Journal of Economic Perspectives—Volume 9, Number 2—Spring 1995—Pages 63–84

The Case for Randomized Field Trials in Economic and Policy Research

Gary Burtless

Social experimentation dates back more than a quarter of a century. Over that time, spending on randomized field trials for social policy has consumed well over a billion dollars (measured in 1994 dollars). In a recent catalogue of social experiments, Greenberg and Shroder (1991) identified more than 90 separate field trials involving a wide variety of distinctive research areas, including health insurance, prisoner rehabilitation, labor supply, worker training, and housing subsidies. Some of the major recent experiments are listed in Table 1. New randomized trials are launched by state and federal agencies each month. Classical experimentation in social policy has the appearance of a flourishing industry.

If social experimentation is an industry, its fortunes are less robust than this survey may suggest. Greater real resources were devoted to experimentation in the 1970s and early 1980s than have been invested since. The large-scale social experiments begun in the 1960s and 1970s were ambitious and costly attempts to estimate basic behavioral parameters—the income and price elasticities of labor supply and housing demand functions and the elasticity of demand for health care in response to alternative insurance arrangements. These lavish experiments generated hundreds of research reports and many articles in leading scholarly journals.[1] Although recent experiments have been much more numerous, they have also been narrower in focus, less ambitious, and less likely to yield major scholarly contributions.

Paradoxically, the findings of the newer and less ambitious experiments have had a larger impact on actual policy decisions. While findings from social

[1]For a critical survey of the early large-scale experiments, see Hausman and Wise (1985).

■ *Gary Burtless is a Senior Fellow, Economic Studies Program, The Brookings Institution, Washington, D.C.*

experiments have sunk out of view of most academic economists, they loom larger than ever for policymakers in state capitals and the federal government.

At the same time social experiments gained new influence in policymaking, some prominent economists grew disenchanted with this research tool and challenged the value of experiments in answering central questions about human behavior and policy effectiveness. Criticisms of such experiments by social scientists, if loud and persistent enough, can affect the willingness of policymakers to support this kind of study. Politicians are naturally suspicious of a research method involving experimentation with and possible harm to human subjects (also known as "voters"). It is important for them to understand the strengths as well as the limitations of this unique research tool.

This paper examines the rationale for field experimentation in economics and considers some of the main criticisms leveled at experiments in recent years. Academics have attacked experiments for a wide range of real and imagined sins. Recent experimental designers have taken past criticisms into account and sought to address some of the most serious ones through improved experimental design. In spite of recent criticisms, classical experimentation on a modest scale has become an accepted part of policy evaluation in the United States. The essential reason is that policymakers and many social scientists find experimental results easier to understand—and ultimately more convincing—than results from most other kinds of policy evaluation.

Definition of Social Experiments

In common parlance, an experiment is any major deviation from past policy or practice. Under this broad definition, the introduction of Social Security in 1935 and the sharp reduction in top marginal income tax rates in 1981 represent policy experiments. The scientific notion of experiments is considerably narrower. It emphasizes the researcher's control of the variables under investigation and over the environment in which those variables are observed. In a typical scientific experiment, the investigator deliberately manipulates the environment or introduces change into the environment to measure the consequences of change.

The income tax reduction passed in 1981 does not represent an experiment under this definition, because policymakers exercised little control over most aspects of the environment that affected economic agents' responses to policy change. For example, U.S. gross domestic product fell in three of the first four quarters after passage of the 1981 tax cuts. One possibility is that tax cuts *caused* the recession. More plausibly, the recession affected consumers' and producers' responses to the tax changes introduced in 1981. The pure effects of the tax cuts on consumer and producer behavior were never directly observed. Instead, they have been inferred by economists after disentangling the effects of other changes in the environment. Since it is unclear how analysts can reliably

Table 1
Selected Social Experiments

Experiment	*Target population*	*Tested treatment(s)*	*Notable publications*
Negative Income Tax (NIT) Experiments (1968–1978)	Low- and moderate income families headed by nonaged adult	NIT plans with alternative income guarantees and tax rates	Keeley et al. (1978); Burtless and Hausman (1978); Munnell (1987)
Housing Allowance Demand Experiment (1973–1977)	Low- and moderate income families	Alternative income supplement plans designed to help low-income households pay for housing costs	Struyk and Bendick (1981); Bradbury and Downs (1981)
RAND Health Insurance Experiment (1974–1982)	Nonaged low- and moderate income persons and families living outside of institutions	Health insurance plans that varied over two dimensions: Upper limit on out-of-pocket medical expenses and copayment rates ranging from 0% (free care) up to 95%	Brook et al. (1983); Manning et al. (1987)
Electricity Time-of-Use Pricing Experiments (1975–1981)	Residential consumers of electricity	Alternative pricing schedules for electricity in which prices vary by time-of-day or season of year	Caves and Christensen (1980); Aigner (1985)
National Supported Work Demonstration (1975–1980)	Long-term AFDC recipients; former drug addicts; ex-offenders; young school dropouts	12–18 months of structured work experience and on-the-job training, using peer-group support and sympathetic supervision	Board of Directors of MDRC (1980)
MDRC Work-Welfare Experiments (1982–1988)	AFDC applicants and recipients	A variety of voluntary and mandatory work-oriented programs, including job search, skills training, and unpaid public employment	Gueron and Pauly (1991)
National Job Training Partnership Act (JTPA) Study (1986–1994)	Disadvantaged adults and out-of-school youth who enroll in programs funded under Title IIA of JTPA	Job search assistance, classroom training, on-the-job training, and other forms of training financed under JTPA, Title IIA	Bloom et al. (1993)

establish the effects of other environmental factors, the effect of the 1981 tax changes on economic behavior remains a subject of intense controversy.

In many scientific experiments, the investigator simply introduces a change in a controlled environment and observes the effect of the change on the material or organism under study. Of course, reliable measurement of the effect requires some basis for comparison. *Consumer Reports* tests the strength of automobile bumpers by subjecting them to a uniform blow and then determining the cost of the necessary repairs. The implicit basis of comparison is the pristine state of the tested vehicle before the blow was delivered. However, a before-and-after comparison is not always appropriate or feasible. The toxic effect of Twinkies cannot be discovered simply by observing that 10 percent of laboratory mice die within one month of eating a Twinkie. No matter how well controlled the environment in which the experiment is conducted, some mice will die whether or not they consume a questionable dessert product. If the usual mortality rate of laboratory mice is 3 percent a month, the extra mortality from consuming a Twinkie is 7 percent. If mortality is usually 12 percent, the dessert is not toxic at all; it reduces mortality by 2 percent a month.

The problem, of course, is establishing a credible basis of comparison. In the previous example, an investigator might compare the mortality experience of Twinkie-eating rodents with that of a handpicked group of mice that is deprived of cream-filled desserts. Naturally, the researcher will want this comparison group to be as similar as possible to the group of mice offered Twinkies. As early as the 1920s, R. A. Fisher (1928, p. 230) argued that the only fully satisfactory method of achieving equivalence between the treatment and comparison groups was to assign subjects to the two groups "wholly at random." This kind of observational study is commonly referred to as a randomized or controlled experiment.

Fisher's case for randomization was widely influential in agricultural, biological, and medical research, but its impact on economic research has been smaller and much slower to develop. Many economists probably believe that the important questions at issue in economics cannot be answered with randomized trials. While this is certainly true of questions involving general equilibrium or movements in economy-wide aggregates, many propositions in economics treat behavioral response at the individual, family, or company level: saving responses to movements in real interest rates, labor supply responses to changes in after-tax wages or unearned income, labor demand responses to tax subsidies, consumption responses to changes in relative price, and so on. In principle, all of these issues can be studied using small- or medium-sized randomized trials. It is certainly feasible for researchers to assign economic agents randomly to different policy regimes in the same way that agricultural scientists assign dairy cattle to different feeding regimes or small plots of land to different varieties of fertilizer. Readers may be uncomfortable with the implied equivalence between human beings and farm livestock, but the statistical rationale for randomization is essentially the same in both cases. However,

whether it is ethical, useful, or cost-effective to carry out such experiments on humans is a matter of debate.

The critical element that distinguishes controlled experiments from all other methods of research is the random assignment of meaningfully different treatments to the observational units of study. In the context of social science, a randomized field trial (or social experiment) is simply a controlled experiment that takes place outside a laboratory setting, in the usual environment where social and economic interactions occur. In the simplest kind of experiment, a single treatment is assigned to a randomly selected subsample (the treatment group) and withheld from the remainder of the enrolled sample (the control or null-treatment group). Many social experiments have tested a variety of different treatments rather than only one. Some have not enrolled a pure control group at all. Instead, the investigators have concentrated on measuring the *differences* in effect of a number of distinctive new treatments. The definition of an experiment can include tests of innovative new policies as well as studies that are intended to measure the effect of current policies relative to a null treatment.

Analysts distinguish between two kinds of social experiments, one of which aims to estimate the underlying parameters of a population response function and a second that attempts to measure the overall effects of one or more distinctive treatments. An example of the first type of experiment is the Seattle-Denver negative income tax (NIT) experiment, which tested a variety of combinations of income guarantees and tax rates in order to estimate labor supply functions in the low-income population (Munnell, 1987). In this kind of "structural" experiment, individual treatments can be defined as points within a continuous policy parameter space, and the experimental objective is to estimate a smooth response surface.

The second kind of experiment tests a sort of "black box," in the sense that each treatment tested represents a unique intervention. Most experiments in employment training policy—indeed, most recent experiments in *any* policy area—have been black box experiments. In this kind of experiment, one or more specific combinations of government services are tested. Even those experiments testing multiple treatments can find no natural way to parameterize the experimental treatments as points along a policy continuum. Thus, the results of black box experiments cannot be easily extrapolated to infer the effects of similar but nonidentical treatments.

In the absence of information from social experiments, economists and other social scientists rely on four main alternatives to experiments to learn about crucial behavioral parameters or the effectiveness of particular programs. One source of information is data on the relationship between economy-wide aggregates, such as interest rates and consumption, either over time or across regions. However, aggregate statistics are inappropriate for analyzing many kinds of microeconomic behavior. A second source is management data collected in the administration of existing programs, but data from an existing program seldom provide any information about what the participants' experi-

ences would have been if they had been enrolled in a different program or in no program at all. A third source is new survey data, which are usually more costly to obtain than programmatic data but which provide information about the experiences of nonparticipants as well as participants in a program, and thus offer some evidence about likely behavior in the absence of treatment. A fourth source is data generated by special demonstration programs. Like experiments, demonstrations involve the special provision of a treatment, collection of information about outcomes, and analysis of treatment effects. Unlike experiments, demonstrations do not involve random assignment. What all experiments have in common, whether based on black box or structural designs, is that the tested treatments are randomly assigned to observational units—that is, to individuals, companies, government offices, or entire communities.

Advantages of Experimentation

The advantages of controlled experimentation over other methods of analysis are easy to describe. Because experimental subjects are randomly assigned to alternative treatments, the effects of the treatments on behavior can be measured with high reliability. The assignment procedure assures us of the direction of causality between treatment and outcome: differences in average outcomes among the several treatment groups are caused by differences in treatment, and differences in average outcome are not the cause of the observed differences in treatment. Causality is not so easy to determine in nonexperimental data. In measuring the response of health spending to alternative health insurance plans, for example, it is unclear in nonexperimental data whether generous insurance coverage causes high health spending or large anticipated health bills cause consumers to purchase generous health insurance policies. In an experiment where insurance plans are assigned to people at random, the direction of causality is certain.

Random assignment also removes any systematic correlation between treatment status and both observed and unobserved participant characteristics. Estimated treatment effects are therefore free from the selection bias that potentially taints all estimates based on nonexperimental sources of information. In a carefully designed and well-administered experiment, there is usually a persuasive case that the experimental data can produce an internally valid estimate of average treatment effect.[2]

[2]An "internally valid" estimate is an unbiased measure of the treatment effect in the sample actually enrolled in an experiment. An "externally valid" estimate is a treatment-effect estimate that can be validly extrapolated to the entire population represented by the sample enrolled in the experiment. Some experimental estimates may not be internally valid, perhaps because treatment is not assigned randomly or because attrition produces noncomparable treatment-group and control samples. In addition, some internally valid estimates may lack external validity. One reason is that a treatment offered to a small experimental sample may not correspond to any treatment that could actually be provided to a broad cross section of the population. Other reasons are described later.

Another advantage of experiments is that they permit analysts to measure —and policymakers to observe—the effects of economic stimuli or new kinds of treatment that have not previously been observed. In many cases the naturally occurring variation in relative prices or policy treatments is too small to allow economists to infer reliably the effects of potential price movements or promising new policies. Many politicians believe, for example, that private employers would hire a greater number of disadvantaged workers if the wages paid to these workers were generously subsidized. If a program of this type has never been tested, it would be difficult or impossible to forecast employer response to a particular wage subsidy level. Of course, new policies can be tested in demonstration programs, too. But in comparison with most sources of nonexperimental information, experiments permit economists to learn about the effects of a much wider range of prices and policies.

Finally, the simplicity of experiments offers notable advantages in making results convincing to other social scientists and understandable to policymakers. A carefully conducted experiment permits analysts to describe findings in extremely straightforward language: "Relative to employers in the control group, employers eligible for government-provided wage subsidies hired X percent more disadvantaged adults and Y percent fewer workers who were not economically disadvantaged." This kind of simplicity in describing results is seldom possible in nonexperimental research, where analytical findings are necessarily subject to a variety of complicated qualifications.

In recent years the last advantage of experiments has turned out to be particularly important for experiments that test practical policy alternatives. Because policymakers can easily grasp the findings and significance of a simple experiment, they concentrate on the implications of the results for changing public policy. They do not become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid. Politicians are more likely to act on results they find convincing.

Social experiments have contributed to important advances in basic knowledge, improved understanding of program effectiveness, and, in rarer cases, significant policy reform. The Health Insurance Experiment, for example, provided convincing evidence about the price sensitivity of the demand for medical care. Even more important, it gave medical practitioners and policymakers unprecedented information about the health consequences of variations in medical care that are induced because consumers face different prices for medical treatment as a result of differences in their insurance coverage (Brook et al., 1983; Manning et al., 1987). Results from the Job Training Partnership Experiment suggest that the government's main training programs for disadvantaged adults yield significant gains in participants' employment and earnings, although programs targeted on out-of-school youth are ineffective and possibly even harmful (Bloom et al., 1993). The administration and Congress scaled back Job Training Partnership Act (JTPA) funding for youth programs partly as a result of these findings. The Manpower Demonstration Research

Corporation (MDRC) is responsible for experiments that produced the most notable and immediate effect on policy. A series of studies known as the Work-Welfare Experiments offered tangible evidence that work-oriented training and job search programs could boost employment and reduce welfare dependency in the AFDC population (Gueron and Pauly, 1991). The experimental findings were persuasive enough to lawmakers to have a significant impact on the design and implementation of the 1988 Family Support Act.

Of course, experimentation does not completely eliminate uncertainty about the correct answer to a well-posed question about economic behavior or policy, as discussed below. But it can dramatically reduce uncertainty. More important, the small number of qualifications to experimental findings can be explained in language that is accessible to people without formal training in statistics or economics. This is a crucial reason for the broad political acceptance of findings from recent labor market experiments.

The analytical advantage of experiments over nonexperimental research methods can be described in terms of a simple model of treatment effect.[3] Suppose the true behavioral model, including treatment effects, is

$$Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i,$$

where Y is the behavioral outcome of interest; T_i is the treatment dosage received by the ith sample member; X_i is a measured or unmeasured characteristic of person i that influences Y; and ε is an error term that captures the effects of random error or measurement error in Y—it is uncorrelated with T and X. T might represent eligibility for a particular government service, say, employment training for the disadvantaged, while X could represent a person's academic achievement as measured on a standardized test. In this example, Y would usually indicate individual earnings, an outcome that training is supposed to improve. Assuming that T and X actually vary and can be measured without error, least squares estimation using information from surveys of people who receive different doses of T can be used to obtain an unbiased estimate of the program treatment effect, β. Under these circumstances there is no need for an experiment.

Where X is not observed, however, the least squares estimates of α and β may be biased, depending on the relationship between X and T.[4] If X and T

[3]This exposition borrows from Leamer (1983).

[4]In particular, suppose X and T are related by $E(X_i/T_i) = r_0 + r_1 T_i$. In this case, the least squares estimates will be biased by:

$$E(Y/T) = \alpha + \beta T + \gamma E(X/T)$$

$$= \alpha + \beta T + \gamma(r_0 + r_1 T)$$

$$= (\alpha + \alpha^*) + (\beta + \beta^*)T,$$

where α^* and β^* are measures of the least squares bias. The bias will increase as γ and the correlation between X and T increase in absolute value.

are correlated with each other, the β coefficient will be biased. This could easily occur, for example, if unmeasured academic ability, X, affects the willingness of a person to enroll in training, T. The estimation problem disappears in a classical experiment. By the definition of an experiment, the treatment doses, T_i, are randomly assigned within the estimation sample and, presumably, are accurately measured. T will therefore be uncorrelated with both X and ϵ, implying that least squares estimation of Y on T can produce an unbiased estimate of β.

A common source of bias in microeconomic statistical studies is sample selection. Nonexperimental studies of education and training, for example, usually rely on observations of naturally occurring variation in treatment doses in order to form estimates of the effects of training. Analysts typically compare measured outcomes in employment or earnings for participants in a training program and for a comparison group of similar people who did not participate in the program. The value of a college degree is often calculated by comparing the earnings of college graduates with the earnings of similar people who graduated from high school but did not attend college. Even if the analysis fully controls for the effects of all measurable characteristics of sample members, it is still possible that average outcomes are influenced by systematic differences in the unmeasured characteristics of individuals in the treatment and comparison groups. In the simplest kind of training program, for example, members of the estimation sample are exposed to just two doses of treatment: $T = 1$ for people enrolled in the training program, and $T = 0$ for people who never enrolled. Program participation represents the sample member's decision to choose one treatment dose or none. Obviously, this decision may be affected by unobserved tastes or other characteristics, which also affect a person's later employment or earnings. Since these factors are unknown and cannot be estimated, the amount of bias in the nonexperimental estimate of β is unknown.

Selection bias is a practical estimation problem in most nonexperimental policy studies. Naive analysts sometimes ignore the problem, implicitly assuming that unmeasured differences between training participants and nonparticipants either do not exist or do not matter. While one or both of these assumptions might be true, the case for believing either of them is usually quite weak. Critics of early training evaluations often pointed out, for example, that people who voluntarily enrolled in employment training for the disadvantaged might be more ambitious than typical disadvantaged workers, yet ambition is an unmeasured personal trait. If personal ambition is correlated with a person's subsequent labor market success, it is unclear what percentage of the average earnings advantage among trainees is due to extra training and what percentage is due to the greater average ambition of trainees.

The selection bias could go in the opposite direction, too. Disadvantaged workers who are less optimistic about their labor market prospects may enroll in job training in disproportionate numbers. If their pessimism is based on a realistic—but unobserved—assessment of their opportunities, we should expect

that their earnings in the absence of training would be lower than those of people with identical observable characteristics who do not enroll in training.

Our uncertainty about the presence, direction, and potential size of selection bias makes it difficult for social scientists to agree on the reliability of estimates drawn from nonexperimental studies. The estimates may be suggestive, and they may even be helpful when estimates from many competing studies all point in the same direction. But if statisticians obtain widely differing estimates or if the available estimates are the subject of strong methodological criticism, policymakers will be left uncertain about the effectiveness of the program. Ashenfelter and Card (1985) and Barnow (1987) have shown that this kind of uncertainty is more than just a theoretical possibility. Both studies found an uncomfortably wide range of estimated impacts of the Comprehensive Employment and Training Act (CETA) on earnings when estimates were based on nonexperimental data. The range of plausible estimates reported in these studies, especially for male trainees, was too wide to permit policymakers to decide whether CETA-sponsored training was cost-effective. It was not even clear for some groups whether the impact of CETA training was positive.

Is There an Econometric Fix?

Econometricians have proposed sophisticated methods to test for the presence of selection bias and to obtain estimates of treatment effects that are purged of selection bias (Heckman, 1976; Maddala and Lee, 1976; Barnow, Cain, and Goldberger, 1980; Heckman and Robb, 1985). The problem with these methods is that they rest on an ultimately untestable assumption about the distribution of the error term or the specification of the equation representing the decision to participate in a program. If critics of a nonexperimental estimator question the reliability of the key assumption, other social scientists (and policymakers) often have no reliable method to decide whether the maintained assumption is a good approximation of reality. In the case of a randomized trial, the key assumption for reliable estimation is that the experimenter has been successful in randomly assigning subjects to different treatments and in measuring the responses of subjects exposed to the treatments. Statisticians and lay observers ordinarily find it much easier to assess the validity of this assumption than to decide whether the key assumptions of nonexperimental studies are valid.

LaLonde (1986) and Fraker and Maynard (1987) studied the reliability of a variety of nonexperimental estimators with an ingenious procedure. Using data from a true randomized trial—the National Supported Work Demonstration—the two sets of authors compared actual estimates obtained in the demonstration with nonexperimental estimates that would have been obtained if no information had been available from the Supported Work Demonstration's control group. To derive their nonexperimental estimates, the authors selected

a variety of nonrandom comparison groups drawn from general population surveys, such as the Current Population Survey and Panel Study of Income Dynamics, and used several methods to control for the problem of sample selection bias. Many of these methods had been used in the earlier evaluation literature on CETA. Neither study found nonexperimental estimators to be reliable. For some groups and most estimators, the nonexperimental estimates of effect differed substantially from the estimate based on a true, randomly selected control group. More disturbingly, LaLonde (1986, p. 617) found that "even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results."

Heckman and Hotz (1989) have shown that some key assumptions of certain nonexperimental estimation methods can be systematically analyzed using available data. Some assumptions of incorrect models can potentially be rejected by formal specification tests. These tests may not have much statistical power to reject incorrect models, however. Then, the nonexperimental estimators that are not rejected by formal specification tests may yield varying estimates of the effectiveness of a particular program. Analysts will still be left with the difficult choice of deciding which model of sample selection is most plausible.

In the case analyzed by Heckman and Hotz (1989), specification tests led to the rejection of several incorrect models of selection. This ruled out some nonexperimental estimates of the effect of the Supported Work program that were clearly implausible in the light of estimates obtained using the classical experimental estimator. However, the nonexperimental estimators that were *not* rejected by the Heckman-Hotz specification tests had much more sampling variability than the classical experimental estimator. In other words, the classical experimental estimator still had a major advantage over the nonexperimental estimators for users who care about the statistical precision of the estimates they use. But the more important advantage is that the validity of the experimental estimator depends upon assumptions that are ordinarily much easier to evaluate—and to believe.

Problems with Experiments

Randomized field trials face numerous problems, of course. Many have been acknowledged since the inception of large-scale social experimentation (Campbell and Stanley, 1966; Rivlin, 1974). Others have come into prominence in recent years (Heckman, 1992; Garfinkel et al., 1992; Levitan, 1992). Several supposed problems turn out to be shortcomings of social research in general or survey research in particular, rather than with experimentation per se. In some cases, of course, difficulties with experimentation can be insuperable. Under those circumstances it makes no sense to conduct an experiment. This still leaves many areas of microeconomic and policy research where social

experimentation represents a cost-effective way to improve basic knowledge. Problems unique to experimentation are often relatively minor in comparison with those that plague nonexperimental research.

Cost

Experiments have three kinds of cost that can make them more expensive than nonexperimental research on the same topic. They consume a great deal of real resources, especially in comparison with nonexperimental analysis of existing data sources. They are almost always costly in terms of time. Several years usually elapse between the time an experiment is conceived or designed and the release of its final report. If policy decisions about a particular public policy cannot be deferred, the usefulness of an experiment may be questionable.

In addition, experiments often involve significant political costs. It is more difficult to develop, implement, and administer a new treatment than it is simply to analyze information about past economic conditions or collect and analyze new information about economic behavior. Voters and policymakers are rightly concerned about possible ethical issues raised by experiments (discussed further below). As a result, it is usually easier to persuade officials to appropriate small amounts for pure research or medium-sized sums for a new survey than it is to convince them that some people should be systematically denied a potentially beneficial intervention in a costly new study.

These disadvantages of experiments are real, but should be placed in perspective. Some forms of nonexperimental research suffer from identical or similar disadvantages. A demonstration program, which lacks a randomly selected control group, can easily cost as much money as a social experiment that tests the same innovative treatment. The demonstration will certainly take as much time to complete as an experiment. If a new survey must be fielded to obtain the needed information, the extra time and money required for an experiment may seem relatively modest.

Ethical Issues of Experimentation with Human Beings

Many observers are troubled by the ethical issues raised by experimentation with human beings, especially when the experimental treatment (or the denial of treatment) has the capacity to inflict serious harm. If the tested treatment is perceived to be beneficial, program administrators may find it hard to deny the treatment to a randomly selected group of people enrolled in a study. Except among philosophers and research scientists, random assignment is often thought to be an unethical way to ration public resources. If, on the other hand, the tested treatment is viewed as potentially harmful, it will be difficult to persuade policymakers to undertake the experiment or to recruit program managers to run the project. It may not be ethical to mount such an experiment in any event. Readers should recall, however, that similar ethical issues arise in studies of new medicines and medical procedures, where the

stakes for experimental participants are usually much greater than they are in a social experiment. Yet randomized field trials have been common in medicine for far longer than they have in social policy. In fact, such trials are often required to prove the efficacy of new medical treatments.

Good experimental design can reduce ethical objections to random assignment (Burtless and Orr, 1986, pp. 621–24; the essays in Rivlin and Timpane, 1975). At a minimum, participants in experimental studies should be fully informed of the risks of participation. Under some circumstances, people offered potentially injurious treatments or denied beneficial services should be compensated for the risks they face. The risks of participation are frequently unclear, however, because it is uncertain whether the tested treatment will be beneficial or harmful.

Of course, uncertainty about the direction or size of the treatment effect is the main reason that an experiment is worthwhile. If successful, an experiment will substantially reduce our uncertainty about the size and direction of the treatment effect. The ethical argument in favor of experimentation is that it is preferable to inflict possible harm on a small scale in an experimental study rather than unwittingly inflict harm on a much larger scale as a result of misguided public policy.

Limited Duration

Social experiments are limited in duration. For some kinds of treatments, this poses a problem for valid inference. As one example, participants may take time to understand the nature of the tested treatment and to react to it. A more serious issue is that participants may react differently to a treatment if they know in advance that it is of limited duration, compared with how they would react if the treatments were expected to last indefinitely. An experimental housing subsidy that will last just one year will presumably have a smaller effect on housing decisions than an equally generous subsidy that is expected to be permanent.

The limited duration of social experiments is not always an issue. The aim of many experiments is to test a short-duration intervention that is supposed to immediately benefit a target population, for example, by improving school achievement, employment rates, or subsequent earnings. The intervention itself lasts only a few weeks or months and is completed long before the end of the experiment. Even where limited duration is a critical issue, as in the housing allowance experiments, research planners can explicitly address the issue through sensible experimental design. The duration of the treatment can itself be experimentally varied to determine whether the length of a subsidy affects the size of response.

Attrition and Interview Nonresponse

Critics of social experiments see statistical problems with experiments in addition to the difficulties connected to cost and ethical propriety. Several

experiments have been criticized, for example, because high attrition in either the treatment or control groups has meant that even though the samples originally enrolled in the two groups were randomly selected from an identical population, the members of the treatment and control groups ultimately used in the analysis were self-selected members of nonidentical populations as a result of different rates of attrition. The crucial advantage of random assignment has been lost. For example, the negative income tax (NIT) experiments suffered high attrition among families enrolled in the control group and in some of the less generous NIT plans. The final analysis samples were thus unrepresentative of the population that was originally enrolled in the experiments. Because attrition differed in the treatment and control groups, it is possible that the average outcome difference between the two groups was partly due to compositional differences between the groups as well as to the pure effect of the NIT treatment.

While this criticism is valid for some social experiments, it is hardly one that applies only or even mainly to experiments. It applies to all research studies, whether experimental or nonexperimental, that rely on longitudinal survey data in which attrition or interview nonresponse is a problem. In the United States, such surveys include the Panel Study of Income Dynamics, the National Longitudinal Surveys, the Retirement History Survey, and even the Current Population Survey, each of which has been used in a large number of nonexperimental studies. Ironically, more is known about the influence of attrition in experimental studies because experimental designers often take extraordinary steps to reduce its effects or to measure its impact.[5]

In fact, many recent social experiments have abandoned longitudinal surveys as a method of gathering information about behavioral outcomes. People enrolled in an experimental sample are given a baseline interview upon enrollment and are never interviewed again. To measure behavioral outcomes, the experiments rely on transfer payment records maintained by public assistance authorities and employment and earnings records supplied by social insurance agencies. It is unlikely that either source of information is seriously affected by differential attrition or nonresponse bias. Thus, the problems of nonrandom attrition and interview nonresponse are not intrinsic to social experimentation.

Partial Equilibrium Results

Experimental findings are often criticized because they do not reflect the general equilibrium effect of a particular price or policy change. In small-scale

[5]In an effort to minimize attrition, for example, the negative income tax (NIT) experiments paid both treatment-group and control-group members for their continued participation. In addition, three of the four experiments checked their interview-based estimates of the treatment effect against an estimate based on information not derived from interviews. Analysts collected earnings records maintained by the social insurance authorities and reestimated the effects of the NIT plans using this information.

training experiments, for example, the advantages of experimental training are conferred on only a small fraction of the people who would benefit under a full-blown national program. The benefits a worker derives from extra training are magnified when few other people in the local labor market receive additional training; employers can choose among only a handful of workers with improved qualifications. If the entire eligible population were offered training, the effects of the program on employment or average earnings would almost certainly be smaller. Similarly, the negative income tax experiments measured the supply-side effects of higher marginal tax rates and more generous income guarantees. But without knowing how employers would alter their wages if more generous income guarantees reduced labor supply across the board, it is impossible to forecast the full general-equilibrium effect of a more generous income transfer system.

While it is true that randomized field trials can measure, at most, partial equilibrium responses to a price or policy change, the partial equilibrium effect is often the response of critical interest.[6] If a training program is intended to improve the job prospects of disadvantaged workers and is found in a small-scale experiment to have no detectable influence on employment or earnings, the general equilibrium effects of the program are probably small and certainly irrelevant. For an experimental program found to raise the employment rate of trainees, analysts are still left with the problem of determining the general equilibrium effect of a full-fledged program, but at least they will be in a better position to predict the general equilibrium effects of the program.

Garfinkel et al. (1992) recently argued that experiments miss at least three other kinds of general equilibrium effects that determine the net impact of a program. First, a small-scale experiment cannot offer participants the information or helpful insights about a program that would be available in a community-wide or nationwide program. Other omitted effects include social interaction and norm formation processes that would be present when a large percentage of the population is affected by a program but which are absent when only a minuscule proportion of the eligible population is enrolled.

These effects are a problem for social experiments—but they often represent an equally serious challenge to nonexperimental research methods. All three require an unknown amount of time to affect observed behavior. This means the statistician cannot assume that the price or policy in effect at a given time determines individual behavior at that time; behavior is also determined by past prices or policies, with an unknown weight on the price or policy in

[6]General equilibrium effects could be measured in field trials if entire communities were assigned at random to one policy regime or another. A large number of communities would have to participate in such an experiment, however, and the administrative and research costs of the experiment would therefore be large. The experimental housing allowance supply experiment was intended to measure general equilibrium effects through the provision of housing allowances to eligible families in two communities. The number of communities offering subsidies was too small to measure the full market response reliably, however.

effect in each past period. Some statisticians are confident they can estimate these weights with naturally occurring data, but the estimation problem is formidable, especially for price or policy variables that remain constant over long periods or vary little from one individual to the next.

A concrete example can illustrate the point. In 1961 federal law was changed to permit men aged 62–64 to draw early Social Security benefits. A randomized trial of alternative early retirement ages conducted in the late 1950s might have shown that the availability of Social Security benefits at progressively younger ages caused labor force withdrawal at younger ages. These results could be criticized. An experiment would have missed the effects of information diffusion, social interaction, and induced shifts in social norms, each of which would presumably magnify the response observed in a small-scale trial.

A nonexperimental statistical study conducted before 1961 would have faced an even more severe problem, however: early retirement benefits for men had never been offered before 1961. A nonexperimental study conducted in a year after 1961 has the advantage of greater variation in the policy variable of interest, but since 1961, there has been no further change in the availability of early pensions. Moreover, economists are uncertain when the general equilibrium effects of the 1961 reform became fully operational. Did retirement norms in 1965 fully reflect the effects of the 1961 reform? Or did norms continue to adjust through 1970? Through 1980? Garfinkel et al. (1992) have not identified a statistical technique available to nonexperimental researchers that would avoid the problem in analysis of nonexperimental data.

Program Entry Effects

A recent criticism of experiments is that the population enrolled in the treatment and control groups is not representative of the population that would be affected by the treatment if it were provided in an ongoing, national program. While occasionally valid, this criticism applies equally to a nonexperimental study when the object of the study is to predict the effects of a new public program or major reform of an old one.

Suppose, for example, that policymakers would like to know the effects of a new kind of training program when some or all public assistance recipients are required to participate in the program as a condition for obtaining welfare. An experiment could enroll a sample of welfare recipients representative of the population that is supposed to participate in the program. A random subsample of this group would be asked to enroll in training; the remainder of the sample would be assigned to a control group. The experiment could accurately measure the effect of the training requirement in a small-scale trial.

However, if the tested program was made permanent and extended to the entire welfare population, it could affect the entry or exit of people from the public assistance rolls (Moffitt, 1992). A training program that is regarded as burdensome by assistance recipients, for example, will deter some people from

applying for welfare and persuade some assistance recipients to exit the rolls faster than they otherwise would.[7] Neither of these effects will have occurred when the experiment begins. It follows that the sample enrolled in the experiment will not be representative of the population that would participate in the program if the treatment were made permanent and were universally available to the eligible population.

In this respect, however, an experiment suffers from no disadvantage in comparison with many kinds of nonexperimental analysis. If no similar training requirement had been imposed in the past, nonexperimental researchers are not in any better position to estimate program entry effects than experimental researchers. They are in a much worse position to estimate program exit effects, since at least the experiment offers evidence on people's likelihood of leaving the program. However, if a similar training requirement had been imposed in the past, a nonexperimental analysis could provide better information about program entry than any small-scale experiment.

Experiments in Ongoing Programs

A new wave of criticisms has been levelled against a relatively new kind of social experiment—one that takes place in an existing program. A prominent experiment of this type is the recent evaluation of JTPA (Bloom et al., 1993). The JTPA is the major U.S. program that pays for employment training and job search assistance for economically disadvantaged workers. Sixteen local areas participated in the experiment. In each site up to a third of the people who applied for job training services were randomly selected and enrolled in a control group. They were not permitted to receive services funded by JTPA during the first 18 months after they initially applied for services. Analysts obtained an estimate of the JTPA treatment effect by measuring the difference in average outcomes between members of the control group and people offered the JTPA services.

Heckman (1992) has strongly criticized the JTPA experiment for at least three reasons. One is that the sites may have enrolled a different group of trainees than would have been enrolled in the absence of the experiment (this is "sample contamination"). Since about a third of applicants were enrolled in the control group rather than JTPA training, the sites enrolled participants who would not have applied for services or who would have been denied services without the experiment. Heckman's second criticism is that normal JTPA services were disrupted as a result of the pressures created by the experimental protocol (this is "treatment contamination"). Finally, Heckman criticized the experiment because only a small handful of potential sites were enrolled in the study, and all of them were self-selected. The 16 sites that

[7]Conversely, a training program that is regarded as exceptionally beneficial might actually attract extra applicants to the public assistance rolls.

agreed to participate may not have been representative of JTPA training sites throughout the United States.

These criticisms have merit. But none of them are a necessary by-product of random assignment. Instead, they resulted from the Labor Department's initial choices about experimental design. The potential problems of sample contamination, treatment contamination, and site self-selection could have been avoided if the department had enrolled a small number of control observations in many representative sites rather than a large number of observations in only a handful of sites. In its design of a new randomized trial to evaluate the Job Corps program, the Labor Department learned from its experience in the JTPA experiment and has decided to enroll small numbers of control observations in many sites.

Of course, it is more costly to enroll experimental participants in many sites rather than only a few. The higher costs of a better design should be weighed against the extra benefits the experiment would provide if its results were more reliable. In case of the JTPA experiment, the findings of the experiment would have been safely generalizable to the entire population served by JTPA. Users of the results could have been assured that the services evaluated in the study were identical to those typically offered by local agencies. The added costs would have been small, in my view, in relation to the benefits from a better design. The federal government spends about $1.8 billion each year on employment and training services for the disadvantaged under JTPA. The cost of the experiment was less than $50 million. Even if including more sites in the study had increased the experimental cost by a factor of five, the cost of the evaluation would still have represented less than 5 percent of the operating budget of the JTPA program over the three-year period covered by the experiment. Since the results of the study suggest that economically disadvantaged youngsters may be hurt by their participation in JTPA, lawmakers and taxpayers—as well as some JTPA participants—would be better off if we had greater confidence in the reliability of the experimental findings.

Other Criticisms

A simple experiment cannot answer some questions about the behavioral response to a treatment (Heckman, 1992). In particular, randomized trials often fail to provide an unbiased estimate of the effect of a program on those who actually participate in the program. If only 30 percent of a sample that is randomly assigned to a training program actually receives services under the program, then 70 percent of the enrolled sample has declined to participate. An experiment can provide a valid estimate of the average impact of the *offer* of training services. It cannot provide a reliable estimate of the average impact of services *among participants who actually receive them*, without additional assumptions about the determinants of participation. Moreover, while an experiment can yield a valid estimate of the mean difference between the treatment and

control groups, without additional assumptions it cannot provide estimates of other critical parameters in the response function, like the median response.

These criticisms of experiments are valid, but rarely decisive. One reason is that nonexperimental data are subject to the same shortcomings (or worse ones). Another is that experiments can usually provide valid estimates of the most important behavioral response. Many policy issues hinge on the *average* response to a particular price or policy change. The median response, the variability of response, and the overall shape of the response function hold intellectual interest, but if policymakers could be confident that the average impact of the treatment offer is positive, most would serenely accept their ignorance of the higher moments of the response distribution. Furthermore, in traditional cost-benefit analysis, the mean effect of the intervention is the crucial determinant of its social usefulness. How the benefits are divided between participants and nonparticipants affects the distribution of social gains but not the net social gain from the treatment. If analysts believe it is important to estimate the distribution of gains across participants and nonparticipants, they can certainly use nonexperimental research methods to analyze evidence collected in an experiment. Random assignment does not prevent researchers from using nonexperimental methods to analyze the data. It does offer them a powerful source of identifying information to measure the average effect of the treatment offer.

Conclusion

Are experiments preferable to nonexperimental evaluations? The answer will differ for different research questions and in different circumstances. The more relevant question for a particular research question is this: what are the costs and benefits of an experiment compared with the available alternatives?

Even if it were true that reliance on an experiment involved risk of drawing incorrect conclusions about policy effectiveness, the alternatives to an experiment often produce conclusions that are much less reliable. As we have seen, nonexperimental studies are usually plagued by more serious statistical problems than those that occur in randomized trials. More fundamentally, the failure to conduct any evaluation research at all can lead to the perpetuation of programs that are less effective and more costly than policy alternatives. In some cases, our failure to evaluate reliably can lead to the perpetuation of policies that are actually harmful to intended beneficiaries.[8]

[8]Evidence from a pair of experiments involving the Targeted Jobs Tax Credit suggests, for example, that this employment subsidy program significantly harms the job-finding opportunities of disadvantaged job seekers (Burtless, 1985; Masters et al., 1982). This finding is exactly the opposite of the one predicted by simple intuition, standard economic theory, and previous nonexperimental studies.

How can we assess whether the greater reliability of experimental results is worth the extra cost of obtaining them? When the direct benefits from improved knowledge are easy to predict and measure, analysts can calculate the financial gains from improved decision making and compare them with the additional costs associated with conducting an experiment. An experiment should be undertaken when the value of the improved decision exceeds the extra cost of the experiment (Stafford, 1979; Burtless and Orr, 1986). Potential benefits are clearest when the focus of study is narrow, as is the case when the government wishes to determine whether a particular current policy is effective or whether a small variation in policy would yield better results. Potential benefits from an experiment are much less obvious when the object of the study is to improve basic knowledge about behavioral parameters that may be important across a range of policies, but that have no clear implications for any single policy decision. The Health Insurance Experiment improved our knowledge about the price sensitivity of demand for medical services in a way that no nonexperimental study has been able to match. The potential value of this kind of knowledge is almost impossible to measure, however. Not surprisingly, narrow policy experiments are now much more common than social experiments aimed at improving our basic knowledge.

When an experiment looks unpromising, for whatever reason, a different research strategy should be adopted. But in comparison with the evaluation alternatives available, an experiment often represents the best combination of reliability, practicality, and cost-effectiveness. Many experiments have produced treatment-effect estimates that are widely believed to be reliable measures of the average outcome difference caused by a program. The confidence of policymakers in simple experimental estimates rests on a solid foundation: random assignment offers a more credible basis for inference than the assumptions needed in most nonexperimental analyses.

■ *I gratefully acknowledge the comments and suggestions of Alan Auerbach, David Greenberg, Judith Gueron, Lynn Karoly, Robert Moffitt, Philip Robins, Carl Shapiro, and Timothy Taylor. The views expressed are solely my own and should not be attributed to any of these people or to the Brookings Institution.*

References

Aigner, Dennis J., "The Residential Electricity Time-of-use Pricing Experiments: What Have We Learned?" In Hausman, Jerry A., and David A. Wise, eds., *Social Experimentation*. Chicago: University of Chicago Press, 1985, pp. 11–48.

Ashenfelter, Orley, and David Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, November 1985, *67*:4, 648–60.

Barnow, Burt S., "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources*, Spring 1987, *22*:2, 157–93.

Barnow, Burt S., Glen C. Cain, and Arthur S. Goldberger, "Issues in the Analysis of Selectivity Bias." In Stromsdorfer, Ernst W., and George Farkas, eds., *Evaluation Studies Review Annual*. Vol. 5. Beverly Hills, Calif.: Sage Publications, 1980, pp. 42–59.

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, and Fred Doolittle, *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. Bethesda, Md.: Abt Associates, January 1993.

Bradbury, Katharine L., and Anthony Downs, eds., *Do Housing Allowances Work?* Washington, D.C.: Brookings Institution, 1981.

Brook, Robert H., John E. Ware, Jr., William H. Rogers, Emmett B. Keeler, and others, "Does Free Care Improve Adults' Health?," *New England Journal of Medicine*, December 8, 1983, *309*:23, 1426–34.

Burtless, Gary, "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment," *Industrial and Labor Relations Review*, October 1985, *39*:1, 105–14.

Burtless, Gary, and Jerry A. Hausman, "The Effect of Taxation on Labor Supply: Evaluating the Gary NIT Experiment," *Journal of Political Economy*, December 1978, *86*:6, 1103–30.

Burtless, Gary, and Larry L. Orr, "Are Classical Experiments Needed for Manpower Policy?," *Journal of Human Resources*, Fall 1986, *21*:4, 606–39.

Campbell, Donald T., and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

Caves, Douglas W., and Lauritis R. Christensen, "Econometric Analysis of Residential Time-of-use Electricity Pricing Experiments," *Journal of Econometrics*, December 1980, *14*:3, 287–306.

Fisher, R. A., *Statistical Methods for Research Workers*. 2nd ed. London: Oliver and Boyd, 1928.

Fraker, Thomas, and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, Spring 1987, *22*:2, 194–227.

Garfinkel, Irwin, Charles F. Manski, and Charles Michalopoulos, "Micro Experiments and Macro Effects." In Manski, Charles F., and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press, 1992, pp. 253–76.

Greenberg, David, and Mark Shroder, *Digest of the Social Experiments*. Madison, Wis.: Institute for Research on Poverty, University of Wisconsin, 1991.

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

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

Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annuals of Economic and Social Measurement*, Fall 1976, *5*:4, 475–92.

Heckman, James J., "Randomization and Social Policy Evaluation." In Manski, Charles F., and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press, 1992, pp. 201–30.

Heckman, James J., and V. Joseph Hotz, "Choosing among Alternative Nonexperimental Methods for Estimating Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, December 1992, *84*:408, 262–80.

Heckman, James J., and Richard Robb, "Alternative Methods for Evaluating the Impact of Interventions." In Heckman, James J., and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press, 1985, pp. 156–245.

Keeley, Michael C., Philip K. Robins, Robert G. Spiegelman, and Richard W. West, "The Estimation of Labor Supply Models using Experimental Data," *American Economic Review*, December 1978, *68*:5, 873–87.

LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, September 1986, *76*:4, 604–20.

Leamer, Edward E., "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, *73*:1, 31–43.

Levitan, Sar A., *Evaluation of Federal Social Programs: An Uncertain Impact*. Washington, D.C.: Center for Social Policy Studies, George Washington University, June 1992.

Maddala, G. S., and Lung-fei Lee, "Recursive Models with Qualitative Endogenous Variables," *Annals of Economic and Social Measurement*, Fall 1976, *5*:4, 525–45.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, and others, "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, June 1987, *77*:3, 251–77.

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration*. New York: Manpower Demonstration Research Corporation, 1980.

Masters, Stanley, et al., "Jobs Tax Credits: The Report of the Wage Bill Subsidy Research Project, Phase II," mimeo, Madison, Wis., Wisconsin Department of Health and Social Services and Institute for Research on Poverty, University of Wisconsin, 1982.

Moffitt, Robert, "Evaluation Methods for Program Entry Effects." In Manski, Charles F., and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press, 1992, pp. 231–52.

Munnell, Alicia H., ed., *Lessons from the Income Maintenance Experiments*. Boston: Federal Reserve Bank of Boston, 1987.

Rivlin, Alice M., "How Can Experiments Be More Useful?," *American Economic Review*, Papers and Proceedings, May 1974, *64*:2, 346–54.

Rivlin, Alice M., and T. Michael Timpane, eds., *Ethical and Legal Issues of Social Experimentation*. Washington, D.C.: Brookings Institution, 1975.

Stafford, Frank P., "A Decision Theoretic Approach to the Evaluation of Training Programs." In Block, Farrell E., ed., *Research in Labor Economics: Evaluating Manpower Training Programs*. Greenwich, Conn.: JAI Press, 1979, pp. 9–35.

Struyk, Raymond J., and Marc Bendick, Jr., eds., *Housing Vouchers for the Poor: Lessons from a National Experiment*. Washington, D.C.: Urban Institute, 1981.