View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

Institute for Research on Poverty
Discussion Paper no. 1061-95

Learning about Social Programs from Experiments with Random Assignment of Treatments

Charles F. Manski
Department of Economics
Institute for Research on Poverty
University of Wisconsin-Madison

March 1995

This research is supported by National Science Foundation Grant SES92-23220. I have benefited from the comments of Josh Angrist, Glen Cain, Jeff Dominitz, Arthur Goldberger, Joe Hotz, John Kennan, Bruce Meyer, Robert Moffitt, John Pencavel, Tomas Philipson, and anonymous reviewers.

Abstract

The importance of social programs to a diverse population creates a legitimate concern that the findings of evaluations be widely credible. The weaker are the assumptions imposed, the more widely credible are the findings. The classical argument for random assignment of treatments is viewed by many as enabling evaluation under weak assumptions, and has generated much interest in the conduct of experiments. But the classical argument does impose assumptions, and there often is good reason to doubt their realism.

Some researchers, finding the classical assumptions implausible, impose other assumptions strong enough to identify treatment effects of interest. In contrast, the recent literature examined in this article explores the inferences that may be drawn from experimental data under assumptions weak enough to yield widely credible findings. This literature has two branches. One seeks out notions of treatment effect that are identified when the experimental data are combined with weak assumptions. The canonical finding is that the average treatment effect within some context-specific subpopulation is identified. The other branch specifies a population of a priori interest and seeks to learn about treatment effects in this population. Here the canonical finding is a bound on average treatment effects.

The various approaches to the analysis of experiments are complementary from a mathematical perspective, but in tension as guides to evaluation practice. The reader of an evaluation reporting that some social program "works" or has "positive impact" should be careful to ascertain what treatment effect has been estimated and under what assumptions.

Learning about Social Programs from Experiments with Random Assignment of Treatments

1. INTRODUCTION

1.1. Credible Evaluation

Program evaluations are efforts to learn from experience in order to improve social decisions. Individuals try to learn from experience in order to improve their private decisions. Yet program evaluation is not the same as private learning from experience. Evaluation and private learning differ in their audiences, and this difference implies important distinctions in empirical practice.

Consider, for example, an individual attempting to learn his private returns to training and a policy analyst seeking to evaluate a training program. The individual need not be concerned about the credibility of his conclusions to anyone other than himself. Hence, he should combine the available empirical evidence about the returns to training with whatever assumptions that he personally is willing to maintain. A policy analyst evaluating a training program must be concerned about the credibility of her findings to a diverse population who may hold varied prior beliefs. She must keep in mind that the weaker are the assumptions imposed in the evaluation, the more widely credible are its findings.

Concern with the credibility of program evaluations has generated considerable interest in the conduct and analysis of experiments, especially ones with random assignment of treatments. Broadly construed, an experiment is any purposeful intervention or natural event that alters the environment in a specified way. Experiments are thought useful because observing outcomes in different environments enriches the empirical evidence available to support evaluations and social science research more generally.

The classical argument for evaluation using experiments with randomly assigned treatments is generally attributed to Ronald Fisher (1935) and could be phrased as follows:

> Let random samples of persons be drawn and formed into treatment groups. Let all members of a treatment group be assigned the same treatment and suppose that each subject complies with the assigned treatment. Then the distribution of outcomes experienced by the members of a treatment group should be the same (up to random sampling error) as would be observed under a program in which the treatment in question is received by all members of the population.

The argument applies equally to *controlled experiments*, in which a researcher purposefully randomizes treatment assignments, and to *natural experiments*, in which randomization is a consequence of some process external to the evaluation. From the perspective of inference, the randomization mechanism is irrelevant.

It is often said that random assignment of treatments minimizes the need for assumptions in evaluations of social programs and, hence, maximizes the credibility of evaluations. A National Research Council report on the evaluation of AIDS prevention programs put it this way: "Well-executed randomized experiments require the fewest assumptions in estimating the effect of an intervention" (Susan Coyle, Robert Boruch, and Charles Turner, 1991, p. 125). A recent New York Times article on the use of random-assignment experiments to evaluate social programs began with this lead: "Use of controls helps put the science into social science" (Peter Passell, 1993, p. C1).

Yet the use of experiments to predict the outcomes of actual social programs certainly does require assumptions. The classical argument assumes that the experimental subjects assigned to a specified treatment are randomly drawn from the relevant population and that these subjects all comply with the assigned treatment. It assumes the social program of interest to be one in which all members of the population would receive the specified treatment. It assumes the experiment is of sufficient duration to emulate this program and to observe the outcomes of interest. And it assumes the absence

of social interactions that may make a full-scale program inherently different from a small-scale experiment.[1]

These assumptions may sometimes be appropriate but very often there is good reason to doubt their realism. In a book that is little known to economists but something of a bible to other social scientists concerned with program evaluation, Donald Campbell and Julian Stanley (1966) discuss various "factors jeopardizing the validity" of experiments. In the economics literature, many of the contributors to the volumes edited by Jerry Hausman and David Wise (1985) and Charles Manski and Irwin Garfinkel (1992) find that the classical argument for random assignment is not credible when applied to recent experimental evaluations of income maintenance, welfare, training, and other social programs.[2]

---

[1]The classical argument assumes that the members of the population are isolated from one another, each person's outcome varying with his own treatment but not with the treatment of others. In particular, the scale of the experiment is assumed not to affect a given subject's outcome. But the scale of an experiment does affect outcomes when persons interact with one another. An example familiar to labor economists is the inappropriateness of using small-scale job training experiments to predict the employment outcomes that would occur in full-scale programs. Unless the demand for labor is perfectly elastic, increasing one unemployed person's human capital alters the employment opportunities of others in his neighborhood. So the outcomes observed when a small fraction of the unemployed population are trained in an experiment are not those that would occur if all unemployed persons were to receive training.

[2]The validity of the classical assumptions was a prominent concern of researchers analyzing data from the income-maintenance experiments of the 1970s. Various analysts sought to use the labor supply and marital outcomes experienced by subjects in the experiments to predict the outcomes that would be experienced in the general population if a tax schedule with negative income tax provisions were to be legislated. Several doubts about the classical assumptions led to caution in drawing inferences from the experiments. One was that persons facing a permanently altered tax schedule might behave differently than did the experimental subjects, who knew that the experiment had limited duration (e.g., Glen Cain, 1986). Another was that persons required by law to compute their taxes using the new tax schedule might behave differently than the experimental subjects, who were offered the new schedule as an alternative to the standard one but were not compelled to use it (e.g., Hausman and Wise, 1979). A third concern was that, due to social interactions, the responses of isolated experimental subjects to the negative income tax might differ from those that would be observed should the program be implemented broadly (see Mordechai Kurz and Robert Spiegelman, 1973).

What can experiments with randomly assigned treatments credibly reveal about social programs when the classical argument for random assignment is not credible? Many authors advocating the use of experiments in evaluation offer no guidance beyond warning that one should be cautious in extrapolating from experiments to programs. Nevertheless, a developing literature is addressing the question in an increasingly systematic way. This article examines the current status of the literature and, in so doing, confronts some basic unsettled issues in program evaluation.

The remainder of this introductory section describes the evolution of thinking about the use of experiments in evaluation and the main themes of the article. These themes are then developed formally in Sections 2 through 5. Section 6 gives conclusions.

1.2. Statutory Randomization

In the United States, concern with the evaluation of social programs has spread rapidly since the 1960s, when attempts were made to evaluate the impacts of programs proposed as part of the War on Poverty.[3] Evaluation requirements now appear in major federal statutes. One of these is the Family Support Act of 1988, which revised the program of Aid to Families with Dependent Children (AFDC). In Title II of this statute, Congress mandated study of the effectiveness of training programs initiated by the states under the new Job Opportunities and Basic Skills Training Program (JOBS). Congress even stipulated the mode of data collection: "a demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program" (Public Law 100-485, October 13, 1988, Section 203, 102 Stat. 2380).

[3]See Henry Aaron (1978, p. 30) and Robert Haveman (1987, Chap. 8). Before the 1960s, evaluation efforts were largely limited to "process" evaluations, describing the administration of programs.

A notable early use of experiments with random assignment of treatments to evaluate anti-poverty programs is the Perry Preschool Project, begun in the early 1960s.[4] Intensive educational and social services were provided to a random sample of about sixty black children aged 3 and 4, living in a low-income neighborhood of Ypsilanti, Michigan. No special services were provided to a second random sample of such children, drawn to serve as a control group. The treatment and control groups have subsequently been followed into adulthood. Among other things, it has been found that 67 percent of the treatment group and 49 percent of the control group were high school graduates by age 19 (see John Berrueta-Clement et al., 1984). This and similar findings for other outcomes have been cited widely as evidence that intensive early childhood educational interventions improve the outcomes of children at risk (see Constance Holden, 1990).

A series of much larger experiments with randomly assigned treatments were carried out by economists in the 1970s. These include four income-maintenance experiments (see David Kershaw and Jerilyn Fair, 1976; Robert Moffitt and Kenneth Kehrer, 1981; and Alicia Munnell, 1986), the RAND national Health Insurance Study (see Wilfred Manning et al., 1987), and the National Supported Work Demonstration (see Manpower Demonstration Research Corporation, 1980). Various experiments of this period are described and critiqued in Hausman and Wise (1985).

The economic experiments of the 1970s were not only much larger than the Perry Preschool Project, but many of them were far more ambitious in their objectives. The Perry Preschool Project, in classical fashion, sought only to describe the life outcomes experienced by the treatment and control groups. In contrast, the economic experiments were designed to provide data that would support estimation of behavioral models--the idea being that such models might be used to evaluate a range of

---

[4]Implemented by educational psychologists, the Perry Preschool Project may be viewed either as an early experimental evaluation of an antipoverty initiative or as part of a long history of experimentation by educational psychologists. Campbell and Stanley (1966, p. 2) report that "a wave of enthusiasm for experimentation dominated the field of education in the Thorndike era, perhaps reaching its apex in the 1920s."

policy alternatives. For example, the income maintenance experiments were designed to provide data for estimation of models of labor supply under alternative income tax schedules. The national health insurance experiment was meant to provide data for estimation of models of the demand for health care under alternative insurance programs.

The experiments of the 1970s undoubtedly were important contributions to economic science, but they do not explain how Congress came to require random assignment of treatments to evaluate the JOBS program in the Family Support Act of 1988. The congressional mandate has two more recent roots.

One was the influence of a set of experiments with welfare reform executed and analyzed by the Manpower Demonstration Research Corporation (MDRC) beginning in the early 1980s (see Judith Gueron and Edward Pauly, 1991). In classical fashion, MDRC sought only to describe the mean outcomes experienced by the various treatment groups. This objective may not have been very ambitious but it enabled MDRC to present its findings in a simple manner that policymakers found compelling.[5] See David Greenberg and Philip Robins (1985) and Greenberg and Michael Wiseman (1992) for critiques of the MDRC research.

The other root was the failure of the nonexperimental evaluations of job training programs performed in the 1970s to produce clear-cut findings (see Burt Barnow, 1987). Frustration with this situation led the U.S. Department of Labor to commission an experimental evaluation of the Job Training Partnership Act in the mid-1980s (see V. Joseph Hotz, 1992). It also led some labor economists to conclude that, as a consequence of the widely recognized *selection problem*, credible evaluation of social programs is not possible unless treatments are randomly assigned (see Lauri Bassi and Orley Ashenfelter, 1986; Robert LaLonde 1986; and Thomas Fraker and Rebecca Maynard, 1987).

---

[5]On the back cover of Gueron and Pauly (1991), Senator Daniel Patrick Moynihan is quoted as follows: "Above all others, Judy Gueron and her colleagues at MDRC did the research that led the Congress to pass the Family Support Act two years ago."

In the context of evaluation, the selection problem arises whenever one observes outcomes under a program in which treatment varies across the population and one wants to predict outcomes under a program in which all members of the population would receive the same treatment. As is well known, the inferences that can logically be made depend critically on what one is willing to assume about the process yielding the observed pattern of treatments and outcomes (see G. S. Maddala, 1983; James Heckman and Richard Robb, 1985; and Manski, 1995, Chap. 2).[6]

By the late 1980s, random assignment of treatments was viewed within the federal government as the preferred approach to data collection for evaluation of welfare and training programs. During 1986 and 1987 the federal Low Income Opportunity Board encouraged experimental evaluation of AFDC reforms proposed by the states (see Michael Fishman and Daniel Weinberg, 1992). In 1988 Congress mandated random assignment of treatments for evaluation of the JOBS program. In 1992, an Assistant Secretary in the U.S. Department of Health and Human Services even criticized the U.S. General Accounting Office for commissioning a nonexperimental evaluation, writing:

"nonexperimental research of training programs has shown such methods to be so unreliable, that Congress and the Administration have both insisted on experimental designs for the Job Training Partnership Act (JTPA) and the Job Opportunities and Basic Skill (JOBS) programs" (letter from Jo Anne B. Barnhart to Eleanor Chelimsky, reproduced as U.S. General Accounting Office (1992), Appendix II).

---

[6]It is common to assume that the treatments persons receive are statistically independent of their outcomes under the alternative treatments. This *exogenous-selection* assumption, which is tantamount to assuming an experiment with randomly assigned treatments, is often thought implausible for two reasons. First, the states or locales that choose to operate a program presumably do so because they expect the program to produce favorable community-wide outcomes. Furthermore, in each location where a program is operated, the people who choose to participate presumably are those who expect the program to have favorable personal outcomes. If expected outcomes are related to actual ones, then the treatments that persons receive are generically dependent on their outcomes.

1.3. Using Experiments to Infer Treatment Effects

Extrapolation from experiments is widely recognized as posing a formidable problem. Aeronautical engineers face this problem when they use wind-tunnel tests to make predictions about the performance of airplanes in flight. Social psychologists face it when they use experiments with college undergraduates to make predictions about general human behavior. Cancer researchers face it when they use high-dosage carcinogen tests performed on laboratory rats to predict the effects of low-dosage exposure on humans. AIDS researchers face it when they use clinical trials with volunteers as subjects to predict the efficacy of new treatments within the population of HIV-positive persons. Researchers evaluating proposed welfare reform programs face it when they use experiments with current AFDC recipients as subjects to predict the outcomes that would be experienced by those who would receive assistance under the proposed reforms.

Recent experimental evaluations of social programs are largely silent on the problem of extrapolating from the experiments performed to the programs of interest. As mentioned earlier, the influential MDRC analyses of welfare reform experiments reported in Gueron and Pauly (1991) only describe the mean outcomes experienced by the various treatment groups. One can use the reported experimental findings to predict program outcomes if one is willing to apply the classical argument for random assignment of treatments. One is at a loss to interpret the findings if one does not think the classical argument credible.

The literature to be examined in this article develops ways to interpret experimental findings when the classical argument is not credible. Our starting point is the idea that, in one way or another, evaluations generally aim to compare the outcomes that the members of some population would experience in alternative programs of interest. In practice, evaluation researchers express this idea by combining the available data with maintained assumptions to infer some form of *treatment effect*; for example, the *average treatment effect*, the *marginal treatment effect*, the *effect of treatment on the*

treated, or the *effect of intent-to-treat* (see Heckman and Robb, 1985; and Anders Björklund and Robert Moffitt, 1987).

Our objective is to determine what may be learned about various treatment effects when data from experiments with randomly assigned treatments are combined with various assumptions. The work to date embraces three more or less distinct approaches:

(a) Select a form of treatment effect deemed of substantive interest. Then impose assumptions strong enough to identify this treatment effect. Examples include Fisher (1935) and Hausman and Wise (1979).

(b) Impose assumptions weak enough to be widely credible. Then determine what forms of treatment effect are identified under these assumptions. Examples include Howard Bloom (1984) and Guido Imbens and Joshua Angrist (1994).

(c) Select a form of treatment effect deemed of substantive interest and assumptions weak enough to be widely credible. Then determine what may be learned about this treatment effect under these assumptions. Examples include Nancy Clements, James Heckman, and Jeff Smith (1994), V. Joseph Hotz and Seth Sanders (1994), and Manski (1993, 1995).

Each approach is logically coherent, and collectively they complement one another. This article focuses on approaches (b) and (c), which aim to achieve widely credible evaluations in different ways. Approach (a), in contrast, trades credibility for stronger conclusions.[7]

To make progress, we need first to formalize the concepts used loosely in this introductory section: treatments, outcomes, programs, treatment effects, experiments, and the classical argument for random assignment. Section 2 presents these essential preliminaries.

---

[7]A parallel literature seeks to determine what may be learned about various treatment effects when nonexperimental data are combined with various assumptions. The work to date embraces the same three distinct approaches. Approach (a) is taken by Heckman and Robb (1985) and Björklund and Moffitt (1987); approach (b) by Imbens and Angrist (1994); and approach (c) by Manski (1994, 1995, Chap. 2).

We then examine what can be learned from experiments in three often-encountered situations where the classical argument fails to hold:

Partial Compliance (Section 3): The treatments actually received by some experimental subjects may differ from the treatments they are assigned. We examine what an experiment with partial compliance reveals about a universal program; that is, one in which everyone actually receives the same treatment. We also discuss the notion of *intention-to-treat.*

Experimentation on a Subpopulation (Section 4): The subjects in an experiment are often drawn from a subpopulation of the relevant population. In particular, the subjects for experiments used to evaluate proposed social programs are often drawn from the participants in some existing program. We examine what experimentation on a subpopulation reveals about outcomes in the population at large.

Treatment Variation (Section 5): The classical argument for random assignment presumes that one wishes to learn about a universal program. In many programs, however, treatment varies across the population. We determine what experiments reveal about the outcomes of such programs.

Other important ways in which the classical argument may fail to hold will not be examined here. Among these, I would call particular attention to the matter of social interactions. It has long been recognized that social interactions invalidate the classical argument (see Jeffrey Harris, 1985; and Garfinkel, Manski, and Charles Michalopolous, 1992). Unfortunately, little progress has been made in determining how experimental data should be interpreted in the presence of social interactions.

2. CONCEPTS OF FORMAL EVALUATION

2.1. Treatments, Outcomes, and Programs

Throughout this article, we shall suppose that each member of the relevant population receives one of several mutually exclusive and exhaustive *treatments*. Each person experiences an *outcome* that may depend on the treatment received. The possible treatments will be numbered t = 1,...,T. The outcomes associated with these treatments will be called y(t), t = 1,...,T.

For example, the relevant population might be unemployed males age 45 to 65. Each member of this population might receive one of four treatments: no public assistance (t = 1); public assistance in job search (t = 2); publicly funded retraining (t = 3); or an offer of publicly funded retraining, allowing a person to choose between treatments 1 and 3 (t = 4). The relevant outcome might be earned income. The outcomes experienced under the four treatments are y(t), t = 1,...,4. We shall assume that y(4) = y(1) when an offer of retraining is declined and y(4) = y(3) when it is accepted, but there is no logical requirement for y(4) to be thus determined.[8]

Social programs influence the treatment each person receives. Let the programs of interest be mutually exclusive and be numbered m = 1,...,M. Let z_m indicate the treatment that a person would receive in program m. Then the outcome a person would experience in program m is $y(z_m)$. Observe that a program is assumed to affect outcomes only through its influence on the treatment received and not through any other channel. That is why we write the outcome in program m as $y(z_m)$ rather than $y_m(z_m)$.

Continuing the example, four programs may be under consideration: no public assistance to unemployed persons (m = 1); retraining for persons age 45 to 54 and assistance in job search for persons over age 55 (m = 2); no assistance for persons with college degrees, but an offer of retraining to persons without college degrees (m = 3); and an offer of retraining to all unemployed persons

[8]Conceivably, a person who chooses to be retrained may be more motivated to learn and, subsequently, earn more than the same person would if he or she were required to undergo retraining.

(m = 4). In the first program, a person experiences outcome y(1). In the second program, a person age 45 to 54 experiences outcome y(3) and one age 55 or over experiences y(2). In the third program, a person with a college degree experiences outcome y(1), and one without a degree experiences either y(1) or y(3). In the fourth program, a person experiences y(1) or y(3).

In the example, m = 1 and m = 4 are *universal programs*, ones in which all members of the population receive the same treatment. In contrast, m = 2 and m = 3 are programs with *treatment variation*, ones in which members of the population vary in the treatments they receive. In each of the four programs, persons who receive a given treatment may vary in the outcomes they experience under that treatment. In program 1, for example, earnings outcomes y(1) would presumably vary across the population, as some unemployed persons would find well-paying jobs, some would find low-paying jobs, and some no job at all.

To express the treatments that different persons receive in the programs of interest, and the outcomes they experience, we could list the members of the population and, for each person i, seek to learn the vector $[y_i(z_{im}), z_{im}; m = 1,...,M]$. But evaluations typically are concerned with the distribution of treatments and outcomes across the population, not with the situations of particular individuals. If individuals are differentiated at all, it is only through their values for some covariates, say x. With this in mind, we shall view $[\{y(z_m), z_m; m = 1,...,M\}, x]$ as a random vector.

In this article, P generically denotes both a probability distribution and the probability of an event. For example, $P(z_m)$ denotes the multinomial distribution of treatments received by members of the population under program m, and $P(z_m = t)$ denotes the probability that a person receives treatment t in this program. (Equivalently, $P(z_m = t)$ is the fraction of the population receiving treatment t.) $P[y(t) \mid z_m = t]$ denotes the distribution of outcomes experienced by those persons receiving treatment t in program m. By the law of total probability, the overall distribution of outcomes in program m is

$$(1)\quad P[y(z_m)] \;=\; \sum_{t=0}^{T} P[y(t) \mid z_m = t]\cdot P(z_m = t).$$

2.2. Comparing Programs

Evaluations seek to compare the outcomes that the population would experience under alternative programs. This objective may seem conceptually straightforward when stated verbally, but formalization reveals two basic unsettled issues in today's literature on evaluation: How should the outcomes of alternative programs be compared? What is the relevant population? We address the first question here and the second in Section 2.3.

Suppose that one wants to compare programs 2 and 1. One way is to compare their outcome distributions $P[y(z_2)]$ and $P[y(z_1)]$. Another is to study the distribution $P[y(z_2) - y(z_1)]$ of the difference between the outcomes of the two programs. More generally, one may compare the joint (outcome, covariate) distributions $P[y(z_2), x]$ and $P[y(z_1), x]$, or study $P[y(z_2) - y(z_1), x]$. To simplify the notation, I omit the covariates in all that follows.

The two approaches are logically distinct. Knowledge of $P[y(z_2) - y(z_1)]$ neither implies nor is implied by knowledge of $P[y(z_2)]$ and $P[y(z_1)]$. The distinction between these ways to compare programs has been blurred for at least two reasons. First, the literature on evaluation has focused almost exclusively on the means of distributions. It is the case that

$$(2)\quad \delta(2,1) \;\equiv\; E[y(z_2) - y(z_1)] \;=\; E[y(z_2)] - E[y(z_1)].$$

Researchers restricting attention to mean values are able to say either that they are comparing $E[y(z_2)]$ and $E[y(z_1)]$ or that they are studying $E[y(z_2) - y(z_1)]$. Either way, the quantity $\delta(2,1)$ is called the *average treatment effect* of program 2 relative to 1.

Second, in settings where outcomes are continuous variables and programs are universal, it has been common to assume that program 2 (i.e., treatment 2) shifts the outcomes of program 1 (i.e., treatment 1) by the same amount for all members of the population. The *shifted-outcomes* or *constant-treatment-effect* assumption is that

(3) $y(2) = y(1) + \alpha$

for some constant α. When (3) is assumed, P[y(2)] and P[y(1)] are the same distribution up to a locational shift of magnitude α, and P[y(2) - y(1)] is a degenerate distribution with all its mass at α. Hence, whichever way one wants to compare programs 2 and 1, the problem is to learn α.

Recent work expands the concern of evaluations beyond the means of distributions and does not reflexively assume that treatment effects are constant across the population. So it has become necessary to distinguish between the two ways to compare programs. Suppose, for example, that one is concerned with the medians of distributions rather than their means. In general,

(4) $\text{Med}[y(z_2) - y(z_1)] \neq \text{Med}[y(z_2)] - \text{Med}[y(z_1)]$.

Researchers referring to the *median treatment effect* of program 2 relative to 1 must be careful to state whether the object of interest is the left or the right side of (3). Clements, Heckman, and Smith (1994) are concerned with $\text{Med}[y(z_2) - y(z_1)]$, while Manski (1993) is concerned with comparison of $\text{Med}[y(z_2)]$ and $\text{Med}[y(z_1)]$.

How should programs 2 and 1 be compared, through $P[y(z_2)]$ and $P[y(z_1)]$, or through $P[y(z_2) - y(z_1)]$? Presuming that the purpose of evaluation is to improve social decisions, we might approach this question from the perspective of a social planner required to choose between the two programs

(e.g., as in Frank Stafford, 1985, pp. 112-114). The standard decision process of welfare economics calls for the planner to adopt a social welfare function $W(\cdot)$, whose argument is the distribution of outcomes in a specified program. The planner should choose program 2 if $W\{P[y(z_2)]\} \geq W\{P[y(z_1)]\}$, and program 1 otherwise. Thus, a planner maximizing a conventional social welfare function wants to learn $P[y(z_2)]$ and $P[y(z_1)]$, not $P[y(z_2) - y(z_1)]$.

Some of the literature on evaluation begins from a different perspective, in which program 1 is the "base" or "default," and program 2 is a proposed alternative. In this setting, it is argued that the object of interest is $P[y(z_2) - y(z_1)]$, which measures the distribution of changes that would be experienced if program 1 were to be replaced by program 2. For example, Clements, Heckman, and Smith (1994) write: "Answers to many interesting evaluation questions require knowledge of the distribution of program gains. From the standpoint of a detached observer of a social program (e.g., a "social planner") who takes the base state values ... as those that would prevail in the absence of a program, it is of interest to know ... the proportion of the total population benefitting from the program" (p. 10). See also Heckman (1992). If y measures the benefits of the treatment received, the proportion of the population that would benefit if program 1 were to be replaced by program 2 is $P[y(z_2) - y(z_1) > 0]$.

2.3. The Relevant Population

We have as yet said nothing about the population with which an evaluation should be concerned. In fact, there is nothing of generality to say if we adopt the perspective of a social planner seeking to maximize a conventional social welfare function. From the perspective of welfare economics, the relevant population is a primitive concept specified prior to the comparison of alternative programs.

Nevertheless, it is common to deviate from the perspective of welfare economics and specify the relevant population within rather than prior to the evaluation. In particular, analysts often begin with a population of welfare-economic interest and then focus attention on a particular subpopulation: those persons who receive a specified treatment t in a specified program m. This done, programs 2 and 1 are compared though the average effect of treatment on the persons for whom $z_m = t$, namely

$$(5)\quad \delta(2,1 \mid z_m = t) \equiv E[y(z_2) - y(z_1) \mid z_m = t]$$
$$= E[y(z_2) \mid z_m = t] - E[y(z_1) \mid z_m = t].$$

$\delta(2,1 \mid z_m = t)$ is called the *average effect of treatment on the treated*, where "the treated" are the persons who receive treatment t in program m.

Some of the many evaluation studies reporting estimates of $\delta(2,1 \mid z_m = t)$ include Bloom (1984), Björklund and Moffitt (1987), Angrist (1990), Gueron and Pauly (1991), Jeffrey Dubin and Douglas Rivers (1993), and Hotz and Sanders (1994). In Gueron and Pauly (1991), for example, program 1 offers AFDC, program 2 offers welfare reform, treatment 1 is the AFDC treatment, treatment 2 is the welfare-reform treatment, and treatment 3 is nonreceipt of welfare. The analysis focuses on $\delta(2,1 \mid z_1 = 1)$, the average effect of welfare reform on current AFDC recipients.

Björklund and Moffitt (1987) stress that, in general, $\delta(2,1 \mid z_m = t)$ does not coincide with the average treatment effect $\delta(2,1)$ computed on the whole population of welfare-economic interest. The algebraic relationship between the two quantities is

$$(6)\quad \delta(2,1) = \delta(2,1 \mid z_m = t)\cdot P(z_m = t) + \delta(2,1 \mid z_m \neq t)\cdot P(z_m \neq t),$$

where $\delta(2,1 \mid z_m \neq t) \equiv E[y(z_2) - y(z_1) \mid z_m \neq t]$ is the *average effect of treatment on the nontreated.* Björklund and Moffitt use nonexperimental data on a Swedish manpower training program to illustrate that $\delta(2,1 \mid z_m = t)$ and $\delta(2,1 \mid z_m \neq t)$ may differ in sign in realistic situations. There are two treatments: training (t = 1) and no training (t = 2). There are three programs of interest: the existing program offering training to certain Swedish workers (m = 3), a hypothetical program expanding training to more such workers (m = 2), and a hypothetical situation in which the government would not offer training to workers (m = 1). Comparing programs 2 and 1, Björklund and Moffitt estimate a positive mean wage gain for those workers who choose to be trained in the existing program, and a negative mean wage gain for those workers who do not accept the existing training offer. That is, they find that $\delta(2,1 \mid z_3 = 1) > 0$ and $\delta(2,1 \mid z_3 \neq 1) < 0$. They assess the potential policy implications as follows:

> "What would be the effect of expanding this program on mean wages (i.e., on productivity growth) in Sweden? Evaluating equation (17) at the means indicates that mean wages would *fall* if the program were expanded by lowering costs (e.g., raising stipends). This is because, at the margin, those who would be brought into the program have negative wage gains according to our estimates" (p. 48).

There are special cases in which $\delta(2,1 \mid z_m = t)$ does coincide with $\delta(2,1)$. Suppose, for example, that programs 1 and 2 are universal, that outcomes are continuous, and that the shifted-outcomes assumption (3) holds. Then $\delta(2,1) = \delta(2,1 \mid z_m = t) = \alpha$ for all values of m and t.

There are also cases in which $\delta(2,1 \mid z_m = t)$ differs from $\delta(2,1)$ only by a factor of scale. Consider again Gueron and Pauly (1991), where program 1 offers AFDC, program 2 offers welfare reform, treatment 1 is the AFDC treatment, treatment 2 is the welfare-reform treatment, and treatment 3 is nonreceipt of welfare. Their analysis assumes that the persons who would receive welfare in the reform program are the same as those who currently receive AFDC; that is, $z_1 = 1 \iff z_2 = 2$, and z_1

$= 3 \iff z_2 = 3$. This assumption of no *entry effects* implies that welfare reform would have no impact on those who do not currently receive AFDC. So $\delta(2,1 \mid z_1 \neq 1) = 0$ and

$$(7) \quad \delta(2,1) = \delta(2,1 \mid z_1 = 1) \cdot P(z_1 = 1).$$

Thus, given knowledge of the current AFDC participation rate $P(z_1 = 1)$ and of the effect of welfare reform on current AFDC recipients, one can determine the effect of welfare reform on the population at large.

Why are evaluation studies so often concerned with the effect of treatment on the treated? In Gueron and Pauly (1991) and similar analyses of welfare reform, I would speculate that the real object of interest is the effect of welfare reform on the population at large, and that the operational focus on $\delta(2,1 \mid z_1 = 1)$ arises only because the assumption of no entry effects makes learning $\delta(2,1 \mid z_1 = 1)$ tantamount to learning $\delta(2,1)$. I say "speculate" because studies evaluating welfare reform typically do not say why they focus on $\delta(2,1 \mid z_1 = 1)$.

It is rare to find an evaluation offering an explicit substantive rationale for interest in the effect of treatment on the treated. One study that does try to motivate the concept is Angrist (1990). Here, program 1 is the Vietnam era military draft, program 2 is a hypothetical situation with no draft, and treatment 1 is military service under the draft. Taking the outcome of interest to be a person's lifetime earnings, Angrist estimates $\delta(2,1 \mid z_1 = 1)$ -- the average effect of mandatory military service on the lifetime earnings of those who were drafted. He motivates interest in this quantity by writing (p. 313): "A central question in the debate over military manpower policy is whether veterans are adequately compensated for their service."

2.4. Inference from Experimental Data

However one decides to compare the outcomes of alternative programs and specify the relevant population, one must use available data to infer the outcome distributions of interest. The central inferential problem confronting all evaluations is the logical impossibility of observing the entire vector $[y(z_m), z_m; m = 1,...,M]$ of treatments that a person might receive and outcomes that a person might experience in alternative programs. In reality, each member of the population receives exactly one treatment in one program. An analyst can at most observe realizations of $[y(z_\mu), z_\mu]$, where μ denotes the actual program faced by the population.

What distinguishes experimental from nonexperimental data? If social programs generically influence the treatments that persons receive, experiments are programs that use *treatment assignments* to influence the treatments received. Let τ_m denote the treatment that a person is assigned in experiment m. We usually think of experiments as assigning treatments only to the sample of persons actually drawn as subjects. It facilitates the analysis, however, if we think of an experiment as assigning a treatment to each member of the population.

Henceforth, we assume that the actual program μ is an experiment. We assume that the analyst draws a random sample of the persons assigned a specified treatment t, and then observes each sample member's realization of $[y(z_\mu), z_\mu]$.[9] These assumptions imply that the analyst can consistently estimate the distribution $P[y(z_\mu), z_\mu \mid \tau_\mu = t]$ of treatments received and outcomes experienced by persons assigned treatment t. Henceforth, we shall abstract from questions of finite-sample inference and simply assume that the analyst knows $P[y(z_\mu), z_\mu \mid \tau_\mu = t]$.

[9]These are standard assumptions in the analysis of experiments, but they are not trivial ones. In some experiments z_μ is not observable; for example, in medical trials where treatments are self-administered by subjects, the analyst observes the treatments subjects are assigned but cannot observe the treatments they actually receive. Furthermore, some experiments have so-called *Hawthorne effects*, wherein outcomes depend not only on the treatments received but also on the treatments assigned.

Now consider the classical assumptions that treatments are randomly assigned and that the persons assigned treatment t all comply with their assigned treatment. A person is said to *comply* with his or her assigned treatment if $z_\mu = \tau_\mu$.[10] Full compliance by those assigned treatment t implies that

(8) $P[y(z_\mu) \mid \tau_\mu = t] = P[y(t) \mid \tau_\mu = t]$.

Random assignment of treatments implies that y(t) is statistically independent of τ_μ; that is,[11]

(9) $P[y(t) \mid \tau_\mu] = P[y(t)]$.

It follows from (8) and (9) that

(10) $P[y(z_\mu) \mid \tau_\mu = t] = P[y(t)]$.

Equation (10) states the classical conclusion that the distribution of outcomes experienced by those persons assigned treatment t is the same as the distribution of outcomes that would be observed if all members of the population were to receive treatment t.

---

[10]In the classical setup, treatments are qualitatively distinct from one another and compliance is a binary variable--a subject does or does not comply with the assigned treatment. In some situations, it is sensible to think of treatments as different values for an ordered variable. In clinical trials, the assigned treatment might be a specified daily dose of a drug and the treatment received might be a different dose. Then the *degree of compliance* measures the closeness of the treatment received to the treatment assigned. Or the assigned treatment might be a specified duration on a weight-loss regime and the treatment received might be the length of time that the subject adheres to the regime. Then the date of attrition from the trial measures the closeness of the treatment received to the treatment assigned.

[11]In traditional econometric language, equation (9) is an *exclusion restriction* and τ_μ is an *instrumental variable*: the treatment τ_μ assigned may affect the treatment z_μ received, but does not affect the outcome y(t) under a specified treatment t.

This completes our development of the concepts of formal evaluation. Our concerns in the remainder of the article can all be seen by reference to equations (8) through (10). Section 3 examines the problem of inference on P[y(t)] when some persons assigned treatment t do not comply, so equation (8) need not hold. Section 4 examines inference on P[y(t)] when the persons assigned treatment t are drawn at random not from the relevant population but from a subpopulation, so equation (9) need not hold. Section 5 examines inference on the distribution P[y(z_m)] of outcomes of a program in which treatment varies across the population, so the classical conclusion (10) does not answer the question of interest.

The analysis in Sections 3 through 5 restricts attention to the means of distributions. Hence, we shall not need to distinguish between the two ways to compare programs discussed in Section 2.2. We shall, however, repeatedly confront the problem discussed in Section 2.3 of specifying the relevant population. It will often be necessary to distinguish between the average treatment effect and the average effect of treatment on the treated.

3. PARTIAL COMPLIANCE

3.1. Mandatory Treatment and Intention-to-Treat

We have said that an experimental subject complies with the assigned treatment if the treatment received coincides with the treatment assigned. This definition of compliance is widely used by evaluation researchers. Nevertheless, different researchers analyzing the same experiment may interpret compliance in different ways.

Consider the Illinois Unemployment Insurance (UI) experiment. Dubin and Rivers (1993) describe the experiment as randomly assigning newly unemployed persons to three treatments: conventional UI (t = 1); conventional UI augmented by a *wage subsidy* paid to the employer if the unemployed person should find a full-time job within eleven weeks (t = 2); and conventional UI

augmented by a *search bonus* paid to the unemployed person if he or she should find a full-time job within eleven weeks (t = 3). They report that 32 percent of those assigned the wage subsidy and 11 percent of those assigned the search bonus did not comply with the assigned treatment, these subjects choosing instead to receive conventional UI. They investigate in some depth the implications of this noncompliance for inference on various treatment effects.

Stephen Woodbury and Robert Spiegelman describe the same experiment as randomly assigning newly unemployed persons to three treatments: conventional UI (t = 1); an offer of a wage subsidy, allowing the unemployed person to choose between conventional UI and the wage-subsidy augmented UI (t = 4); and an offer of a search bonus, allowing the unemployed person to choose between conventional UI and the search-bonus augmented UI (t = 5). They view all subjects as complying with their assigned treatments and they analyze the experimental data in the classical manner.[12]

The differing perspectives of Dubin and Rivers (1993) and Woodbury and Spiegelman (1987) are not contradictory. Consider the wage subsidy. Both studies seek to compare the conventional UI program with an alternative program involving a wage subsidy, but the two studies differ in the alternative programs they pose. In Woodbury and Spiegelman, the alternative is a choice between conventional UI, and conventional UI augmented by the subsidy. The relevant average treatment effect, often called the *effect of intention to treat*, is

[12]Noncompliance is logically possible in the Woodbury-Spiegelman setup. It would occur, for example, if a person assigned the conventional UI treatment would somehow receive the search bonus. In practice, this did not occur.

$$(11)\quad \delta(4,1) \equiv E[y(4)] - E[y(1)]$$
$$= E[y(2) \mid z_4 = 2]\cdot P(z_4 = 2) + E[y(1) \mid z_4 = 1]\cdot P(z_4 = 1) - E[y(1)]$$
$$= \{E[y(2) \mid z_4 = 2] - E[y(1) \mid z_4 = 2]\}\cdot P(z_4 = 2).$$

In Dubin and Rivers, the alternative is conventional UI augmented by a mandatory wage subsidy. Here the relevant average treatment effect is

$$(12)\quad \delta(2,1) \equiv E[y(2)] - E[y(1)].$$

From the perspective of inference on $\delta(4,1)$, subjects fully comply with their assigned treatments and the classical argument for randomization holds. From the perspective inference on $\delta(2,1)$, there is substantial noncompliance and the classical argument does not hold.

3.2. Compliance as Selection

Let us now examine the implications of noncompliance for inference on $\delta(2,1)$. The general problem is that one wants to learn the distribution P[y(t)] of outcomes that would be experienced in a program in which all members of an a priori specified population receive a specified treatment t. One observes the treatments received and outcomes experienced by experimental subjects who are randomly assigned this treatment. Some of these subjects comply with the assigned treatment but others do not. What do the experimental data reveal about P[y(t)]?

Partial compliance generates the same *selection problem* as arises routinely in the analysis of nonexperimental data. To see that partial compliance is a selection problem, we begin with the fact that random assignment of treatments makes y(t) statistically independent of the assigned treatment τ_μ, as stated in equation (9). This holds whether or not subjects comply with their assigned treatments.

By (9) and the law of total probability,

$$(13)\quad P[y(t)] = P[y(t) \mid \tau_\mu = t]$$
$$= P[y(t) \mid z_\mu = t,\ \tau_\mu = t]\cdot P(z_\mu = t \mid \tau_\mu = t)$$
$$+ P[y(t) \mid z_\mu \neq t,\ \tau_\mu = t]\cdot P(z_\mu \neq t \mid \tau_\mu = t).$$

Observation of the treatments received by subjects assigned treatment t reveals the compliance probability $P(z_\mu = t \mid \tau_\mu = t)$ and the noncompliance probability $P(z_\mu \neq t \mid \tau_\mu = t)$. Observation of the outcomes of the subjects who comply with the assigned treatment reveals $P[y(t) \mid z_\mu = t,\ \tau_\mu = t]$. The experimental data reveal nothing about $P[y(t) \mid z_\mu \neq t,\ \tau_\mu = t]$, the distribution of y(t) among the subjects who do not comply. The outcomes y(t) are censored for these subjects. This is precisely the selection problem (see Manski, 1994; 1995, Chapter 2).

The selection problem has been studied in some depth. As is well known, P[y(t)] is identified if one is willing to impose sufficiently strong assumptions. Certainly the most common and longstanding practice in evaluation studies is to assume that the distribution of y(t) observed among those subjects who do comply with treatment t is the same as the distribution of y(t) in the population at large. That is,

$$(14)\quad P[y(t)] = P[y(t) \mid z_\mu = t,\ \tau_\mu = t].$$

This *exogenous compliance* assumption identifies P[y(t)] because the right side of (14) is revealed by the experiment.

Another common practice, at least among economists, is to view compliance as a decision made by each subject acting in his or her self-interest. Assuming that subjects care about the outcomes they experience, researchers connect P[y(t)] to the observed distribution of $[y(z_\mu), z_\mu, \tau_\mu]$ through models jointly explaining compliance and outcomes. For example, Hausman and Wise (1979)

use such a model to predict the earnings outcomes that would occur if a tax schedule offered in the Gary Income Maintenance experiment were to be made mandatory. Philipson (1995) suggests that comparison of the compliance decisions made by subjects assigned different treatments may be used to infer treatment effects as perceived by those subjects. He writes:

> "Subjects reveal the economic value of a treatment by their desire to consume it, which in turn is revealed by their decisions regarding continued participation in the experiment with their assigned treatment. Differences in subjects' willingness to stay on their assigned treatments should therefore reveal differences in subjects' economic valuations of the treatments." (p. 3)

Exogenous compliance and decision models connecting compliance and outcomes trade credibility for strong conclusions. Given that credibility is a central concern in evaluations, it makes sense to ask what can be learned about P[y(t)] using the experimental evidence alone.

The key result is this: Let f[y(t)] be any real-valued function of the outcome y(t). Use (9) and the law of iterated expectations to write

$$\begin{aligned}(15)\quad E\{f[y(t)]\} &= E\{f[y(t)] \mid \tau_\mu = t\} \\ &= E\{f[y(t)] \mid z_\mu = t,\ \tau_\mu = t\} \cdot P(z_\mu = t \mid \tau_\mu = t) \\ &\quad + E\{f[y(t)] \mid z_\mu \neq t,\ \tau_\mu = t\} \cdot P(z_\mu \neq t \mid \tau_\mu = t).\end{aligned}$$

The experiment reveals all the quantities on the right side except for the censored mean $E\{f[y(t)] \mid z_\mu \neq t,\ \tau_\mu = t\}$. This censored mean can take any value in the interval $[K_0, K_1]$, where K_0 and K_1 are the lower and upper endpoints of the logical range of f[y(t)]. Hence $E\{f[y(t)]\}$ must lie within this bound:

$$(16)\quad E\{f[y(t)] \mid z_\mu = t,\ \tau_\mu = t\} \cdot P(z_\mu = t \mid \tau_\mu = t) + K_0 \cdot P(z_\mu \neq t \mid \tau_\mu = t)$$

$$\leq E\{f[y(t)]\}$$

$$\leq E\{f[y(t)] \mid z_\mu = t,\ \tau_\mu = t\} \cdot P(z_\mu = t \mid \tau_\mu = t) + K_1 \cdot P(z_\mu \neq t \mid \tau_\mu = t).$$

The bound, whose width is $(K_1 - K_0) \cdot P(z_\mu \neq t \mid \tau_\mu = t)$, expresses all that can be learned about $E\{f[y(t)]\}$ given the experimental data alone (see Manski, 1994).[13]

We may use (16) to bound the probability $P[y(t) \in B]$ that the outcome y(t) lies in any set B. This probability can be interpreted as the expected value of the indicator function $1[y(t) \in B]$, whose logical lower and upper bounds are $K_0 = 0$ and $K_1 = 1$. Thus, (16) implies that

$$(17)\quad P[y(t) \in B \mid z_\mu = t,\ \tau_\mu = t] \cdot P(z_\mu = t \mid \tau_\mu = t) \leq P[y(t) \in B]$$

$$\leq P[y(t) \in B \mid z_\mu = t,\ \tau_\mu = t] \cdot P(z_\mu = t \mid \tau_\mu = t) + P(z_\mu \neq t \mid \tau_\mu = t).$$

For example, we can bound the distribution function of y(t). For each outcome value Y, (17) implies that

---

[13]This finding assumes that the only data available are observations on the subjects assigned treatment t. Data on other treatment groups may yield information about $E\{f[y(t)]\}$ if some members of these groups do not comply with their assigned treatments and receive treatment t instead. Let s denote any treatment and let S be a specified set of treatments. Suppose that, for each $s \in S$, one observes the treatments received and outcomes experienced by the subjects assigned treatment s. The same reasoning that yielded (16) now shows that $E\{f[y(t)]\}$ must lie within this intersection of bounds (see Manski, 1994, Proposition 6):

$$\sup_{s \in S} E\{f[y(t)] \mid z_\mu = t,\ \tau_\mu = s\} \cdot P(z_\mu = t \mid \tau_\mu = s) + K_0 \cdot P(z_\mu \neq t \mid \tau_\mu = s) \leq E\{f[y(t)]\}$$

$$\leq \inf_{s \in S} E\{f[y(t)] \mid z_\mu = t,\ \tau_\mu = s\} \cdot P(z_\mu = t \mid \tau_\mu = s) + K_1 \cdot P(z_\mu \neq t \mid \tau_\mu = s).$$

(18) $P[y(t) \leq Y \mid z_\mu = t, \tau_\mu = t] \cdot P(z_\mu = t \mid \tau_\mu = t) \leq P[y(t) \leq Y]$

$\leq P[y(t) \leq Y \mid z_\mu = t, \tau_\mu = t] \cdot P(z_\mu = t \mid \tau_\mu = t) + P(z_\mu \neq t \mid \tau_\mu = t).$

We may also use (16) to bound the average treatment effect $\delta(2,1)$ given in (12). Let $[K_0, K_1]$ be the logical range of the outcome variable y. The lower bound on $\delta(2,1)$ is the lower bound on E[y(2)] minus the upper bound on E[y(1)]. Similarly, the upper bound on $\delta(2,1)$ is the upper bound on E[y(2)] minus the lower bound on E[y(1)]. Thus

(19) $E[y(2) \mid z_\mu = 2, \tau_\mu = 2] \cdot P(z_\mu = 2 \mid \tau_\mu = 2) + K_0 \cdot P(z_\mu \neq 2 \mid \tau_\mu = 2)$
$- E[y(1) \mid z_\mu = 1, \tau_\mu = 1] \cdot P(z_\mu = 1 \mid \tau_\mu = 1) - K_1 \cdot P(z_\mu \neq 1 \mid \tau_\mu = 1)$

$\leq \delta(2,1)$

$\leq E[y(2) \mid z_\mu = 2, \tau_\mu = 2] \cdot P(z_\mu = 2 \mid \tau_\mu = 2) + K_1 \cdot P(z_\mu \neq 2 \mid \tau_\mu = 2)$

$- E[y(1) \mid z_\mu = 1, \tau_\mu = 1] \cdot P(z_\mu = 1 \mid \tau_\mu = 1) - K_0 \cdot P(z_\mu \neq 1 \mid \tau_\mu = 1).$

This bound expresses all that can be learned about $\delta(2,1)$ given the experimental data alone.

To illustrate these findings, consider the Illinois UI experiment. Let y(t) indicate whether a newly unemployed person receiving treatment t is reemployed within 11 weeks of filing for unemployment insurance. Suppose that one wishes to learn P[y(2) = 1], the probability that a newly

unemployed person would be reemployed within 11 weeks in a program of UI augmented by a mandatory wage subsidy.

Recall that 32 percent of those assigned the wage subsidy did not comply. Thus, abstracting from random sampling error, $P(z_\mu = 2 \mid \tau_\mu = 2) = .68$ and $P(z_\mu \neq 2 \mid \tau_\mu = 2) = .32$. Dubin and Rivers (1993) report that, among the subjects who did comply, 38 percent were re-employed within 11 weeks. Hence $P[y(2) = 1 \mid z_\mu = 2, \tau_\mu = 2] = .38$. Applying the bound (17), we may therefore conclude that $.26 \leq P[y(2) = 1] \leq .58$. The lower (upper) bound is the probability of re-employment under the extreme assumption that persons not complying with the wage-subsidy treatment always take more (less) than 11 weeks to find a job.

Going on, consider the problem of inference on the average treatment effect $\delta(2,1) = P[y(2) = 1] - P[y(1) = 1]$. All experimental subjects assigned to treatment 1 complied with their assignment, so the experimental data identify $P[y(1) = 1]$. Dubin and Rivers report that $P[y(1) = 1] = .35$. Hence we may conclude that $-.09 \leq \delta(2,1) \leq .23$.

These conclusions may seem modest. One might be tempted to use the midpoint of the interval [.26,.58] as a point estimate for $P[y(2) = 1]$ and then conclude that $\delta(2,1) = .42 - .35 = .07$. Given the experimental evidence alone, however, there is no justification for doing so.

The usefulness of the bound [-.09, .23] on $\delta(2,1)$ is that it establishes a domain of consensus about the average effect on newly unemployed persons of a UI program mandating wage subsidies.[14] One may draw stronger conclusions by combining the experimental evidence with assumptions about the determination of compliance and outcomes (see Heckman and Robb, 1985; and Manski, 1995, Chap. 2). For example, assuming compliance to be exogenous implies that $\delta(2,1) = .03$. The price of a stronger conclusion, however, is diminished credibility.

---

[14]Not everyone agrees that bounds can be useful. Several years ago I was informed by a prominent econometric consultant that "You can't give the client a bound. The client needs a number."

3.3. The Effect of Treatment on Compliers

Researchers analyzing experiments with partial compliance sometimes focus attention not on the whole population but rather on those persons who comply with their assigned treatment. Let two treatments be labeled t = 1 and t = 2, and consider those experimental subjects who are assigned treatment 2 and comply with this assignment. The *effect of treatment on compliers*, a version of the effect of treatment on the treated, is

$$(20)\quad \delta(2,1 \mid z_\mu = 2,\ \tau_\mu = 2) \equiv E[y(2) \mid z_\mu = 2,\ \tau_\mu = 2] - E[y(1) \mid z_\mu = 2,\ \tau_\mu = 2].$$

The problem of inference on $\delta(2,1 \mid z_\mu = 2,\ \tau_\mu = 2)$ seems to have been first studied by Bloom (1984). Inference on $E[y(2) \mid z_\mu = 2,\ \tau_\mu = 2]$ is not problematic because y(2) is observed for subjects who are assigned treatment 2 and comply. Inference on $E[y(1) \mid z_\mu = 2,\ \tau_\mu = 2]$ is problematic because y(1) is not observed for these subjects. Bloom found that $E[y(1) \mid z_\mu = 2,\ \tau_\mu = 2]$ is identified if treatments 1 and 2 are the only feasible treatments and if all subjects assigned treatment 1 comply with their assignment. In this case, the experimental data alone suffice to identify the effect of treatment on compliers. Nothing need be assumed about the determination of compliance and outcomes.

Here is the result: If treatments 1 and 2 are the only feasible treatments, we may use the law of iterated expectations to write

$$\begin{aligned}(21)\quad E[y(1) \mid \tau_\mu = 2] \;=\; & E[y(1) \mid z_\mu = 1,\ \tau_\mu = 2]\cdot P(z_\mu = 1 \mid \tau_\mu = 2) \\ & + E[y(1) \mid z_\mu = 2,\ \tau_\mu = 2]\cdot P(z_\mu = 2 \mid \tau_\mu = 2).\end{aligned}$$

Solving for $E[y(1) \mid z_\mu = 2,\ \tau_\mu = 2]$ yields

(22) $E[y(1) \mid z_\mu = 2, \tau_\mu = 2] = E[y(1) \mid \tau_\mu = 2]/P(z_\mu = 2 \mid \tau_\mu = 2)$

$- E[y(1) \mid z_\mu = 1, \tau_\mu = 2] \cdot P(z_\mu = 1 \mid \tau_\mu = 2)/P(z_\mu = 2 \mid \tau_\mu = 2).$

Inspect the right side of (22). The experiment reveals $E[y(1) \mid z_\mu = 1, \tau_\mu = 2]$, $P(z_\mu = 1 \mid \tau_\mu = 2)$, and $P(z_\mu = 2 \mid \tau_\mu = 2)$. Randomization of treatments implies that

(23) $E[y(1) \mid \tau_\mu = 2] = E[y(1) \mid \tau_\mu = 1].$

If all subjects assigned treatment 1 comply with their assignment, the experiment reveals $E[y(1) \mid \tau_\mu = 1]$, and so $E[y(1) \mid \tau_\mu = 2]$ is identified. Thus $E[y(1) \mid z_\mu = 2, \tau_\mu = 2]$ is identified.

To compute the effect of treatment on compliers, one may directly apply (20), (22) and (23), thereby obtaining

(24) $\delta(2,1 \mid z_\mu = 2, \tau_\mu = 2) \equiv E[y(2) \mid z_\mu = 2, \tau_\mu = 2]$

$- E[y(1) \mid \tau_\mu = 1]/P(z_\mu = 2 \mid \tau_\mu = 2)$

$+ E[y(1) \mid z_\mu = 1, \tau_\mu = 2] \cdot P(z_\mu = 1 \mid \tau_\mu = 2)/P(z_\mu = 2 \mid \tau_\mu = 2).$

An alternative form is[15]

---

[15]By the law of iterated expectations,

$$E[y(z_\mu) \mid \tau_\mu = 2] = E[y(2) \mid z_\mu = 2, \tau_\mu = 2] \cdot P(z_\mu = 2 \mid \tau_\mu = 2)$$
$$+ E[y(1) \mid z_\mu = 1, \tau_\mu = 2] \cdot P(z_\mu = 1 \mid \tau_\mu = 2).$$

If all persons assigned to treatment 1 comply with their assignment,

$$E[y(z_\mu) \mid \tau_\mu = 1] = E[y(1) \mid \tau_\mu = 1].$$

(25) $\delta(2,1 \mid z_\mu = 2, \tau_\mu = 2)$

$$= \{E[y(z_\mu) \mid \tau_\mu = 2] - E[y(z_\mu) \mid \tau_\mu = 1]\}/P(z_\mu = 2 \mid \tau_\mu = 2).$$

The quantity $E[y(z_\mu) \mid \tau_\mu = 2] - E[y(z_\mu) \mid \tau_\mu = 1]$ is the effect of intention to treat. So the effect of treatment on compliers turns out to be a rescaled version of the effect of intention to treat.

Whereas Bloom restricted attention to the case in which treatments 1 and 2 are the only feasible treatments and all subjects assigned treatment 1 comply with their assignment, Hotz and Sanders (1994) examine the general problem of inference on the effect of treatment on compliers. They show that, given the experimental data alone, the case considered by Bloom is essentially the only one in which $\delta(2,1 \mid z_\mu = 2, \tau_\mu = 2)$ is identified. In general, it is at most possible to bound this treatment effect.

Imbens and Angrist (1994) introduce a treatment effect that is identified when treatments 1 and 2 are the only feasible treatments, whether or not all subjects assigned treatment 1 comply with their assignment. They define the *local average treatment effect* to be (p. 467): "the average treatment effect for individuals whose treatment status is influenced by changing an exogenous regressor that satisfies an exclusion restriction." In an experiment with randomized assignment of treatments, the "exogenous regressor that satisfies an exclusion restriction" is the treatment assignment τ_μ. Hence, the subpopulation on which Imbens and Angrist compare programs consists of those persons for whom the treatment received varies with the treatment assigned.

To formalize this, imagine a hypothetical experiment $\mu-$ that reverses the treatment assignments of the actual experiment μ. Imbens and Angrist consider the subpopulation of persons for whom $z_\mu \neq z_{\mu-}$. In principle, this subpopulation consists of (i) persons who would comply with

---

Hence the right sides of (24) and (25) coincide.

whatever treatment they are assigned; and (ii) persons who would refuse to comply with whatever treatment they are assigned. Let us refer to the former as *compliant persons* and to the latter as *perverse persons*. Imbens and Angrist show that $\delta(2,1 \mid z_{\mu} \neq z_{\mu-})$ is identified under the assumption that the subpopulation for whom $z_{\mu} \neq z_{\mu-}$ contains only compliant persons.

Imbens and Angrist (1994) do not argue that an analyst should, from a welfare economic perspective, be specifically concerned with the subpopulation of compliant persons. To motivate interest in the local average treatment effect, they stress that this treatment effect is identified in circumstances where the effect of treatment on compliers is not identified. The local average treatment effect coincides with the effect of treatment on compliers in the case considered by Bloom, where all persons assigned treatment 1 comply with the assignment. In that case, the compliant persons are those who comply when assigned treatment 2.

4. STRATIFICATION: EXPERIMENTATION ON A SUBPOPULATION

The classical analysis of experiments assumes that subjects are drawn at random from the whole relevant population, but it is common in practice for subjects to be drawn at random from a subpopulation. Experimental psychologists seeking to learn general principles of human behavior use college undergraduates as subjects. Medical researchers wanting to compare alternative treatments advertise for volunteers for clinical trials and draw subjects from those who respond. Subjects in the Illinois UI experiment were drawn from the subpopulation of persons entering the conventional UI program. Subjects in the experimental evaluation of JTPA were drawn from applicants to the program. The standard protocol in the welfare experiments performed since 1980 has been to draw subjects from the subpopulation presently receiving AFDC or, perhaps, from those entering a spell on AFDC. As quoted in Section 1.2, Congress required that subjects in the experimental evaluation of the JOBS program be drawn from current AFDC recipients.

Experimentation on a subpopulation is not problematic if one wants to learn treatment effects within this subpopulation. Assuming full compliance with assigned treatments, clinical trials performed on volunteers reveal the effects of treatments on those who volunteer. Experiments with welfare reform reveal the effects of reform on current AFDC recipients. More generally, stratification on the treated reveals the effect of treatment on the treated.

Experimentation on a subpopulation obviously is problematic if one wants to learn treatment effects in the whole relevant population -- the experiment reveals nothing about outcomes outside of the subpopulation. Let A denote the subpopulation from which experimental subjects are drawn, let $\bar{A}$ denote the complement of A, and let m denote a program of interest. Use the law of total probability to write

$$(26)\quad P[y(z_m)] = P[y(z_m) \mid A]\cdot P(A) + P[y(z_m) \mid \bar{A}]\cdot P(\bar{A}).$$

The experiment may reveal $P[y(z_m) \mid A]$ but reveals nothing about $P[y(z_m) \mid \bar{A}]$. Moreover, the experiment reveals nothing about the size P(A) of subpopulation A. Given the experimental evidence alone, one cannot dismiss the possibility that P(A) is close to 0 and, as a consequence, that $P[y(z_m)]$ is close to $P[y(z_m) \mid \bar{A}]$. Hence the experimental data alone reveal nothing about $P[y(z_m)]$.

4.1. Stratification and Compliance as Selection

The conclusion just reached is overly pessimistic. One sometimes does know the size P(A) of subpopulation A, or one at least can estimate P(A). For example, medical researchers concerned with the population of HIV-positive persons may be able to estimate the fraction of such persons who volunteer to participate in AIDS-treatment trials. Analysts evaluating the effect of welfare reform on the children of unmarried mothers may know the fraction of such children currently receiving AFDC.

So let us examine what can be learned when the experimental evidence is combined with knowledge of P(A).

As in Section 3.2, suppose that in the program of interest, all members of an a priori specified population would receive a specified treatment t. We again focus on the problem of learning the mean E{f[y(t)]} of a function f[y(t)]. Whereas partial compliance was the only problem faced earlier, we now face the additional problem that the experiment is performed on a subpopulation.[16]

Use the law of iterated expectations to write

$$(27)\quad E\{f[y(t)]\} = E\{f[y(t)] \mid A\}\cdot P(A) + E\{f[y(t)] \mid \bar{A}\}\cdot P(\bar{A}).$$

The experiment on subpopulation A reveals nothing about $E\{f[y(t)] \mid \bar{A}\}$, so this quantity can take any value in the interval $[K_0, K_1]$, where K_0 and K_1 are the lower and upper endpoints of the logical range of f[y(t)]. Applying (16) to subpopulation A, the experiment reveals $E\{f[y(t)] \mid A\}$ to lie within the bound given in (28):

$$(28)\quad E\{f[y(t)] \mid A, z_\mu = t, \tau_\mu = t\}\cdot P(z_\mu = t \mid A, \tau_\mu = t) + K_0\cdot P(z_\mu \neq t \mid A, \tau_\mu = t)$$

$$\leq E\{f[y(t)] \mid A\}$$

$$\leq E\{f[y(t)] \mid A, z_\mu = t, \tau_\mu = t\}\cdot P(z_\mu = t \mid A, \tau_\mu = t) + K_1\cdot P(z_\mu \neq t \mid A, \tau_\mu = t).$$

---

[16]Partial compliance and experimentation on a subpopulation are distinct problems but are sometimes confused with one another. Consider clinical trials performed on volunteers. The experiment is performed on a subpopulation because some members of the population do not volunteer to be subjects. Noncompliance occurs when persons who do volunteer are assigned treatments and then do not comply with their assignments. Thus, refusals to volunteer take place before treatments have been assigned and noncompliance occurs after treatments have been assigned. See Dubin and Rivers (1993) and Maddala (1983, p. 266) for further discussion.

It follows that E{f[y(t)]} must lie within this bound:

$$(29)\quad E\{f[y(t)] \mid A,\ z_\mu = t,\ \tau_\mu = t\}\cdot P(z_\mu = t \mid A,\ \tau_\mu = t)\cdot P(A) + K_0\cdot[P(z_\mu \neq t \mid A,\ \tau_\mu = t)\cdot P(A) + P(\bar{A})]$$

$$\leq\ E\{f[y(t)]\}$$

$$\leq\ E\{f[y(t)] \mid A,\ z_\mu = t,\ \tau_\mu = t\}\cdot P(z_\mu = t \mid A,\ \tau_\mu = t)\cdot P(A) + K_1\cdot[P(z_\mu \neq t \mid A,\ \tau_\mu = t)\cdot P(A) + P(\bar{A})].$$

The bound (29) has the same form as (16), except that the selection probability is $P(z_\mu = t \mid A, \tau_\mu = t)\cdot P(A)$ rather than $P(z_\mu = t \mid \tau_\mu = t)$. Thus, we find that the inferential problems created by partial compliance and by experimentation on a subpopulation compound one another multiplicatively.

4.2. Exogenous Stratification and the Absence of Entry Effects

$P[y(z_m)]$ is identified if one is willing to impose sufficiently strong assumptions. Analyses of clinical trials often assume that the distribution of outcomes in the subpopulation of persons who volunteer to participate in a trial is the same as the distribution of outcomes in the larger population of persons for whom the treatment is designed. That is,

$$(30)\quad P[y(z_m)] = P[y(z_m) \mid A].$$

Under this *exogenous stratification* assumption, experimentation on subpopulation A reveals as much about $P[y(z_m)]$ as would experimentation on the whole population.

Analyses of welfare-reform experiments often assume the absence of entry effects. Recall our discussion of Gueron and Pauly (1991) in Section 2.3. Program 1 is AFDC and program 2 is welfare reform. Treatment 1 is the AFDC treatment, treatment 2 is the welfare-reform treatment, and treatment 3 is nonreceipt of welfare. Experimental subjects are drawn from $A = \{z_1 = 1\}$; that is, from the subpopulation of current AFDC recipients. Assuming full compliance with assigned treatments, the experiment reveals $\delta(2,1 \mid z_1 = 1)$, the average effect of welfare reform on current AFDC recipients. But the experiment does not reveal the average effect of welfare reform on the whole population, namely

$$(31) \quad \delta(2,1) = \delta(2,1 \mid z_1 = 1)\cdot P(z_1 = 1) + \delta(2,1 \mid z_1 = 3)\cdot P(z_1 = 3).$$

Assuming no entry effects implies that $\delta(2,1 \mid z_1 = 3) = 0$. This assumption, the experimental data, and knowledge of the current AFDC participation rate $P(z_1 = 1)$ do reveal $\delta(2,1)$.

Although exogenous stratification and the absence of entry effects are often assumed, these assumptions are also often difficult to justify. Consider the assumption of exogenous stratification in clinical trials. The persons who volunteer for a clinical trial presumably are those who expect to benefit from doing so. If expected outcomes are related to actual ones, it is not credible to assume that the outcomes of a trial on volunteers are the same as would be observed if the treatment under study were applied more widely.

Consider the assumption of no entry effects in AFDC reform. The persons currently receiving AFDC presumably are those who expect to benefit from AFDC. The persons who would be recipients of welfare in a reform program presumably would be those who expect to benefit from the reform program. If welfare reform constitutes a meaningful departure from AFDC, it is not credible to

assume that the same persons would participate in both programs. See Heckman (1992) and Moffitt (1992) for extended discussions and for suggestions on modeling the entry process.

Similarly, the persons currently entering the Illinois UI program presumably are those who expect to benefit from UI. Augmentation of conventional UI with a wage subsidy or search bonus may induce additional persons to enter the UI program. Bruce Meyer (1995) puts it this way: "By increasing the financial reward for short UI spells, a permanent bonus would probably increase the number of people unemployed between job changes and increase the number of UI filers." Judging the possible magnitude of this entry effect, he concludes that "the permanent adoption of a reemployment bonus could have important unintended negative effects. The key drawback of the experiments is that they cannot account for the effect of a reemployment bonus on the size of the claimant population.

Thus, experimentation on a subpopulation confronts the researcher with stark alternatives. The experimental data alone reveal nothing about outcomes in the whole population. The experimental data combined with knowledge of the size of the subpopulation implies bounds on mean outcomes, but the bounds are wide if either the subpopulation is small or if noncompliance is widespread. The experimental data combined with knowledge of the size of the subpopulation and with assumptions about the entry process (e.g., no entry effects) implies stronger conclusions, at the expense of the credibility of the evaluation.

5. TREATMENT VARIATION

5.1. The Mixing Problem

The classical analysis of experiments assumes that one wants to predict the outcomes of universal programs, ones in which all members of the population receive the same treatment. The standard definition of a treatment group calls for all the subjects in this group to be assigned the same

treatment. The notion of a treatment may be construed broadly -- a treatment may offer each person a choice among several options. Nevertheless, all the subjects in a treatment group are offered the same choice. In the intention-to-treat interpretation of the Illinois UI experiment, for example, all subjects assigned treatment t = 4 were allowed to choose between conventional UI and the wage-subsidy-augmented UI.

Universal programs are of interest, but so are ones in which treatment varies across the population. Treatment variation occurs when resource constraints prevent universal implementation of desirable treatments. For example, health insurance may be wanted by everyone in the United States but there presently is no program of universal coverage. Treatment variation also occurs when society mandates a treatment but is unable to compel compliance by persons not wanting to be treated. For example, some children do not comply with compulsory schooling requirements and some adults do not pay their taxes.

Let t = 1 and t = 2 be two treatments. Suppose that one draws subjects at random from the whole population and takes pains to ensure that all subjects comply with their assigned treatments, so the experiment reveals P[y(1)] and P[y(2)]. In this section we examine what the experiment reveals about the distribution $P[y(z_m)]$ of outcomes in a program with treatment variation, one in which some members of the population receive treatment 1 and the rest receive treatment 2.

The present question reverses the one addressed in Section 3, where there was partial compliance in the experiment but full compliance in the program of interest. Here there is full compliance in the experiment but partial compliance in the program of interest. Formally, $y(z_m)$ is a probability mixture of y(0) and y(1): $y(z_m) = y(1)$ with probability $P(z_m = 1)$ and $y(z_m) = y(2)$ with

probability $P(z_m = 2)$. The problem of inference on $P[y(z_m)]$ given knowledge of $P[y(1)]$ and $P[y(2)]$ is called the *mixing problem* (Manski, 1995, Chap. 3).[17]

5.2. Inference Using the Experimental Data Alone

I shall develop the main features of the mixing problem through an extended illustration, drawn from Manski (1995, Chap. 3). Recall the Perry Preschool Project discussed in Section 1. Let $t = 2$ be the educational and social services provided to children participating in the project and let $t = 1$ be whatever services as were available to children in the control group. Let y be a binary variable indicating high school graduation by age 19.

Ignoring attrition and sampling error, the experimental data reveal that the high school graduation probability would be .67 if all children in the relevant population were to receive the services provided by the Perry Preschool Project and would be .49 if none of them were to receive these services. Thus $P[y(2) = 1] = .67$ and $P[y(1) = 1] = .49$. The question is this: What do the experimental data reveal about the probability $P[y(z_m) = 1]$ of high school graduation under a program in which some children receive the Perry Preschool services and the rest do not?

It might be conjectured that, if some children were to receive the Perry Preschool services and the rest not, the high school graduation probability would necessarily lie between those observed in the Perry Preschool control and treatment groups, namely .49 and .67. This conjecture is correct under some assumptions but not in general. The experimental data reveal only that $P[y(z_m) = 1]$ must lie between .16 and 1.

To understand this result, observe that each member of the population has one of these four values for [y(2), y(1)]:

---

[17]The mixing problem should not be confused with the converse problem: What does knowledge of $P[y(z_m)]$ imply about $P[y(1)]$, $P[y(2)]$, and $P(z_m)$? The latter is commonly called the *mixture problem*.

[y(2) = 0, y(1) = 0] [y(2) = 1, y(1) = 1]

[y(2) = 1, y(1) = 0] [y(2) = 0, y(1) = 1].

Program m has no impact on persons for whom y(2) = y(1) but determines the outcomes of persons for whom y(2) ≠ y(1). The highest feasible graduation rate is attained by a hypothetical program that always selects the treatment with the better graduation outcome, and so gives treatment 2 to each person with [y(2) = 1, y(1) = 0] and treatment 1 to each person with [y(2) = 0, y(1) = 1]. Then the only persons who do not graduate are those with [y(2) = 0, y(1) = 0], so the graduation rate is 1 - P[y(2) = 0, y(1) = 0]. Symmetrically, the lowest feasible graduation rate is attained by a hypothetical program that, by design or error, gives treatment 1 to each person with [y(2) = 1, y(1) = 0] and treatment 2 to each person with [y(2) = 0, y(1) = 1]. Then the only persons who graduate are those with [y(2) = 1, y(1) = 1], so the graduation rate is P[y(2) = 1, y(1) = 1].

The experiment cannot reveal the joint probabilities P[y(2) = 0, y(1) = 0] and P[y(2) = 1, y(1) = 1]. Treatments 2 and 1 are mutually exclusive, so we never observe both y(2) and y(1) for the same subject. The experiment does reveal the marginal probabilities P[y(2) = 1] = .67 and P[y(1) = 1] = .49. Among all joint distributions P[y(2), y(1)] that are consistent with these marginals, there is one that minimizes both P[y(2) = 0, y(1) = 0] and P[y(2) = 1, y(1) = 1]. This is

P[y(2) = 0, y(1) = 0] = 0 P[y(2) = 0, y(1) = 1] = .33

P[y(2) = 1, y(1) = 0] = .51 P[y(2) = 1, y(1) = 1] = .16.

Hence, the highest graduation rate consistent with the experimental evidence is 1 and the lowest is .16.[18]

---

[18]The general result (see Manski, 1995, Chap. 3) is this: Define C ≡ P[y(2) = 1] + P[y(1) = 1]. Then

$$\max(0, C - 1) \leq P[y(z_m) = 1] \leq \min(C, 1).$$

Observe that the bound is informative from the left or the right but not from both sides simultaneously. The width of the bound narrows toward 0 as C approaches 0 or 2 but widens toward

5.3. Inference under Various Assumptions

The experimental evidence from the Perry Preschool Project reveals little about the probability of high school graduation under a program in which some children receive the Perry Preschool services and the rest do not. We can conclude only that $P[y(z_m) = 1]$ falls in the interval [.16, 1]. Stronger conclusions may be drawn if the experimental evidence is combined with assumptions that reveal the joint distribution P[y(2), y(1)] or the way treatments are selected under program m. Here are some examples.

Given assumptions that identify P[y(2), y(1)], one can tighten the bound on $P[y(z_m) = 1]$ to {P[y(2) = 1, y(1) = 1], 1 - P[y(2) = 0, y(1) = 0]}. Suppose one assumes that receiving the preschool intervention can never harm a child's high school graduation prospects; that is, there are no persons with [y(2) = 0, y(1) = 1]. Then

$$P[y(2) = 1, y(1) = 1] = P[y(1) = 1] = .49$$

$$P[y(2) = 0, y(1) = 0] = P[y(2) = 0] = .33.$$

Hence $P[y(z_m) = 1]$ must lie in the interval [.49, .67]. Or suppose one assumes that y(2) and y(1) are statistically independent. Then

$$P[y(2) = 1, y(1) = 1] = P[y(2) = 1]\cdot P[y(1) = 1] = .33$$

$$P[y(2) = 0, y(1) = 0] = P[y(2) = 0]\cdot P[y(1) = 0] = .17.$$

Hence $P[y(z_m) = 1]$ must now lie in the interval [.33, .83].

One might know the way treatments are selected under program m. Perhaps one optimistically believes that treatment decisions would be made by omniscient parents who choose for each child the treatment yielding the better outcome. Then, as shown earlier, the graduation rate would be 1 - P[y(2) = 0, y(1) = 0]. We already know that zero is the smallest value of P[y(2) = 0, y(1) = 0] that is

---

1 as C approaches 1. Thus, the experiment may reveal a lot or a little about $P[y(z_m) = 1]$, depending on the value of C. In the Perry Preschool illustration, C = .67 + .49 = 1.16.

consistent with the experimental data. The largest feasible value of P[y(2) = 0, y(1) = 0] is min{P[y(2) = 0], P[y(1) = 0]} = .33. Hence, $P[y(z_m) = 1]$ must lie in the interval [.67, 1].

Finally, assume that y(2) and y(1) are statistically independent, and also that each child receives the treatment yielding the better outcome. The experimental data and these two assumptions identify the probability of high school graduation under program m: $P[y(z_m) = 1] = 1 - P[y(2) = 0] \cdot P[y(1) = 0] = .83$.

What should an analyst conclude about the probability of high school graduation under a program in which some children receive the Perry Preschool services and the rest do not? Should the analyst conclude that $P[y(z_m) = 1]$ is in the interval [.16, 1], or [.49, .67], or [.33, .83], or [.67, 1], or that $P[y(z_m) = 1] = .83$? Of course there is no "right" answer to this question. Here, as elsewhere, the strength and credibility of the conclusions drawn depend on the assumptions imposed.

6. CONCLUSION

The importance of social programs to a diverse population creates a legitimate concern that the findings of evaluations be widely credible. The weaker are the assumptions imposed, the more widely credible are the findings. The classical argument for random assignment of treatments is viewed by many as enabling evaluation under weak assumptions, and has generated much interest in the conduct of experiments. But the classical argument certainly does impose assumptions, and there often is good reason to doubt their realism.

An initially appealing but ultimately empty response to this situation is to say that we should strive to design and execute better experiments. This response is ultimately empty because it is logically impossible to observe outcomes under alternative programs. Each member of the population actually receives exactly one treatment in one program. All attempts to compare programs, whether based on experimental or nonexperimental data, face the problem of counterfactual inference.

This article has examined aspects of the problem of counterfactual inference from experimental data. Many analysts analyze experimental data in classical fashion. Some researchers, finding the classical assumptions implausible, impose other assumptions strong enough to identify treatment effects of interest. In contrast, the recent literature on which we have focused explores the inferences that may be drawn from experimental data under assumptions weak enough to yield widely credible findings.

We have seen that this recent literature has two branches. One seeks out notions of treatment effects that are identified when the experimental data are combined with weak assumptions. The canonical finding is that the average treatment effect within some context-specific subpopulation is identified. The other branch specifies a population of a priori (perhaps welfare-economic) interest and seeks to learn about treatment effects in this population. Here the canonical finding is a bound on average treatment effects.

The various approaches to the analysis of experiments are complementary from a mathematical perspective, but in tension as guides to evaluation practice. My own research, whether based on experimental or nonexperimental data, reveals a preference to maintain weak assumptions and to keep attention focused on treatment effects in populations of substantive interest. If that means one can only bound the treatment effects of interest, so be it.

Different preferences are revealed in the work of researchers who view identification of treatment effects as a necessary condition for fruitful evaluation. Some impose assumptions strong enough to identify treatment effects of interest. Others seek out notions of treatment effects that are identified under weak assumptions. The reader of an evaluation reporting that some social program "works" or has "positive impact" should be careful to ascertain what treatment effect has been estimated and under what assumptions.

References

Aaron, Henry. Politics and the Professors: The Great Society in Perspective. Washington, D.C.: The Brookings Institution, 1978.

Angrist, Joshua. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." American Economic Review, 1990, 80, pp. 313-336.

Barnow, Burt. "The Impact of CETA Programs on Earnings: A Review of the Literature." Journal of Human Resources, 1987, 22, pp. 157-193.

Bassi, Laurie and Ashenfelter, Orley. "The Effect of Direct Job Creation and Training Programs on Low-Skilled Workers." In Sheldon Danziger and Daniel Weinberg, eds., Fighting Poverty: What Works and What Doesn't. Cambridge, Mass.: Harvard University Press, 1986.

Berrueta-Clement, John, et al. Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19. Ypsilanti, Michigan: High/Scope Press, 1984.

Björklund, Anders and Moffitt, Robert. "Estimation of Wage Gains and Welfare Gains in Self-Selection Models." Review of Economics and Statistics, 1987, 69, pp. 42-49.

Bloom, Howard. "Accounting for No-Shows in Experimental Evaluation Designs." Evaluation Review, 1984, 8, pp. 225-246.

Cain, Glen. "The Issues of Marital Stability and Family Composition and the Income Maintenance Experiments." In Alicia Munnell, ed., Lessons from the Income Maintenance Experiments. Boston: Federal Reserve Bank of Boston, 1986.

Campbell, Donald and Stanley, Julian. Experimental and Quasi-Experimental Designs for Research. Boston: Houghton Mifflin, 1966.

Clements, Nancy, Heckman, James, and Smith, Jeffrey. "Making the Most out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence from Randomized Trials with an Application to

the National JTPA Study." Cambridge, Mass.: National Bureau of Economic Research, Technical Working Paper 149, 1994.

Coyle, Susan, Boruch, Robert, and Turner, Charles, eds. Evaluating AIDS Prevention Programs. Washington, D.C.: National Academy Press, 1991.

Dubin, Jeffrey and Rivers, Douglas. "Experimental Estimates of the Impact of Wage Subsidies." Journal of Econometrics, 1993, 56, pp. 219-242.

Fisher, Ronald. The Design of Experiments. London: Oliver and Boyd, 1935.

Fishman, Michael and Weinberg, Daniel. "The Role of Evaluation in State Welfare Reform Waiver Demonstrations." In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Fraker, Thomas and Maynard, Rebecca. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." Journal of Human Resources, 1987, 22, pp. 194-227.

Garfinkel, Irwin, Manski, Charles, and Michalopolous, Charles. "Micro-Experiments and Macro Effects." In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Greenberg, David and Robins, Philip. "The Changing Role of Social Experiments in Policy Analysis." Journal of Policy Analysis and Management, 1985, 5, pp. 340-362.

Greenberg, David and Wiseman, Michael. "What Did the OBRA Demonstrations Do?" In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Gueron, Judith and Pauly, Edward. From Welfare to Work. New York: Russell Sage Foundation, 1991.

Harris, Jeffrey. "Macroexperiments versus Microexperiments for Health Policy." In Jerry Hausman and David Wise, eds., Social Experimentation. Chicago: University of Chicago Press, 1985.

Hausman, Jerry and Wise, David. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," Econometrica, 1979, 47, pp. 455-473.

Hausman, Jerry, and Wise, David, eds. Social Experimentation. Chicago: University of Chicago Press, 1985.

Haveman, Robert. Poverty Policy and Poverty Research: The Great Society and the Social Sciences. Madison: University of Wisconsin Press, 1987.

Heckman, James. "Randomization and Social Policy Evaluation." In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Heckman, James and Robb, Richard. "Alternative Methods for Evaluating the Impact of Interventions." In James Heckman and Burton Singer, eds., Longitudinal Analysis of Labor Market Data. New York: Cambridge University Press, 1985.

Holden, Constance. "Head Start Enters Adulthood." Science, 1990, 247, pp. 1400-1402.

Hotz, V. Joseph. "Designing an Evaluation of the Job Training Partnership Act." In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Hotz, V. Joseph and Sanders, Seth. "Bounding Treatment Effects in Controlled and Natural Experiments Subject to Post-Randomization Treatment Choice." Harris School of Public Policy, University of Chicago, 1994.

Imbens, Guido and Angrist, Joshua. "Identification and Estimation of Local Average Treatment Effects." Econometrica, 1994, 62, pp. 467-476.

Kershaw, David, and Fair, Jerilyn. The New Jersey Income Maintenance Experiment. New York: Academic Press, 1976.

Kurz, Mordechai and Spiegelman, Robert. "Social Experimentation: A New Tool in Economics and Policy Research." Research Memorandum 22, Stanford Research Institute, 1973.

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

Maddala, G. S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press, 1983.

Manning, Wilfred et al. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." American Economic Review, 1987, 77, pp. 251-277.

Manpower Demonstration Research Corporation. Summary and Findings of the National Supported Work Demonstration. Cambridge, Mass.: Ballinger, 1980.

Manski, Charles. "The Mixing Problem in Program Evaluation." Social Systems Research Institute, Paper 9309R, University of Wisconsin-Madison, 1993.

Manski, Charles. "The Selection Problem." In Christopher Sims, ed., Advances in Econometrics: Sixth World Congress. Cambridge, UK: Cambridge University Press, 1994.

Manski, Charles. Identification Problems in the Social Sciences. Cambridge, Mass.: Harvard University Press, 1995.

Manski, Charles F. and Garfinkel, Irwin, eds. Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Meyer, Bruce. "Lessons from the U.S. Unemployment Insurance Experiments." Journal of Economic Literature, 1995, forthcoming.

Moffitt, Robert. "Evaluation Methods for Program Entry Effects." In Charles Manski and Irwin Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992.

Moffitt, Robert and Kehrer, Kenneth. "The Effect of Tax and Transfer Programs on Labor Supply: The Evidence from the Income Maintenance Experiments." In Ronald Ehrenberg, ed., Research in Labor Economics. Greenwich, Conn.: JAI Press, 1981.

Munnell, Alicia, ed. Lessons from the Income Maintenance Experiments. Boston: Federal Reserve Bank of Boston, 1986.

Passell, Peter. "Like a New Drug, Social Programs Are Put to the Test." New York Times, March 9, 1993, p. C1.

Philipson, Tomas. "Self-Interested Treatment Evaluation in Experiments." Population Research Center, Discussion Paper 95-1, NORC and the University of Chicago, 1995.

Stafford, Frank. "Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor-Market Studies." In Jerry Hausman and David Wise, eds., Social Experimentation. Chicago: University of Chicago Press, 1985.

U.S. General Accounting Office. Unemployed Parents. GAO/PEMD-92-19BR. Gaithersburg, Maryland: U.S. General Accounting Office, 1992.

Woodbury, Stephen and Spiegelman, Robert. "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois." American Economic Review, 1987, 77, pp. 513-530.