROCESS

Phase of Site Selection Process	Total SDAs Contacted	SDAs Participating	SDAs Rejecting	SDAs Dropped[a]	Participation Rate (%)
Date of Initial Contact:					
Phase 1: Initial Design[b]/ Probabilistic[c] Selection (Before 1/6/87)	83[d]	5	56	22	6.0
Phase 2: Initial Design/ Expanded Recruitment[e] (Between 1/6/87 and 4/30/87)	61	4	49	8	6.6
Phase 3: Final Design[b]/ Expanded Recruitment (After 4/30/87)	85	7	65	13	8.2
Date of Final Decision:					
Phase I: Initial Design/ Probabilistic Selection (Before 1/6/87)	48	1	34	13	2.1
Phase 2: Initial Design/ Expanded Recruitment (Between 1/6/87 and 4/30/87)	61	3	47	11	4.9
Phase 3: Final Design/ Expanded Recruitment (After 4/30/87)	120	12	89	19	10.0
Total	229	16	170	43	7.0

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 2.1

PERCENT OF SDAs CITING SPECIFIC CONCERNS ABOUT THE STUDY

Concern	Percent of SDAs Citing the Concern
Ethical and Public Relations Implications of: Random Assignment in Social Programs	61.8
Denial of Services to Controls	54.4
Potential Negative Effect of Creation of a Control Group on Achievement of Client Recruitment Goals	47.8
Potential Negative Impact on Performance Standards	25.4
Implementation of the Study When Service Providers Do Intake	21.1
Objections of Service Providers to the Study	17.5
Potential Staff Administrative Burden	16.2
Possible Lack of Support by Elected Officials	15.8
Legality of Random Assignment and Possible Grievances	14.5
Procedures for Providing Controls with Referrals to Other Services	14.0
Special Recruitment Problems for Out-of-School Youth	10.5
Sample Size	228

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.
Percents may add to more than 100.0 because SDAs could raise more than one concern.

FIGURE 1.1

RANDOM ASSIGNMENT MODEL FOR THE NATIONAL JTPA STUDY

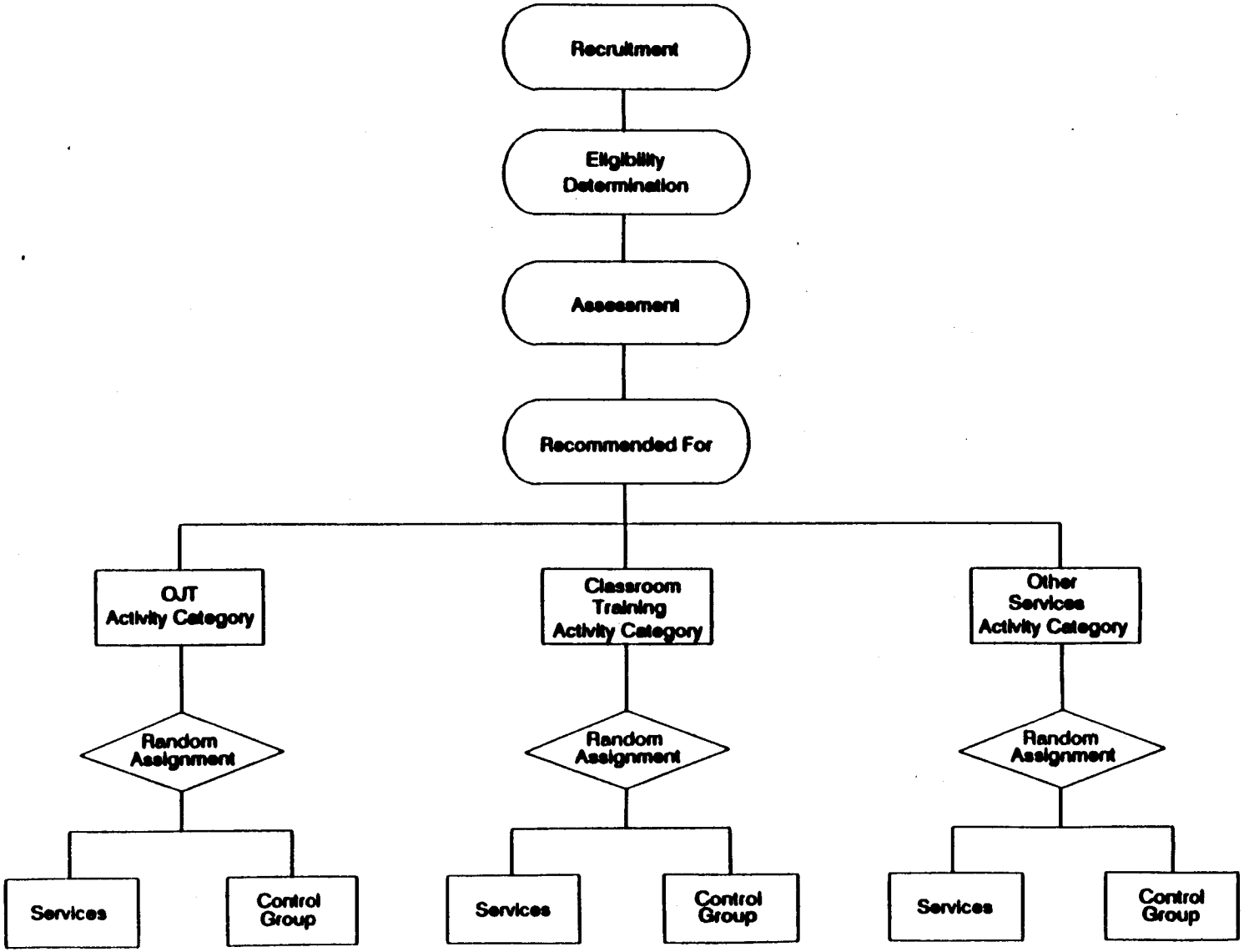

FIGURE 4.1

RANDOM ASSIGNMENT TO TREATMENTS
OR TO A CONTROL GROUP

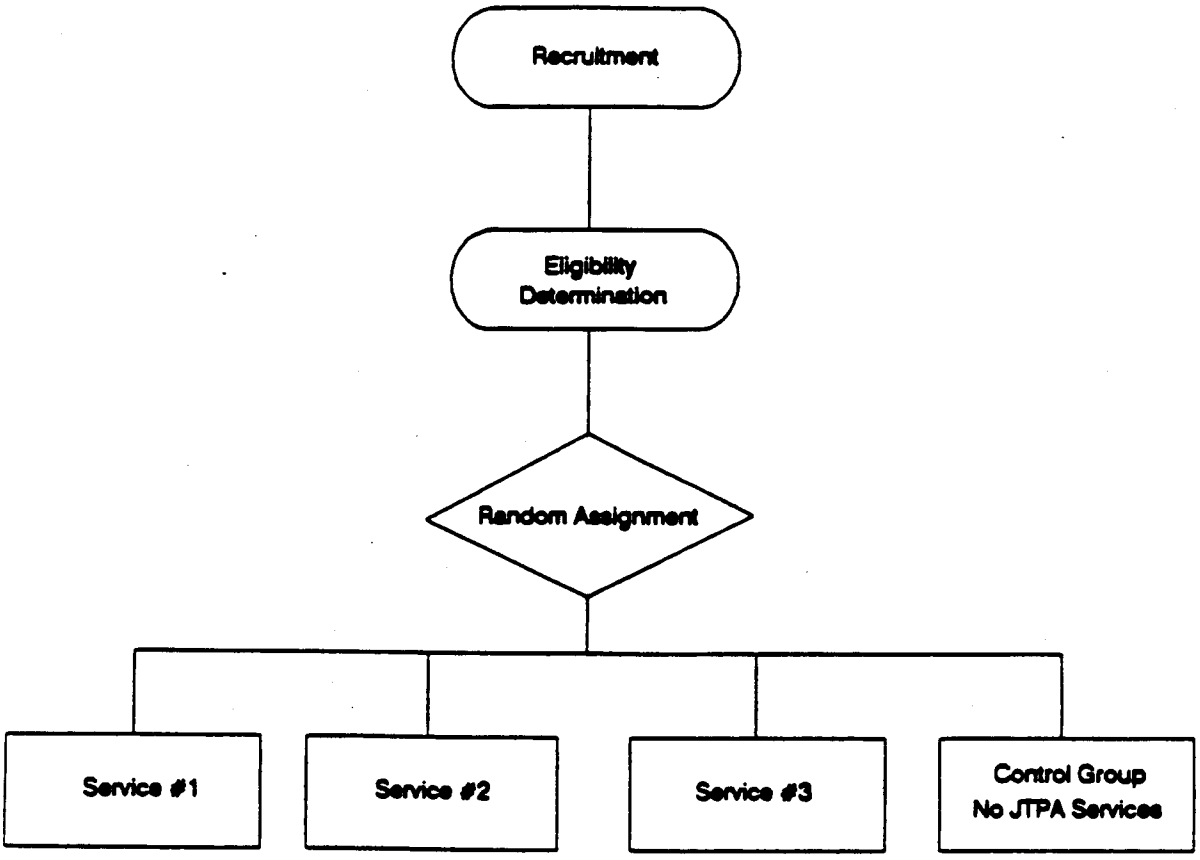

TABLE 4.2

ACTIVITIES AVAILABLE TO APPLICANTS ASSIGNED TO EACH TREATMENT CATEGORY

Assigned Treatment Category	Activities Available to Applicants						
	On-the-Job Training (OJT)	Classroom Training in Occupational Skills (CT-OS)	Combined OJT and CT-OS (Customized Training)	Basic Education	Job Search Assistance/ Job Placement	Work Experience	Other Activities
On-the-Job Training	Yes	No	No	Yes	Yes	Yes	Yes
Classroom Training in Occupational Skills	No	Yes	No	Yes	Yes	Yes	Yes
Other Activities[a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: [a]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 5.8

FINAL SAMPLE, BY TARGET GROUP AND TREATMENT CATEGORY

Target Group	Treatment Category			Total
	OJT	CT-OS	Other Services	
Adult Males	3,190	1,592	2,078	6,860
Adult Females	2,672	3,417	1,980	8,069
Out-of-School Youth	1,571	2,097	2,009	5,677
White	1,027	981	873	2,881
Minority	544	1,116	1,136	2,796
Total	7,433	7,106	6,067	20,606

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

SDA	Target Sample	Actual Sample	Difference
Capital Area, MS (Jackson)	1,220	1,478	+258
Concentrated Employment Program, MT (Butte)	825	683	-142
Coosa Valley, GA (Rome)	1,800	1,840[a]	+40
Corpus Christi/Nueces County, TX	1,500	1,609[a]	+109
Crawford/Hancock/Marion/ Wyandot Counties, OH	1,150	1,154	+4
East Central Iowa (Cedar Rapids)	2,963	498	-2,465
Greater Omaha, NE	1,600	1,362	-238
Heartland, FL (Lakeland)	4,850	597	-4,253
Jersey City, NJ	1,600	1,686[a]	+86
Larimer County, CO (Fort Collins)	1,200	1,027	-173
Macon/De Witt Counties, IL (Decatur)	750	471	-279
Northeast, IN (Fort Wayne)	3,600	3,608	+8
Northwest, MN (Crookston and Thief River Falls)	550	560	+10
Oakland, CA	1,065	1,072[a]	+7
Providence/Cranston, RI	1,750	1,759[a]	+9
Springfield, MO	2,000	1,202	-798
Total	28,423	20,606	-7,817

NOTE: [a]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

Specific Program Services Received	*Percentage of Treatment/Control Group that Received Particular Services:*[a]					
	Classroom Training		On-the-Job Training/Job Search Ass.		Other Services	
	Treatment Group	Control Group	Treatment Group	Control Group	Treatment Group	Control Group
Never enrolled	27.60%		43.50%		37.70%	
Classroom training in occupational skills	56.20	27.70%	3.30	11.00%	9.40	16.30%
Basic education[b]	12.90	8.70	3.10	4.70	15.70	9.40
On-the-job training	3.80	0.20	28.00	0.60	4.70	0.20
Job search assistance	19.50		28.90		19.70	
Work experience	4.00	0.40	2.90	0.30	2.30	0.40
Miscellaneous[c]	9.90		6.50		31.00	
Sample Size	4119	1769	4291	1851	3064	1183

Source: Exhibit 3.18, Bloom, *et al.* (1992).

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

Let the *potential outcome*, Y_{it}^0, be characterized as:

$$Y_{it}^0 = g_t^0(X_i) + U_{it}^0 \tag{5.0a}$$

and Y_{it}^1 by

$$Y_{it}^1 = g_t^1(X_i) + U_{it}^1 \tag{5.0b}$$

where $g_t^j(X_i) \equiv E(Y_{it}^j \,|\, X_i), j = 1, 0.$

In the general, ***heterogeneous treatment effect*,** case,

$$\begin{aligned} Y_{it} &= D_i Y_{it}^1 + (1 - D_i) Y_{it}^0 \\ &= Y_{it}^0 + (Y_{it}^1 - Y_{it}^0) D_i \\ &= g_t^0(X_i) + \alpha_{it}(X_i) D_i + U_{it}^0 \\ &= g_t^0(X_i) + \alpha_t(X_i) D_i + [U_{it}^0 + D_i(U_{it}^1 - U_{it}^0)] \\ &= g_t^0(X_i) + \alpha_t(X_i) D_i + U_{it}^* \end{aligned} \tag{5.1a}$$

where the effect of the treatment, *D*, for individual *i* is defined to be:

$$\alpha_{it}(X_i) \equiv Y_{it}^1 - Y_{it}^0 = \left(g_t^1(X_i) - g_t^0(X_i)\right) + \left(U_{it}^1 - U_{it}^0\right) \tag{5.1b}$$

and the *expected treatment effect* in period *t* conditional on X_i is:

$$\alpha_t(X_{it}) \equiv E(\alpha_{it} \,|\, X_i) = g_t^1(X_i) - g_t^0(X_i) \tag{5.1c}$$

Note that a special case of (5.1d) – the ***homogeneous treatment effect*** case – is characterized by:

$$\tilde{\alpha}_t(X_i) \equiv \alpha_{it}(X_i) = g_t^1(X_i) - g_t^0(X_i) \tag{5.2a}$$

which arises when $U_{it}^1 = U_{it}^0$ and implies the following specification of the outcome equation:

$$\begin{aligned} Y_{it} &= g_t^0(X_i) + \tilde{\alpha}_t(X_i)D_i + U_{it}^0 \\ &= g_t^0(X_i) + \tilde{\alpha}_t(X_i)D_i + \tilde{U}_{it} \end{aligned} \tag{5.2b}$$

Yet another special case of (5.1d) – the ***common*** or ***constant treatment effect*** case – is characterized by:

$$\alpha_t \equiv \alpha_{it}(X_i), \text{ for all } i, X_i \tag{5.3a}$$

which implies that the outcome equation can be written as:

$$\begin{aligned} Y_{it} &= g_t^0(X_i) + \alpha_t D_i + U_{it}^0 \\ &= g_t^0(X_i) + \alpha_t D_i + \hat{U}_{it} \end{aligned} \tag{5.3b}$$

Finally, note that yet a further specialization of the specifications of the potential outcomes in (5.1a) and (5.1b) restricts

$$g_t^j(X_i) \equiv X_i \beta_t^j, j = 1, 0, \tag{5.3c}$$

which gives rise to linear (in X) versions of the observed outcome equations in (5.1a), (5.2b) and (5.3b) above.

Selection bias arises when the disturbance terms, $U_{it}^*, \tilde{U}_{it}$ or $\hat{U}_{it}$ in the outcome equations, (5.1a), (5.2b) and (5.3b), respectively, are correlated with the treatment status, D_i. Note that this bias will arise when the treatment status, D:

(a) depends upon U_{it}^0, the pre-treatment *level* of U, and/or

(b) depends on $U_{it}^1 - U_{it}^0$ $[= \alpha_{it}(X_i) - \alpha_t(X_i)]$, the unobserved *gain* associated with the treatment relative to no treatment.

1.1 Origins of Selection Bias

Statistically-Based Approaches:

Let the index, IN_i, be a function of both observed (Z_i) and unobserved (V_i) variables.

$$IN_i = Z_i \gamma + V_i \tag{5.4}$$

Then the i^{th} individual's training status is

$$D_i = \begin{cases} 1 \text{ if and only if } IN_i > 0, \\ 0 \text{ otherwise.} \end{cases} \tag{5.5}$$

The error term 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:

$$\Pr(D_i = 1 \mid Z_i) = 1 - F(-Z_i\gamma) \equiv p(Z_i) \tag{5.6a}$$

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

To the extent that either Z_i and/or V_i are correlated with U_{it}, selection bias will be present in nonexperimental settings, i.e., $E(U_{it}|X_{it},D_i)$ will not be zero. Its presence implies that

$$E(Y_{it} \mid D_i, Z_i) = \beta_0 + X_{it}\beta + D_i\alpha_t + E(U_{it} \mid D_i, X_i) \neq X_{it}\beta + D_i\alpha_t \tag{5.6b}$$

so that an ordinary (or nonlinear) least squares (OLS) regression of Y_{it} on X_{it} and D_i will *not* yield consistent estimates of α_t (or β). (Why not?)

Model-Based Approaches:

Suppose that the objective of the agent is to maximize the Present Value of their life time earnings, where

Y_{0it} for $t = 1,\ldots,k$. (pre-training earnings)

(Y_{0it},Y_{1it}) for $t = k+1,\ldots,T$. (post-training earnings)

c_i direct cost of training in period k.

$$\max_{D_i} E\left[\sum_{j=1}^{T-k}\frac{Y_{1i,k+j}}{(1+r)^j} - c_i - \sum_{j=0}^{T-k}\frac{Y_{0i,k+j}}{(1+r)^j}\middle| I_{ik}\right]$$

or

$$\max_{D_i} E\left[\sum_{j=1}^{T-k}\frac{\alpha_{i,k+j}}{(1+r)^j} - c_i - Y_{0ik}\middle| I_{ik}\right]$$

which implies the following decision-rule for taking training:

$$D_i = \begin{cases} 1, \text{ if } E\left[\sum_{j=1}^{T-k}\frac{\alpha_{ik+j}}{(1+r)^j} - c_i - Y_{0ik}\middle| I_{ik}\right] > 0, \\ 0, \text{ otherwise.} \end{cases}$$

Different estimators of program participation and training effects play off of different assumptions about the structure of the earnings processes, costs of training, and information sets.

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

2.1 General Considerations

One class of methods are *control function estimators*, where $h(d_i,X_i,Z_i,\pi)$ is the *control function* and where π is a vector of parameters. The control function adjusts for the dependence between d_i and U_{it} so that

$$Y_{it} = \beta_0 + X_{it}\beta + D_i\alpha_t + h(d_i, X_i, Z_i, \pi) + U'_{it} \tag{5.7}$$

where, when $h(D_i,X_i,Z_i,\pi) = E(U_{it}|D_i,X_i,Z_i)$, $E(U'_{it}\,|D_i, X_i, Z_i) = 0$.

2.2 Selection on Observables

One variant within the control function class arises when the dependence between U_{it} 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

$$E(U_{it}\,|D_i, X_i) \neq 0 \text{ and } E(U_{it}\,|D_i, X_i, Z_i) \neq 0$$

It is the case that

$$E(U_{it}\,|D_i, X_i, Z_i) = E(U_{it}\,|, X_i, Z_i)\,.$$

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 (5.2) with an appropriate control function, i.e.,

$$Y_{it} = \beta_0 + X_{it}\beta + D_i\alpha_t + h(X_i, Z_i) + U'_{it} \tag{5.8}$$

and utilizing least squares to estimate (5.8), will yield consistent estimates of α_t. Thus the selection bias problem (i.e., the correlation between U_{it} and D_i) can be eliminated by accounting for the observable factors that influence the selection process.

One functional form for $h(\,)$—see : Barnow, Cain and Goldberger (1980)—is:

$$h(X_{it}, Z_i) = X_{it}\theta_{1t} + Z_i\theta_{2t} \tag{5.9}$$

where θ_{1t} and θ_{2t} are parameter vectors. This also called the *regression discontinuity design quasi-experimental estimator,* which has frequently been used in the educational evaluation literature.

A related strategy has been proposed by Rosenbaum and Rubin (1983). They use $p(Z_i) = 1 - F(-Z_i\gamma)$, the *propensity score control function:*

$$h(X_{it}, Z_i) = p(Z_i) \tag{5.10}$$

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.

2.3 The Mills Ratio (or "Heckman") Procedure

Historically, a commonly used control function estimator proposed by Heckman (1976), is based on the assumption that the joint distribution of U_{it} and V_i is bivariate normal. Under this set of assumptions, $h(D_i,X_i,Z_i,\pi) = E(U_{it}|D_i,X_i,Z_i)$ is proportional to the Mills ratio, i.e.,

$$h(D_i, X_i, Z_i, \pi) = \frac{\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)} \tag{5.11}$$

where ϕ and Φ are the standardized normal density and distribution functions, respectively. Under the joint normality assumption, the inclusion of the Mills Ratio in the outcome equation (5.2), i.e.,

$$Y_{it} = \beta_0 + X_{it}\beta + D_i\alpha_t + \theta_t\left(\frac{\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)}\right) + U'_{it} \tag{5.12}$$

where θ_t is a parameter to be estimated.

A consistent estimate of α_t when (5.12) is estimated by least squares. In practice, a two-stage procedure is used since γ is generally unknown and must also be estimated.

3. Longitudinal and/or Repeated Cross-Section Data Estimators

3.1 Before and After Estimators

Suppose we compare the outcomes of trainees (the treated group) *after* training (treatment) with their outcomes *before* receipt of treatment. That is, we use pre-training outcomes of treated to proxy for their counterfactual post-training outcomes. Suppose that training occurs in period l and $t' < l < t$ and we have data on Y_{it} for trainees in periods t' and t. Recall that

$$Y_{1it} \equiv Y_{it} \,|\, D{=}1 \tag{5.13}$$

but we are missing $Y_{0it} \,|\, D{=}1$ Suppose we use $Y_{it'} \,|\, D{=}1$ to measure it. For this to be valid, we must assume that

$$E(Y_{0t} - Y_{0t'} \,|\, D = 1) = 0\,. \tag{5.14}$$

If (5.14) holds, then we can use

$$\overline{Y}_{Tt} - \overline{Y}_{Tt'} \tag{5.15}$$

to estimate $\tilde{\alpha} = E(\alpha \,|\, D = 1)$ since $E(\overline{Y}_{Tt} - \overline{Y}_{Tt'}) = \tilde{\alpha} = E(\alpha \,|\, D = 1)$.

Note that Y_{ik} can be written as:

$$Y_{ik} = \beta_0 + \beta X_i + \alpha D_i + \varepsilon_{ik} \tag{5.16}$$

where $k = t,t'$ and note that $D_i = 0$ in period t' but $D_i = 1$ for $t > l$.

3.2 Fixed Effect Estimators

Suppose that Y_{ik} changes over time due to factors other than training. (These would include factors that change over a person's life cycle and/or temporal changes in environmental conditions, such as the state of the labor market.) Recall the specification in (5.2). Let's generalize it in the following way

$$Y_{it} = \beta_0 + X_{ik}\beta + D_i\alpha_k + \lambda_k\phi_i + \nu_{ik} \tag{5.17}$$

where U_{ik} is now assumed to have the form

$$U_{ik} = \lambda_k\phi_{1i} + \nu_{ik}\,, \tag{5.18}$$

and λ is a fixed parameter, ϕ_{1i} is a zero mean, person-specific component or "fixed effect," and ν_{ik} is serially uncorrelated random variable that is independent of ϕ_{1i}. [Note that λ is often set to 1 in the literature on fixed effects estimation.]

In this specification, ϕ_i, but not ν_{ik}, assumed to influence program participation decision. Thus,

$$E(U_{it} - U_{it'} \,|\, D_i, X_{it'}, X_{it}) = 0, \text{ for all } t,t',t > k > t'. \tag{5.19}$$

Furthermore, suppose that we assume that

$$\lambda_{t'} = \lambda_t \tag{5.20}$$

It follows that consistent estimates of α_t obtained by estimating

$$Y_{it} - Y_{it'} = d_i\alpha_t + X_{it}\beta_t - X_{it'}\beta_{t'} + (\nu_{it} - \nu_{it'}), \text{ for } t > k > t'. \tag{5.21}$$

Under the assumptions of the fixed effect model, estimating (5.21) by least squares yields a consistent estimator of α_t.

Note that one does not need to have data on a comparison group, so long as one has before and after training data for the trainees. At the same time, one can use data on trainees (D = 1) and a comparison group (D = 0).

3.3 The Difference-in-Difference Estimator

Consider the following version of the model for outcomes:

$$Y_{ik} = \beta_0 + D_{ik}\alpha + \mu_{ik} + \phi_i + \nu_{ik} \tag{5.22}$$

where μ_{ik} denotes time varying variables and the other parameters and random variables are as before. The fixed-effects estimator can be applied to (5.22) when one has longitudinal data on the same individuals. However, suppose that we only have data from *repeated cross-sections*. That is, we have samples of individuals in periods t' and t who:

(a) received treatment in period l, i.e., $D_{it'}=0$ but $D_{it}=1$ (trainees or T's)
(b) did not receive it before or after l, i.e., $D_{it'}=0$ and $D_{it}=0$ (non-trainees or N's)
but the samples are drawn from cross-sections of each group in t' and t.

The *Difference-in-Difference* (Diff-in-Diff) estimator assumes that

(i) the relationship between Y_{ik} and D is given by (5.22)

(ii) $E(\bar{\phi}_{Tt} - \bar{\phi}_{Tt'}) = 0$ and $E(\bar{\phi}_{Nt} - \bar{\phi}_{Nt'}) = 0$ (*time invariant group composition*)

(iii) $E(\bar{\mu}_{Tt} - \bar{\mu}_{Tt'}) = E(\bar{\mu}_{Nt} - \bar{\mu}_{Nt'})$ (*common group trends*)

Then it follows that

$$\begin{aligned} E(\Delta\bar{Y}_T) &= \alpha + \Delta\mu \\ E(\Delta\bar{Y}_N) &= \Delta\mu \end{aligned} \tag{5.23}$$

and

$$E(\Delta\bar{Y}_T - \Delta\bar{Y}_N) = \alpha \tag{5.24}$$

where $\Delta\bar{x} \equiv \bar{x}_t - \bar{x}_{t'}$.

One also can apply OLS regression methods to implement the Diff-in-Diff method.

When one has *longitudinal* data, the fixed effect estimator amounts to differencing the dependent and independent variables, i.e.,

$$(Y_{it'} - Y_{it}) = \alpha(D_{it'} - D_{it}) + (\nu_{it'} - \nu_{it}) \tag{5.25}$$

or

$$(Y_{it'} - Y_{it}) = \alpha(D_{it'} - D_{it}) + \beta(X_{it'} - X_{it}) + (\nu_{it'} - \nu_{it}) \tag{5.25'}$$

If one controls for X_{ik}, and then and applying OLS to either of the above equations.

When one has repeated *cross-sectional* data on the treatment and non-treatment groups, one estimates the following equation

$$Y_{ik} = \beta_0 + \beta_1 P_k + \beta_2 T_i + \beta_3' X_{ik} P_k + \alpha P_k T_i + \beta_5' X_{ik} P_k + \nu_{ik} \tag{5.26}$$

using OLS, where

$$P_k = \begin{cases} 1, & \text{if } k = t' \\ 0, & \text{otherwise} \end{cases}.$$

Violations of Assumptions (i), (ii) or (iii):

Violations of these assumptions will invalidate the Diff-in-Diff estimator.

Violation of Assumption (ii):

In this case, consider the possibility that the sample of cross-sections changes over time. In the case of longitudinal data, suppose that there is sample attrition, i.e., data on same firms is not available for both time periods. In the case of repeated cross-sectional data, the composition may change between the two time periods and this change may be due to changes in the law, i.e., firms enter or exit in response to treatment. Then, either $E(\bar{\phi}_{Tt} - \bar{\phi}_{Tt'}) \neq 0$ or $E(\bar{\phi}_{Nt} - \bar{\phi}_{Nt'}) \neq 0$.

Violation of Assumptions (i) and (iii):

It is easier to think about violations of Assumption (iii), but this could come about because of the assumed functional form in (5.22) does not hold, i.e., there are non-linearities involving D_{ik} and μ_i's. The *key assumption* is that the change in treatment group over time, net of the influence of the treatment, is captured by the change in the outcomes for the comparison group. Suppose this is not the case. Consider, for example, a revised version of (5.26).

$$Y_{ik} = \beta_0 + \beta_1 P_k + \beta_2 T_i + \beta_3' X_{ik} P_k + \alpha P_k T_i + \beta_5' X_{ik} P_k + \gamma T_i X_{ik} P_k + \nu_{ik} \tag{5.27}$$

Here the assumption is that the there is a change in the outcome for the treatment group over time, *over and above the impact of the treatment itself*. As a result, differencing in longitudinal data or differencing the averages in repeated cross-sectional data will not eliminate the bias.

3.4 The Random Growth Estimator

Suppose U_{it} is of the following form,

$$U_{it} = \phi_{1i} + t\phi_{2i} + \nu_{it} \tag{5.28}$$

where ϕ_{1i} is as before and ϕ_{2i} is a person-specific growth rate for the outcome variable Y_{it}. Again, suppose that (ϕ_{1i}, ϕ_{2i}) uncorrelated with ν_{it} for all i and t. In this model, individual outcomes are allowed to differ both in levels and in rates of growth. Program participation decisions depend on ϕ_{1i} and ϕ_{2i} so that U_{it} correlated with d_i. Consider the following transformation of the outcome

equation,

$$[Y_{it} - Y_{it'}] - (t - t')[Y_{it'} - Y_{it'-1}] = \tag{5.29}$$
$$d_i\alpha_t + [X_{it}\beta_t - X_{it'}\beta_{t'}] - (t - t')[X_{it'}\beta_{t'} - X_{it'-1}\beta_{t'-1}]$$
$$+ [\nu_{it} - \nu_{it'}] - (t - t')[\nu_{it'} - \nu_{it'-1}]$$

where $t > k > t'$. Estimation of (5.29) by least squares will yield consistent α_t.

3.5 The Autoregressive Disturbance Estimator

Historically, 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,

$$U_{it} = \rho U_{i,t-1} + \nu_{it}\,, \tag{5.30}$$

where ρ is a parameter (assumed to not equal ± 1) and ν_{it} is a mean zero independently distributed random disturbance. Once again a transformation of the outcome equation can be used to eliminate the selection bias problem.

$$Y_{it} = \rho^{t-t'}Y_{it'} + X_{it}\beta_t - \rho^{t-t'}X_{it'}\beta_{t'} + (1 - \rho^{t-t'})\alpha_t d_i + \left[\sum_{j=0}^{t-(t'+1)} \rho^j \nu_{i,t'-1}\right] \tag{5.31}$$

Under the assumptions for the autoregressive model, nonlinear least squares methods applied to (5.31) will yield consistent estimates of α_t.

4. Instrumental Variables (IV) Estimators

IV estimators presume the existence of variables (elements of Z_i, for example) which are independent of the outcome equation disturbances, U_{it}^0, U_{it}^1, but are correlated with the training/treatment status of individuals, d_i.

- Assumptions for Instrumental Variable:

  A1: Conditional on X, $W_i \in Z_i$ is uncorrelated with unobservables (U_{it}^0, V_i) and (U_{it}^1, V_i).

  A2: Conditional on X, D_i is non-trivial function of W_i.

- Implications of Assumptions:

  A2 implies

  $$E(D|X,W) = \Pr(D = 1|X,W) \neq \Pr(D = 1|X)$$

  A1 states that *W* has *no* impact on *Y* through unobservables, *U*'s, but only through its influence on *D*. That is, *W* helps "trace out" influence of just *D* on *Y*.

4.1 Homogeneous Treatment Effect Case

Case where $\alpha_{it}(X_i) = \alpha_t$, for all i and X_i. Standard result is:

$$\alpha_{IV} = \frac{Cov(Y_i, W_i)}{Cov(W_i, d_i)} \tag{5.32}$$

Or consider (linear) projection of D on W:

$$D_i = \pi_0 + \pi_1 W_i + e_i \tag{5.33}$$

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

$$\hat{D}_i = \pi_0 + \pi_1 W_i \tag{5.34}$$

Then:

$$Y_{it} = X_{it}\beta_t + \hat{D}_i \alpha_t + U'_{it}, \tag{5.35}$$

where $U'_{it} = U_{it} - e_i$. Use OLS on (5.35) to estimate α_{IV}.

Note that with IV estimators, no explicit distributional assumptions about U_{it} or V_i need to be made.

4.2 Heterogeneous Treatment Effect

Case where $\alpha_{it}(X_i)$ varies with i. Need further assumptions.

A3: Selection of D by agents does not depend on $\alpha_{it}(X_i) - \alpha_t(X_i)$ $[= U^1_{it} - U^0_{it}]$, *the (unobserved) gain from treatment*.

This assumption holds if individuals are no more knowledgeable about gain from treatment than is the econometrician.

It follows that:

$$E[U^1_{it} - U^0_{it} \,|\, X_i, W_i, D_i] = E[D_i[\alpha_{it}(X_i) - \alpha_t(X_i)] \,|\, X_i, W_i] = 0 \tag{5.36}$$

and given A1 and A2, the IV estimator defined in (5.32) identifies the average treatment effect, $E(\alpha_t|X)$.

However, if A3 fails to hold, agents know and use gain, $\alpha_{it}(X_i) - \alpha_t(X_i)$, in selection of D. As a result, the error in the outcome equation – U^*_{it} in

$$Y_{it} = g_t^0(X_i) + \alpha_t(X_i)D_i + U_{it}^* \tag{5.1a}$$

where $U_{it}^* \equiv U_{it}^0 + D_i(U_{it}^1 - U_{it}^0)$ is correlated with *W*, since *W* is correlated with *D*. In particular, now shifts in *W* not only cause shifts in *D*, but also cause shifts in *Y*, through U_{it}^*, which confounds being able to identify $\alpha_t(X_i)$.

Thus, in this more general case, IV estimator of the treatment effect is inconsistent without further assumptions.

4.3 Local Average Treatment Effect (LATE)

Imbens and Angrist (1994) propose a way to deal with the above problem. They do so, in essence, by changing the parameter of interest and then adding an additional assumption about the nature of how *W* affects *D*.

A2′: Conditional on X, the decision rule governing D is a monotonic function of W.

The idea now is that we assume that changes in the instrument, *W*, result in changes in *D* in a monotonic way, i.e., either always increases (never decreases) or always decreases (never increases) the probability that *D* = 1.

In addition, we define a "localized" version of A1:

A1′: Conditional on X and $W_i = w_i$, $W_i \in Z_i$ *is uncorrelated with unobservables* (U_{it}^0, V_i) *and* (U_{it}^1, V_i).

The difference between A1 and A1′ is that now we condition on a particular value of *w* and only require that the instrument, *W*, is uncorrelated with unobservables, (U_{it}^0, V_i) *and* (U_{it}^1, V_i), determining potential outcomes and *D*.

Then it follows that:

$$\begin{aligned}
&E(Y_{it} \mid X_i, W_i = w) - E(Y_{it} \mid X_i, W_i = w') \\
&\quad = E[D_i(w)Y_{it}^1 + (1 - D_i(w))Y_{it}^0 \mid X_i, W_i = w] - \\
&\qquad E[D_i(w')Y_{it}^1 + (1 - D_i(w'))Y_{it}^0 \mid X_i, W_i = w'] \\
&\quad = E[(D_i(w) - D_i(w'))(Y_{it}^1 - Y_{it}^0)] \\
&\quad = E[Y_{it}^1 - Y_{it}^0 \mid X_i, D_i(w) - D_i(w') = 1] \Pr[D_i(w) - D_i(w') = 1] \\
&\qquad + E[Y_{it}^1 - Y_{it}^0 \mid X_i, D_i(w) - D_i(w') = -1] \Pr[D_i(w) - D_i(w') = -1]
\end{aligned} \tag{5.37}$$

where the second line follows from A1′. Now it follows from A2′, i.e., that either $D_i(w) \geq D_i(w')$ or $D_i(w) \leq D_i(w')$, that either

$$\Pr[D_i(w) - D_i(w') = 1] \text{ or } \Pr[D_i(w) - D_i(w') = -1]$$

equals zero for everyone, i.e., the change from $W = w$ to $W = w'$ shifts people to treatment ($D = 1$) or not treatment ($D = 0$), *but not both*. Thus, if we suppose that $D_i(w) \geq D_i(w')$, then $\Pr[D_i(w) - D_i(w') = -1] = 0$ and it follows from (5.37) that:

$$E[Y_{it}^1 - Y_{it}^0 \,|\, X_i, D_i(w) - D_i(w') = 1] = \frac{E(Y_{it} \,|\, X_i, W_i = w) - E(Y_{it} \,|\, X_i, W_i = w')}{\Pr(D_i = 1 | W_i = w) - \Pr(D_i = 1 | W_i = w')} \tag{5.38}$$

which is, by definition, the Local Average Treatment Effect (LATE).

First, note that this treatment effect depends on "changers," i.e., those who would change from one value of D to another, in response to a shift in W from w to w'. In general, we don't observe which individuals are changers!

Second, note that the LATE depends on the values of W, i.e., on w and w'. We get different treatment effects if these values change. This is why LATE is local.

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 idea is to match the members of these two groups based on their observables and, under conditions noted below, whether an agent received the treatment is random, much like a randomized experiment.

Matching does not require exclusion restrictions or particular specifications of the treatment decision rules or of the functional forms of the outcome equations.

5.1 Assumptions Required for Matching Estimators

A1: (Unconfoundedness or Conditional Independence) *Conditional on the set of observables, X, the potential outcomes are independent of treatment status, i.e.,*

$$(Y^0, Y^1) \perp D | X \tag{5.39}$$

A2: (Overlap or Common Support):

$$0 < \Pr(D = 1|X) < 1. \tag{5.40}$$

Assumption A1 is just the conditional independence invoked in the selection on observables models discussed earlier. Assumption A2 simply says that in our data, the probability of treatment, given X, cannot be 0 or 1. With these two assumptions, the Average Treatment Effect (ATE), $E(\alpha_{it}(X))$, is identified.

To estimate the Average Treatment on the Treated (ATT), $E(\alpha_{it}(X)|D = 1)$, one requires a weaker version of A1:

A1′: (Unconfoundedness or Conditional Independence for Non-Treated Group) *Conditional on the set of observables, X, the non-treated potential outcomes are independent of treatment status, i.e.,*

$$Y^0 \perp D \big| X \tag{5.39′}$$

5.2 Propensity Score Matching

Because matching on all elements of X is problematic, the greater the dimension of X, one often exploits a result due to Rosenbaum and Rubin (1983, 1984), which in place of A1 or A1′, one can condition on the *propensity score*, i.e., $p(X_i) = \Pr(D_i = 1|X_i)$ to get new versions of these assumptions:

A3: Conditional on the propensity score, $p(X)$, the *potential outcomes are independent of treatment status, i.e.,*

$$(Y^0, Y^1) \perp D \big| p(X) \tag{5.41}$$

A3′: Conditional on the propensity score, $p(X)$, the non-treated potential outcomes are independent of treatment status, i.e.,

$$Y^0 \perp D \big| p(X) \tag{5.41′}$$

Rosenbaum and Rubin show that the unconfoundedness or conditional independence assumptions in A1 and A1′ hold when one conditions on the propensity score, rather than X, i.e., they prove that conditioning on $p(X)$ is equivalent to conditioning on X.

5.3 Estimation

Alternative ways to do this, but here is a example using matching, based on X or $p(X)$ for estimating the Average Treatment on the Treated (ATT):

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. 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 ATT can be formed either by a simple mean of the pairs, i.e.,

$$\hat{\alpha}_t = \frac{1}{N_E} \sum_{i=1}^{N_E} (Y_{it} - Y_{j^*(i)}) \tag{5.42}$$

6. Bounds on Treatment Effects

(See Manski, 1989, 1990)

Consider the Average Treatment Effect:

$$\alpha^* \equiv E(Y_T - Y_C \,|\, X = x) \tag{5.43}$$

where we condition on a set of observables, x. α^* could be estimated (identified) if we *randomly assigned* individuals to either treatment $Z^* = T$ or $Z^* = C$.

But, consider case in which we don't have experimental data and individuals selectively choose treatment $D = 1$ or $D = 0$, where the numbers 1 and 0 correspond with the treatments T and C. In general, the average treatment effect is given by:

$$\begin{aligned} \alpha^* &\equiv E(Y_1 \,|\, x) - (Y_0 \,|\, x) \\ &= E(Y_1 \,|\, x, Z = 1)P(Z = 1 \,|\, x) \\ &\quad + E(Y_1 \,|\, x, Z = 0)P(Z = 0 \,|\, x) \\ &\quad - E(Y_0 \,|\, x, Z = 1)P(Z = 1 \,|\, x) \\ &\quad + E(Y_0 \,|\, x, Z = 0)P(Z = 0 \,|\, x) \end{aligned} \tag{5.44}$$

The problem is that we do not observe $E(Y_1 \,|\, x, D = 0)$ or $E(Y_0 \,|\, x, D = 1)$. (The rest of the stuff can be estimated (identified) from observable data.)

Without further restrictions, we can't learn much about the treatment effect in (5.43).

But, suppose either Y is bounded or a discrete random variable.

Bounds when Outcomes are Bounded:

Suppose $Y_k \in [K_{kLx}, K_{kUx}]$, for $k = 0, 1$. Then it follows that:

$$\begin{aligned} E(Y_1 \,|\, x) \in [&E(Y_1 \,|\, x, D = 1)P(D = 1 \,|\, x) + K_{1Lx}P(D = 0 \,|\, x), \\ &E(Y_0 \,|\, x, D = 0)P(D = 1 \,|\, x) + K_{1Ux}P(D = 0 \,|\, x)] \end{aligned} \tag{5.45}$$

The lower bound is the value of $E(Y_1 \,|\, x)$ is the value it takes if Y_1 equals its lower bound (K_{1Lx}) for all those who choose treatment $D = 0$ and similarly the upper bound on $E(Y_1 \,|\, x)$ is given by using the upper bound (K_{1Lx}) for Y_1 for those who choose treatment $D = 0$.

The same logic applies to bounding $E(Y_0 \,|\, x)$, using K_{0Lx} and K_{0Ux}.

It follows that

$$\begin{aligned}\alpha^* \in [&K_{0Lx}P(D=1|x)+E(Y_0|x,D=0)P(D=0|x)\\&-E(Y_1|x,D=1)P(D=1|x)-K_{1Ux}P(D=0|x),\\&K_{0Ux}P(D=1|x)+E(Y_0|x,D=0)P(D=0|x)\\&-E(Y_1|x,D=1)P(D=1|x)-K_{1Lx}P(D=0|x)]\end{aligned} \tag{5.46}$$

Note that the width of the bound is

$$\begin{aligned}w(x) \equiv &(K_{0Ux}-K_{0Lx})P(D=1|x)\\&+(K_{1Ux}-K_{1Lx})P(D=0|x)\end{aligned} \tag{5.47}$$

and if the bounds on Y_1 and Y_0 are the same the width of the bound is:

$$w(x) \equiv K_{Ux}-K_{Lx}$$

Note that these bounds are not all that informative, in that they necessarily cover zero. That is, it does not identify the sign of the treatment effect.

Bounds when outcomes are binary:

Suppose that $Y_k = 0$ or 1. It follows that $K_{Lx} = 1$ and $K_{Ux} = 0$, so that α^* must lie in the interval [-1,1]. But note that the expected value of a binary outcome is itself the probability that the indicator = 1. So, it follows that the bound on the treatment effect reduces to:

$$\begin{aligned}\alpha^* \in [&0\cdot P(D=1|x)+\Pr(Y_0=1|x,D=0)P(D=0|x)\\&-\Pr(Y_1=1|x,D=1)P(D=1|x)-1\cdot P(D=0|x),\\&1\cdot P(D=1|x)+\Pr(Y_0=1|x,D=0)P(D=0|x)\\&-\Pr(Y_1=1|x,D=1)P(D=1|x)-0\cdot P(D=0|x)]\\\in [&\Pr(Y_0=1|x,D=0)P(D=0|x)\\&-\Pr(Y_1=1|x,D=1)P(D=1|x)-P(D=0|x),\\&P(D=1|x)+\Pr(Y_0=1|x,D=0)P(D=0|x)\\&-\Pr(Y_1=1|x,D=1)P(D=1|x)]\end{aligned}$$

where the width of the bound is 1, in which case the bound is one-half the size of the difference between the maximum width of the bound.

Tightening the Bounds with Additional Assumptions about Selection Process or other restrictions:

Manski and others consider cases in which one wishes to impose additional restrictions on the selection process and/or outcomes. These add additional information and thus, tighten the

bounds. If sufficient information is added, then the bounds collapse to points and one achieves point identification, as in the non-experimental estimators we considered last class.

Imposing Assumption of Treatment Choice following Comparative Advantage:

Suppose it is the case that treatment-selection is based on an individual selecting based on comparative advantage in the following sense:

$$D = k \text{ iff } Y_k \geq Y_{\sim k} \tag{5.48}$$

Then it follows that the bounds can be tightened. Condition (5.48) implies that:

$$\begin{aligned} E(Y_1 \mid x, D=0) &= E(Y_1 \mid x, Y_1 \leq Y_0) \\ &\leq E(Y_1 \mid x, Y_1 > Y_0) \\ &= E(Y_1 \mid x, D=1) \end{aligned}$$

and

$$\begin{aligned} E(Y_0 \mid x, D=1) &= E(Y_0 \mid x, Y_0 \leq Y_1) \\ &\leq E(Y_0 \mid x, Y_0 > Y_1) \\ &= E(Y_0 \mid x, D=0) \end{aligned}$$

Thus, $E(Y_1 \mid x, D=1)$ and $E(Y_0 \mid x, D=0)$ are upper bounds on $E(Y_1 \mid x, D=0)$ and $E(Y_0 \mid x, D=1)$, respectively.

Continuing to assume the existence of bounds on Y_k, the treatment effect bounds under this comparative advantage assumption tighten and are given by:

$$\begin{aligned} \alpha^* \in [&K_{0Lx} P(D=1 \mid x) + E(Y_0 \mid x, D=0) P(D=0 \mid x) - E(Y_1 \mid x, D=1), \\ &E(Y_0 \mid x, D=0) - E(Y_1 \mid x, D=1) P(D=1 \mid x) - K_{1Lx} P(D=0 \mid x)] \end{aligned} \tag{5.49}$$

which may (or may not) be tight enough on one side of zero to allow the sign of the treatment effect to be identified.

Manski (1989, 1990) considers other examples of tighten the bounds by imposing additional information.

References

Barnow, B., G. Cain, and A. Goldberger (1980). "Issues in the Analysis of Selectivity Bias," *Evaluation Studies*, Vol. 5, ed. by E. Stromsdorfer and G. Farkas, 1980, pp. 42-59.

Barros, R. (1987). Two Essays on the Nonparametric Estimation of Economic Models with Selectivity Using Choice-Based Samples, Unpublished Ph.D. Dissertation, University of Chicago, May 1987.

Heckman, J. (1978). "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, Vol. 46, 1978, pp. 931-961.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, 1979, pp. 153-161.

Heckman, J., and V. J. Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, Vol. 84, No. 408, December 1989, pp. 862-880. (Includes comments by Paul Holland and Robert Moffitt and Rejoinder.)

Heckman, J. and R. Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer, Cambridge University Press, 1985.

Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70, 1983, pp. 41-55.

Rosenbaum, P. and D. Rubin (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, Vol. 79, 1984, pp. 516-524.

Rosenbaum, P. and D. Rubin (1985). "Constructing A Control Group Using Multivariate Matched Sample Methods that Incorporate the Propensity Score," *The American Statistician*, Vol. 39, 1985, pp. 33-38.

Rubin, D. (1973a). "Matching to Remove Bias in Observational Studies, *Biometrics*, Vol. 29, 1973, pp. 159-183.

Rubin, D. (1973b). "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics*, Vol. 29, 1973, pp. 185-203.

Rubin, D. (1979). "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association*, Vol. 74, 1979, pp. 318-328.

Rubin, D. (1980). "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics*, Vol. 36, 1980, pp. 293-298.

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,

α_{it} be the impact of training on person i at time t,

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

$$\begin{aligned} Y_{it} &= Y^*_{it} + d_i\alpha_{it}, \quad t > k, \\ Y_{it} &= Y^*_{it}, \qquad\quad t < k. \end{aligned} \tag{2.1}$$

■ We focus on estimating the mean impact of training on the trained,

$$E(\alpha_{it}|d_i=1) = E(Y_{it} - Y^*_{it}|d_i=1), \tag{2.2}$$

■ The mean post-program earnings of trainees is

$$E(Y_{it}|d_i=1) = E(\alpha_{it}|d_i=1) + E(Y^*_{it}|d_i=1). \tag{2.3}$$

The mean post-program earnings of nontrainees is

$$E(Y_{it}|d_i=0) = E(Y^*_{it}|d_i=0). \tag{2.4}$$

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

$$E(Y_{it}|d_i=1) - E(Y_{it}|d_i=0) = E(\alpha_{it}|d_i=1) + \left\{E(Y^*_{it}|d_i=1) - E(Y^*_{it}|d_i=0)\right\} \tag{2.5}$$

■ Selection Bias is present if

$$E(Y^*_{it}|d_i=1) \neq E(Y^*_{it}|d_i=0), \tag{2.6}$$

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

$$E(Y^*_{it}|d_i=1) = E(Y^*_{it}|d_i=0) = E(Y^*_{it}), \tag{2.7}$$

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:*

$$\Rightarrow \quad Y_{it}^{*} = X_{it}\beta + U_{it}, \tag{3.1}$$

where X_{it} vector of observed characteristics, U_{it} unobserved characteristics, and β vector of parameters.

Thus:

$$\Rightarrow \quad Y_{it} = X_{it}\beta + d_i\alpha_t + U_{it}, \text{ for } t = 0,\ldots,T. \tag{3.2}$$

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

- Selection Bias is present if

$$\Rightarrow \quad E(U_{it}|d_i,X_i) \neq 0 \quad \text{so} \quad E(Y_{it}|d_i,X_i) \neq X_{it}\beta + d_i\alpha_t.$$

Ordinary least squares regression of Y_{it} on X_{it} and d_i yields inconsistent estimates of α_t (or β).

- ALTERNATIVE NONEXPERIMENTAL ESTIMATORS AND THE ASSUMPTIONS ON WHICH THEY ARE BASED:

B. *Selection on Observables*

- Suppose:

$$\Rightarrow \quad E(U_{it}|d_i,X_i) \neq 0 \quad \text{and} \quad E(U_{it}|d_i,X_i,Z_i) \neq 0$$

but

$$\Rightarrow \quad E(U_{it}|d_i,X_i,Z_i) = E(U_{it}|X_i,Z_i). \tag{3.6}$$

Controlling for the observed selection variables (Z_i) solves the selection bias problem.

■ Therefore, using (3.6), one can form estimators by noting that:

$$E(Y_{it}|d_i,X_i,Z_i) = X_{it}\beta + d_i\alpha_t + E(U_{it}|X_i,Z_i) \tag{3.7}$$

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

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

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

• Variant I: [Assuming $E(\alpha_{it}|d_i=1,X_i,Z_i) = \alpha_t$]

$$\Rightarrow \quad Y_{it} = C_i\delta_t + d_i\alpha_t + \tilde{U}_{it} \tag{3.8}$$

where C_i denotes the vector of all variables included in either X_i or Z_i, $\tilde{U}_{it} = U_{it} - E(U_{it}|d_i,C_i) = U_{it} - E(U_{it}|C_i)$ and δ_t is a parameter vector.

• Variant II: [Assuming $E(\alpha_{it}|d_i=1) = C_i\theta_t$]

$$\Rightarrow \quad Y_{it} = C_i\delta_t + d_i(C_i\theta_t) + \tilde{U}_{it}. \tag{3.8'}$$

• Consistent estimators of α_t can be obtained by ordinary least squares.

C. *Selection on Unobservables*

■ Suppose dependence between d_i and U_{it} not eliminated even after controlling for Z_i.

$$\Rightarrow \quad E(U_{it}|d_i,X_i) \neq 0 \text{ and } E(U_{it}|d_i,X_i,Z_i) \neq E(U_{it}|X_i,Z_i). \tag{3.9}$$

■ Fixed Effect (or First Difference) Estimator:

• Suppose

$$\Rightarrow \quad U_{it} = \phi_{1i} + \nu_{it},$$

where ϕ_{1i} is a zero mean, person-specific component and ν_{it} is random component. Selection assumed to depend on permanent component ϕ_{1i}.

Consistent estimates of α_t are obtained by regressing $Y_{it} - Y_{it'}$ on d_i and

$X_{it} - X_{it'}$.

- Variant I:

$$\Rightarrow \quad Y_{it} - Y_{it'} = d_i\alpha_t + (X_{it} - X_{it'})\beta + (\nu_{it} - \nu_{it'}), \text{ for } t > k > t'. \quad (3.11)$$

- Variant II:

$$\Rightarrow \quad Y_{it} - Y_{it'} = d_i(C_i\theta_t) + (X_{it} - X_{it'})\beta + (\nu_{it} - \nu_{it'}), \text{ for } t > k > t'. \quad (3.11')$$

■ Random Growth Estimator:

- Suppose

$$\Rightarrow \quad U_{it} = \phi_{1i} + t\phi_{2i} + \nu_{it}, \quad (3.12)$$

where ϕ_{2i} a person-specific growth rate component. Dependence between U_{it} and d_i assumed due to dependence between d_i and (ϕ_{1i},ϕ_{2i}).

OLS Regression on Transformation of (3.12), where pre-training earnings proxy for (ϕ_{1i},ϕ_{2i}), will yield consistent estimates. [Pudney (1982)]

- Variant I:

$$\Rightarrow \quad (Y_{it} - Y_{it'}) - (t-t')(Y_{it'} - Y_{i,t'-1}) = d_i\alpha_t + [(X_{it} - X_{it'}) - (t-t')(X_{it'} - X_{i,t'-1})]\beta + [(\nu_{it} - \nu_{it'}) - (t-t')(\nu_{it} - \nu_{i,t'-1})], \quad (3.13)$$

for $t > k > t'$.

- Variant II:

$$\Rightarrow \quad (Y_{it} - Y_{it'}) - (t-t')(Y_{it'} - Y_{i,t'-1}) = d_i(C_i\theta_t) + [(X_{it} - X_{it'}) - (t-t')(X_{it'} - X_{i,t'-1})]\beta + [(\nu_{it} - \nu_{it'}) - (t-t')(\nu_{it} - \nu_{i,t'-1})], \quad (3.13')$$

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 α_t (or θ_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.
- Use comparison groups for each drawn from the March 1976 and March 1977 Current Population Survey (CPS).
- 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 nonlinear 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:
  - High School Dropouts: -$48 in 1978 and $9 for 1979. Neither statistically significant.
  - 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

Variable	Description
Earnings variables	
SSEARN72	SSA earnings in 1972 (in 1978 dollars)
SSEARN73	SSA earnings in 1973 (in 1978 dollars)
SSEARN74	SSA earnings in 1974 (in 1978 dollars)
SSEARN75	SSA earnings in 1975 (in 1978 dollars)
SSEARN78	SSA earnings in 1978 (in 1978 dollars)
SSEARN79	SSA earnings in 1979 (in 1978 dollars)
Background variables in *B1*	
BLKHIS	1 if black or Hispanic and 0 otherwise
SEX	1 for men and 0 for females
MARRIAGE	1 if married at enrollment for NSW participants or at March interview for CPS respondents and 0 otherwise
AGE	Age in years at enrollment for NSW participants or at March interview for CPS respondents
EDUC	Years of schooling completed at enrollment for NSW participants or at March interview for CPS respondents
URBAN	1 if in central-city standard metropolitan statistical area and 0 otherwise
7677ENR	1 if enrolled in 1977 for NSW participants or if interviewed in March 1977 for CPS respondents and 0 otherwise
Background variables in *B2*	
BLACK	1 if black and 0 otherwise
HISPANIC	1 if Hispanic and 0 otherwise
AGESQ	AGE squared
HOUSESIZE	Number of household members at enrollment for NSW participants or at March interview for CPS respondents
DEPEND	Number of dependents at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]
AGEKID	Age of youngest dependent at enrollment for NSW participants or at March interview for CPS respondents [used only for AFDC recipient (women) results]
Work history variables in *W1*	
SSEARNL1	Annual earnings (from SSA data) one year prior to enrollment for NSW participants or one year prior to interview of CPS respondents
SSEARNL2	Annual earnings (from SSA data) two years prior to enrollment for NSW participants or two years prior to the interview for CPS respondents
WORKWKS	Number of weeks worked in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
UEWKS	Number of weeks unemployed in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
AVEHRS	Average hours per week (when worked) in year prior to enrollment for NSW participants or in year prior to interview for CPS respondents
WELFARE	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
Work history variables in *W2*	
SSEARNL3	Annual earnings (from SSA data) three years prior to enrollment for NSW participants or three years prior to interview for CPS respondents
SSEARNL4	Annual earnings (from SSA data) four years prior to enrollment for NSW participants or four years prior to interview for CPS respondents
CLERSALE	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
SERVICE	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
PROFESSION	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
AFDC	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]

Table 2. Sample Means

Variable	High-school dropouts (youths)			AFDC recipients (women)		
	NSW samples		CPS sample	NSW samples		CPS sample
	Trainees	Controls		Trainees	Controls	
Earnings variables						
SSEARN72	192.7	228.9	201.3	971.3	1,085.6	1,041.0
SSEARN73	329.9	401.1	548.4	1,087.6	1,206.3	1,192.7
SSEARN74	581.1	630.6	1,036.6	895.3	1,000.8	1,201.9
SSEARN75	532.4	504.9	1,465.9	541.4	638.9	1,045.8
SSEARN78	1,704.0	1,751.5	3,654.8	2,007.8	1,588.9	1,841.8
SSEARN79	1,838.2	1,825.5	3,787.0	2,039.9	1,798.3	1,959.6
Background variables in *B1*						
BLKHIS	.918	.909	196	.955	.945	.500
SEX	.883	.864	483	.000	.000	.000
MARRIAGE	.044	.033	.285	.023	.042	.186
AGE	18.200	18.347	18.080	33.375	33.615	31.460
EDUC	9.616	9.677	10.658	10.307	10.272	11.133
URBAN	1.000	1.000	.240	.979	.981	.440
7677ENR	.645	.651	433	.729	.719	.485
Background variables in *B2*						
BLACK	.736	.706	110	.835	.817	.381
HISPANIC	.182	.203	.086	.120	.128	.120
HOUSESIZE	4.704	4.746	3.335	3.613	3.779	3.636
DEPEND				2.167	2.292	2.506
AGEKID				9.341	9.215	10.124
AGESQ				1,169.06	1,181.43	1,078.44
Work history variables in *W1*						
SSEARNL1	559.0	539.3	1,545.4	459.0	462.1	905.3
SSEARNL2	436.4	447.4	944.8	508.8	607.6	862.8
WORKWKS	9.3	9.3	21.4	3.3	3.2	11.0
UEWKS	10.3	11.2	2.2	11.7	13.3	1.9
HOURS	3.3	3.2	15.9	1.3	.9	7.2
WELFARE	33.8	33.7	121.6	93.9	91.6	169.7
Work history variables in *W2*						
SSEARNL3	357.1	400.6	521.9	711.8	816.5	872.5
SSEARNL4	180.6	198.1	194.7	711.9	769.1	741.1
CLERSALE	.101	.108	126	.072	.081	.109
SERVICE	.256	.212	228	.100	.084	.164
PROFESSION	.057	.040	016	.013	.016	.022
AFDC	.045	.043	146			
Number of observations	566	678	2,369	800	802	1,995
Number of cells	69	87	321	110	107	266

Table 3. Estimates of Training Effects for High-School Dropouts (youths)

Model and control variable sets	1978 earnings: Variant 1 (α_t)	1978 earnings: Variant 2 ($\bar{C}\theta_t$)	1979 earnings: Variant 1 (α_t)	1979 earnings: Variant 2 ($\bar{C}\theta_t$)
Nonexperimental estimates				
Linear control function estimates				
No control variables	−1,910		−1,917	
	(243)		(191)	
B1	−1,884	−1,827	−2,119	−2,092
	(247)	(246)	(342)	(300)
B1 + B2	−1,279	−1,079	−1,569	−1,498
	(273)	(295)	(239)	(319)
B1 + W1	−1,117	−1,146	−1,539	−1,447
	(246)	(263)	(343)	(372)
B1 + B2 + W1 + W2	−889	−889	−996	−1,331
	(328)	(380)	(442)	(388)
Fixed-effect estimates constructed with t' = 1972 pretraining earnings				
No control variables	−1,904		−1,910	
	(236)		(266)	
B1	−1,886	−1,831	−2,172	−2,070
	(242)	(201)	(277)	(275)
B1 + B2	−1,360	−1,227	−1,644	−1,647
	(270)	(291)	(309)	(330)
Fixed-effect estimates constructed with t' = 1974 pretraining earnings				
No control variables	−1,456		−1,462	
	(203)		(186)	
B1	−1,411	−1,370	−1,663	−1,636
	(227)	(228)	(301)	(269)
B1 + B2	−1,035	−964	−1,330	−1,383
	(255)	(276)	(326)	(312)
Random-growth estimates constructed with t' = 1973 and $t' - 1$ = 1972 pretraining earnings				
No control variables	−649		−446	
	(336)		(386)	
B1	−231	−235	−241	−236
	(414)	(416)	(475)	(477)
B1 + B2	−23	76	−85	−126
	(476)	(515)	(547)	(589)
Weighted average of estimates	−24		−154	
	(185)		(212)	
Random-growth estimates constructed with t' = 1974 and $t' - 1$ = 1973 pretraining earnings				
No control variables	−499		−267	
	(328)		(307)	
B1	−614	−589	−701	−659
	(431)	(436)	(410)	(515)
B1 + B2	−624	−777	−806	−850
	(497)	(537)	(586)	(630)
Weighted average of estimates	−616		−724	
	(426)		(502)	
Experimental estimates	−48		9	
	(144)		(173)	

NOTE: Standard errors are in parentheses.

pothesis $\bar{C}0_t = 0$ and for the sake of brevity are not reported.

Under the headings "Postprogram tests," we present P values for tests of the hypotheses $\alpha_t = 0$ and $\bar{C}0_{t \vee \bar{r}} 0$ ($t > k$), for earnings models fit on a pooled sample of experimental controls from the experiment ($d_i = 1$) and comparison-group members ($d_i = 0$). Again, tests of the vector hypothesis $\theta_t = 0$ are consistent with the test based on $\bar{C}\theta_t$ and are not reported here.

Under the heading "Model-restriction tests," we report P values for the hypotheses that extraneous Y values do not have statistically significant coefficients in the fixed-

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

Model and control variable sets	1978 earnings: Variant 1 (α_t)	1978 earnings: Variant 2 ($\bar{C}0_t$)	1979 earnings: Variant 1 (α_t)	1979 earnings: Variant 2 ($\bar{C}0_t$)
Nonexperimental estimates				
Linear control function estimates				
No control variables	157		79	
	(164)		(155)	
B1	686	726	494	534
	(192)	(194)	(193)	(195)
B1 + B2	231	638	−195	546
	(282)	(358)	(286)	(360)
B1 + W1	653	715	496	370
	(203)	(260)	(230)	(289)
B1 + B2 + W1 + W2	937	907	441	586
	(263)	(335)	(303)	(386)
Weighted average of estimates	374		238	
	(146)		(152)	
Fixed-effect estimates constructed with t' = 1972 pretraining earnings				
No control variables	231		153	
	(152)		(156)	
B1	699	736	508	544
	(185)	(188)	(193)	(195)
B1 + B2	938	1,124	512	1,032
	(275)	(353)	(287)	(362)
Fixed-effect estimates constructed with t' = 1974 pretraining earnings				
No control variables	475		397	
	(135)		(150)	
B1	713	693	522	500
	(168)	(170)	(179)	(184)
B1 + B2	946	800	520	708
	(250)	(321)	(268)	(328)
Random-growth estimates constructed with t' = 1973 and $t' - 1$ = 1972 pretraining earnings				
No control variables	494		460	
	(367)		(433)	
B1	78	44	−217	−263
	(463)	(473)	(546)	(557)
B1 + B2	486	−936	−15	−1,372
	(692)	(898)	(816)	(1,965)
Random-growth estimates constructed with t' = 1974 and $t' - 1$ = 1973 pretraining earnings				
No control variables	1,276		1,398	
	(356)		(471)	
B1	1,183	981	1,109	860
	(453)	(453)	(576)	(576)
B1 + B2	1,278	880	935	808
	(677)	(869)	(778)	(918)
Experimental estimates	440		267	
	(142)		(162)	

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, ϕ_{1i} and ϕ_{2i}, 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)

Control variable set	Probability values: Preprogram tests using preprogram earnings		Model-restriction tests: Preprogram earnings		1978 earnings		1979 earnings		Postprogram tests: 1978 earnings		1979 earnings	
	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$
	1975 earnings as dependent variable						*1978 or 1979 earnings as dependent variable*					
Linear control function estimators												
No control variables	.000								.000		.000	
B1	.000	.000							.000	.000	.000	.000
B1 + *B2*	.000	.012							.000	.000	.000	.000
B1 + *W1*	.000	.208							.000	.012	.000	.000
B1 + *B2* + *W1* + *W2*	.016	.336							.005	.632	.033	.000
	t = 1974 and t′ = 1972 earnings						*t = 1978 or 1979 and t′ = 1972 earnings*					
Fixed-effect estimators												
No control variables	.000		.000		.000		.000		.000		.000	
B1	.000	.000	.000	.000	.000	000	.000	.000	.000	.000	.000	.000
B1 + *B2*	.000	.019	.000	.000	.000	000	.000	.000	.000	.000	.000	.000
	t = 1975 and t′ = 1972 earnings						*t = 1978 or 1979 and t′ = 1974 earnings*					
Fixed-effect estimators												
No control variables	.000		.000		.502		.715		.000		.000	
B1	.000	.000	.000	.000	.076	081	.661	.629	.000	.002	.000	.001
B1 + *B2*	.000	.000	.000	.000	.023	021	.524	.817	.000	.002	.000	.001
	t = 1975, t′ = 1973, t′ − 1 = 1972 earnings						*t = 1978 or 1979, t′ = 1973, t′ − 1 = 1972 earnings*					
Random-growth estimators												
No control variables	.000		.035		.000		.000		.007		.042	
B1	.375	.173	.316	.080	.000	000	.000	.000	.329	.113	.427	.443
B1 + *B2*	.558	.128	.614	.042	.000	000	.000	.000	.798	.695	.622	.608
	t = 1975, t′ = 1974, t′ − 1 = 1973 earnings						*t = 1978 or 1979, t′ = 1974, t′ − 1 = 1973 earnings*					
Random-growth estimators												
No control variables	.000		.126		.000		.003		.139		.474	
B1	.301	.172	.809	.817	.090	105	.281	.353	.567	.398	.821	.417
B1 + *B2*	.352	.121	.909	.659	.070	146	.276	.546	.696	.312	.698	.169

effect and random-growth models fit on preprogram earnings [for a pooled sample of future trainees and controls ($d_i = 1$) and comparison-group members ($d_i = 0$)] and postprogram earnings [for a pooled sample of controls from the experiment ($d_i = 1$) and comparison-group members ($d_i = 0$)]. We do not see any compelling model restrictions for the linear control function estimator, so no test is reported.

The preprogram and postprogram tests and the model-restriction tests generally produce consistent findings. Linear control function and fixed-effect models are decisively rejected; the random growth model is not. (Though not reported in Table 5, the coefficient estimates associated with the preprogram and postprogram tests range from −2,128 to −167 for the linear control function models, from −2,186 to −274 for the fixed-effect models, and from −894 to −37 for the random-growth models.) Nevertheless, the tests for the model restrictions applied to the random-growth model fit on postprogram data are mixed. Using 1973 and 1972 earnings to proxy the unobserved components, ϕ_{1i} and ϕ_{2i}, the model is rejected on the postprogram data. Using 1973 and 1974 earnings to proxy the unobserved components, the model is not rejected on the postprogram data. Neither version of the random-growth model with regressors is rejected when it is fit on the preprogram sample that combines future participants (trainees and controls) and comparison-group members.

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_{it} ($t - t' = 6$ and 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, ϕ_{3i}, 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_{it} 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

$$U_{it} = \phi_{1i} + b_t\phi_{2i} + v_{it}, \qquad b_t \neq b_{t'},\ t \neq t'.$$

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.

Table 7. Estimates of Training Effects Using Modified Random-Growth Estimators for High-School Dropouts (youths)

Control variable set	1978 earnings				1979 earnings			
	Variant 1		Variant 2		Variant 1		Variant 2	
	α_t	$\omega_{t,t'-1}$	$\bar{C}\theta_t$	$\omega_{t,t'-1}$	α_t	$\omega_{t,t'-1}$	$\bar{C}\theta_t$	$\omega_{t,t'-1}$
	$t' = 1973$, $t' - 1 = 1974$ pretraining earnings							
Modified random-growth estimators								
B1	−191	5.836	−183	5.942	−277	6.749	−270	6.927
	(329)	(1.021)	(298)	(1.253)	(351)	(.918)	(379)	(1.124)
	$t' = 1974$, $t' - 1 = 1973$ pretraining earnings							
Modified random-growth estimators								
B1	−237	4.691	−201	4.573	−237	5.213	−213	5.011
	(367)	(1.001)	(361)	(1.116)	(385)	(1.023)	(370)	(.927)

NOTE: Standard errors are in parentheses.

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

Table 8. Specification Tests of Nonexperimental Estimators for AFDC Recipients (women)

Control variable set	Probability values											
	Preprogram tests using preprogram earnings		Model-restriction tests						Postprogram tests			
			Preprogram earnings		1978 earnings		1979 earnings		1978 earnings		1979 earnings	
	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$	$\alpha_t = 0$	$\bar{C}\theta_t = 0$
	1975 earnings as dependent variable								1978 or 1979 earnings as dependent variable			
Linear control function estimators												
No control variables	.000								.082		.246	
B1	.274	.817							.196	.098	.210	.121
B1 + *B2*	.000	.436							.217	.404	.192	.265
B1 + *W1*	.199	.021							.204	.358	.402	.806
B1 + *B2* + *W1* + *W2*	.139	.010							.435	.232	.973	.429
	$t = 1974$ and $t' = 1972$ earnings								$t = 1978$ or 1979 and $t' = 1972$ earnings			
Fixed-effect estimators												
No control variables	.000		.000		.000		.000		.031		.141	
B1	.606	.622	.000	.000	.000	.000	.000	.000	.455	.251	.490	.311
B1 + *B2*	.561	.131	.000	.000	.000	.000	.000	.000	.837	.074	.898	.045
	$t = 1975$ and $t' = 1972$ earnings								$t = 1978$ or 1979 and $t' = 1974$ earnings			
Fixed-effect estimators												
No control variables	.000		.000		.144		.014		.574		.893	
B1	.128	.824	.000	.000	.022	.047	.004	.009	.383	.340	.430	.423
B1 + *B2*	.307	.299	.000	.000	.019	.014	.001	.001	.701	.402	.772	.280
	$t = 1975$, $t' = 1973$, $t' - 1 = 1972$ earnings								$t = 1978$ or 1979, $t' = 1973$, $t' - 1 = 1972$ earnings			
Random-growth estimators												
No control variables	.021		.000		.000		.000		.738		.989	
B1	.016	.055	.000	.000	.000	000	.000	.000	.227	.243	.208	.208
B1 + *B2*	.183	.069	.000	.000	000	000	.000	000	.375	.074	.358	.075
	$t = 1975$, $t' = 1974$, $t' - 1 = 1973$ earnings								$t = 1978$ or 1979, $t' = 1974$, $t' - 1 = 1973$ earnings			
Random-growth estimators												
No control variables	.999		.102		VII - .002		004		.022		.008	
B1	.827	.974	.033	.021	.000	000	000	000	.124	.267	.135	.303
B1 + *B2*	.686	.985	.040	.037	.000	000	.000	.000	.267	.659	.277	.627

REFERENCES

Ashenfelter, O. and Card, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." Review of Economics and Statistics, 67, 648-660.

Barnow, B. (1987), "The Impact of CETA Programs on Earnings: A Review of the Literature." Journal of Human Resources, XXII, 157-193.

Barnow, B., Cain, G., and Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias." Evaluation Studies, eds. E. Stromsdorfer and G. Farkas, 1980, 5, 42-59.

Burtless, G. and Orr, L. (1986), "Are Classical Experiments Needed for Manpower Policy?" Journal of Human Resources, 21, 606-639.

Fraker, T. and Maynard, R. (1984), An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs. Princeton: MPR, Inc.

Fraker, T. and Maynard, R. (1987), "Evaluating Comparison Group Designs with Employment-Related Programs." Journal of Human Resources, 22, 194-227.

Heckman, J. and Robb, R. (1986), "Alternative Identifying Assumptions in Econometric Models of Selection Bias." in Advances in Econometrics: Innovations in Quantitative Economics, Essays in Honor of Robert L. Basmann, 5, eds. D. Slottje, Greenwich, CT: JAI Press, 243-287.

LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." American Economic Review, 76, 604-620.

Pudney, S. (1982), "Estimating Latent Variable Systems When Specification is Uncertain: Generalized Component Analysis and the Eliminant Method." Journal of the American Statistical Association, 77, 883-889.