Chapter 1

Precedents and Prospects for Randomized Experiments

When families move to low-poverty neighborhoods, their teenage children are less likely to commit crimes (Ludwig, Hirschfield, and Duncan 2001). Couples therapy and family therapy are equally effective at improving marital relationships (Shadish et al. 1995). Increasing welfare benefit amounts by 10 percent discourages 1 percent of low-income parents from working (Burtless 1987). Each of these statements answers a question about the effect, or impact, of a social policy or intervention on people's behavior. Does helping low-income families move to low-poverty neighborhoods affect their children's development? Does couples counseling bring more benefits than family therapy? What proportion of low-income parents would stop working if welfare benefits were increased by a certain amount?

These conclusions, like many others of importance to individuals and society as a whole, are based on studies that used a powerful research paradigm known as random assignment. In this book, such studies are referred to as randomized experiments or, simply, experiments (in medicine and other disciplines they are often called randomized trials or clinical trials). In randomized experiments, the units of analysis—usually individuals, but sometimes clusters of individuals—are assigned randomly either to a group that is exposed to the intervention or treatment being studied or to a control group that is not exposed to it. Because the groups do not differ systematically from one another at the outset of the experiment, any differences between them that subsequently arise can be attributed to the intervention or treatment rather than to preexisting differences between the groups. Random assignment also provides a means of rigorously determining the likelihood that subsequent differences could have occurred by chance rather than because of differences in treatment assignment.

Consider how researchers used random assignment in the studies cited in the opening paragraph. In the study of neighborhoods and juvenile crime, families living in public housing developments, all of them situated in high-poverty neighborhoods, were randomly assigned to three groups. Two groups received one of two types of housing voucher: one that could be used toward the cost of private housing provided that the family moved to a low-poverty neighborhood (this was called the conditional-voucher group), or another that could be used to pay for private housing in any neighborhood (the unconditional-voucher group). The third group received no private housing vouchers (the no-voucher group).

Random assignment ensured that the group a family was assigned to depended neither on the family's observable characteristics, such as the parents' age and race or the number of children in the family, nor on characteristics that would be difficult or impossible for a researcher to observe, such as the parents' motivation, ability, and preferences about where to live. In other words, because all the families in the study had an equal chance of being assigned to each of the three groups, the only systematic difference among them was the type of voucher (or no voucher) received. Consequently, when the researchers found a larger proportion of moves to low-poverty neighborhoods and a lower incidence of juvenile crime in the conditional-voucher group than in the other two groups, they were confident about attributing these changes to that single systematic difference.

By itself, however, random assignment would not have allowed these researchers to conclude that moving to a low-poverty neighborhood reduced criminal activity among children. The conditional vouchers affected many other outcomes that might have reduced juvenile crime. For example, they increased work and reduced welfare receipt among parents, and they also increased family income. To find out whether juvenile crime stemmed from living in a low-poverty neighborhood, the researchers went beyond random assignment in two ways: They relied on a strong theory linking neighborhoods to crime, and they used a statistical technique called instrumental variables to identify the effect on juvenile crime of a family's moving out of public housing to a low-poverty neighborhood.

The other two results mentioned in the opening paragraph likewise came from studies that built on random assignment. In the study exploring the effectiveness of psychotherapy in helping couples resolve their problems, the researchers looked at seventy-one random-assignment studies of two types. In one type, couples were randomly assigned to receive marital therapy or no therapy at all. In the other type, couples were randomly assigned to receive family therapy or no therapy at all.

Although the two types of therapy were never compared in a single experiment, the researchers were able to conduct cross-study comparisons using meta-analysis (Glass 1976), a statistical technique for systematically synthesizing quantitative results from multiple primary studies on the same topic. Though not as rigorous as a randomized experiment comparing the two types of therapy directly, this meta-analysis provided an immediate plausible answer to a question that had not yet been addressed using random assignment.

The study of the effect of welfare benefit levels on employment looked at a diverse low-income sample that included families who were not receiving public assistance. After being randomly assigned to one of several different welfare levels, each family was informed of the benefit amount for which it was eligible. A comparison of parental employment rates over a three-year period revealed that parents in families that were eligible for larger benefit amounts were less likely to work than were parents in families that were eligible for smaller amounts. The random-assignment design allowed the researchers to conclude that higher benefit levels caused parents to work less, but the design could not on its own indicate what proportion of parents would be moved to stop working if they were eligible for benefit levels not directly tested in the experiment. Using regression analysis to relate families' responses to the treatments they were assigned, the researchers extrapolated estimates of this proportion to unexamined benefit levels from the results of several different experiments.

Randomized experiments were a rarity in the social sciences until the second half of the twentieth century, but since then they have rapidly become a common method for testing the effects of policies, programs, and interventions in a wide range of areas, from prescription medicines to electricity prices. Written for a diverse audience of social scientists, the present book makes the case for enhancing the scope and relevance of social research by combining randomized experiments with nonexperimental statistical methods that leverage the power of random assignment. This chapter provides context for the rest of the book by describing how and why researchers in a variety of disciplines have gradually shifted from nonstatistical to statistical approaches and from nonexperimental to experimental methods. It also discusses the kinds of questions that are difficult to answer by means of randomized experimentation alone and explores ways in which nonexperimental statistical techniques can complement experimental designs to provide more or better answers than can either approach on its own. The authors of each of the four following chapters use detailed examples from their areas of research—welfare, education, and employment policy—to present four nonexperimental statistical approaches to maximizing the knowledge generated by experiments in the social sciences.

The Evolution of Impact Analysis

Ideas, individuals, and institutions all helped shape the evolution of statistics in science, especially statistical inference—which is the process of determining what data say about the world that generated them. In most disciplines, statistical inference did not become the norm until the twentieth century. Nowadays, not only scientific reports but popular-press accounts of discoveries in a wide range of fields mention statistical considerations, from significance levels in clinical medicine to margins of error in opinion research. To shed light on impact analysis today, this section presents a selective review of its historical development.[1]

Nineteenth-Century Approaches

> Despite historical roots that extend to the seventeenth century, the widespread use of systematic, data-based evaluations is a relatively modern development. The application of social research methods to evaluation coincides with the growth and refinement of the methods themselves, as well as with ideological, political, and demographic changes that have occurred during this century.
>
> —Peter H. Rossi and Howard E. Freeman (1993, 9–10)

One reason why statistical inference was less commonly used in nineteenth-century social science than today was the widely followed school of thought called determinism. In the deterministic view, uncertainty merely reflects human ignorance of the sequence of events that cause each subsequent event. Even in analyses of games of chance, where probability was perhaps most frequently applied in the nineteenth century, it was believed that every roll of a die or flip of a coin has a predetermined, if unknown, outcome.

Determinism led researchers to use statistical concepts such as the average to measure what they thought were fixed, knowable constants. For example, in estimating planetary orbits, astronomers took the average of calculations based on different observations of the planets because they saw each observation as providing information on constants implied by Isaac Newton's theories (Stigler 1999). They believed that different observations provided different information solely because errors had been made in those measurements.

This belief paved the way for major advances in the natural sciences, but it contributed little to the utility of statistics in the social sciences. Social scientists found that statistical aggregates describing humanity were stable over time and across places, and these social aggregates became the object of study for researchers such as the statistician Adolphe Quetelet (Gigerenzer et al. 1989). But social aggregates lacked the theoretical foundation that Newtonian theory provided astronomers, and they were therefore deemed inadequate for describing human behavior and outcomes.

Quetelet's 1835 study of conviction rates in French courts is a good example of the challenges faced by nineteenth-century social scientists. Quetelet compared criminal cases with respect to a number of factors hypothesized to influence conviction, including whether the crime had been committed against property or a person and whether the defendant was male or female and was under or over thirty years of age. Such comparisons led him to conclude that well-educated females over the age of thirty were least likely to be convicted (Stigler 1986).

Quetelet's approach met with criticism (Stigler 1986). Baron de Keverberg, who was advising the state on the matter, argued that data had to be separated into every possible category of characteristics that might be related to the outcome of interest. Not doing so would mean comparing groups that were otherwise not equivalent and perhaps drawing unwarranted conclusions. The economist Jean-Baptiste Say likewise argued that averaging measurements across people resulted in a meaningless jumble of disparate factors (Gigerenzer et al. 1989). In medicine, physicians argued that averages were useless because each patient was unique and each disease was specific to an individual (Gigerenzer et al. 1989). Perhaps for this reason, "As late as 1950, most physicians still thought of statistics as a public health domain largely concerned with records of death and sickness" (Marks 1997).

The German physicist and philosopher Gustav Fechner felt that his method of testing his theory of psychophysics (the science concerned with quantitative relations between sensations and the stimuli producing them) provided him with the theoretical justification for averaging observations that the social sciences had hitherto lacked (Stigler 1999). Specifically, Fechner developed the psychological measure known as the "just noticeable difference," which is the smallest observable difference between two stimuli. To measure the just noticeable difference between two weights, for example, Fechner had subjects pick up pairs of weights more and less similar in weight and observed the proportion of the time they correctly identified the heavier weight in each pair. For differences smaller than the just noticeable difference, they would be forced to guess, in which case they would correctly identify the heavier weight only half the time. Fechner's theory provided an objective, theoretically derived constant, .5, to which he could compare the proportion of times the heavier weight was correctly identified. Even more important for the present purpose, Fechner could manipulate environmental conditions such as the amount of weight that was lifted.

Although experimentation provided the phenomena to be measured, the deterministic view led researchers to believe that probability-based statistical inference was unnecessary. In other words, the experimental method was considered at odds with the use of statistical inference. Because uncertainty was thought to be due to measurement error, researchers focused on removing that uncertainty by increasing the quality and rigor of their studies. The physiologist Emil du Bois-Raymond thought that natural processes are constant over time and across individuals and therefore can be measured adequately by observing a small number of subjects in a well-designed, well-run study (Coleman 1987). The psychologist Wilhelm Wundt similarly argued that all individuals responded to psychological stimuli in the same way, justifying the study of the effects of a stimulus in a mere handful of subjects (Danziger 1987).

Nonexperimental Statistical Approaches

> Do slums make slum people or do slum people make the slums? Will changing the living conditions significantly change the social behavior of the people affected?
>
> —A. S. Stephan (quoted in Rossi and Freeman 1993, 12)

> The method of least squares is the automobile of modern statistical analysis: despite its limitations, occasional accidents, and incidental pollution, this method and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all.
>
> —Stephen M. Stigler (1999, 320)

As innovative as Fechner's method was, its utility did not immediately benefit social scientific fields, in which environmental conditions could not be manipulated as readily as in psychophysics (although economists have recently found ways to manipulate conditions in classroom settings to test basic economic theories; see Smith 1994). For example, Quetelet, in his investigations of who would be convicted of criminal behavior, could measure conviction rates for different groups, but he could not manipulate the judicial system to test his hypotheses about the factors that affected conviction rates.

What were the scientific and political changes that fostered the use of statistical inference in the social sciences? An important breakthrough for social scientists came in the 1890s with G. Udny Yule's theoretical reinterpretation of least squares, a method that the French mathematician Adrien-Marie Legendre had introduced to find the average of a series of observations that differed because of measurement error (Stigler 1986).[2] The technique was named "least squares" because it minimizes the sum of the squared deviations between the observed data and the estimated mean. Building on several decades of research by social scientists and statisticians, Yule realized that the method could also be used to estimate the parameters of a linear relationship among any number of factors. Further, he argued that the multivariate linear relationship provided a means of investigating how one factor varied with a second factor, with all other factors held constant. In this new framework, least squares provided social scientists with three essential ingredients of statistical inference: parameters that could be subjected to estimation and statistical inference (the parameters resulting from least squares), a way to adjust for all other observed differences across individuals (including them in the least-squares regression), and a technology for estimating a multivariate relationship (the least-squares method itself).

To appreciate the implications of these developments, imagine that it is 1900 and you are investigating the link between neighborhoods' economic prosperity and rates of juvenile crime. On the basis of your discovery that children in low-poverty neighborhoods commit fewer crimes than other children, you conclude that a family's living in a high-poverty neighborhood makes the children in that family more likely to commit crimes. Your critics object that families and children in the two types of neighborhoods may differ in other respects as well. For instance, the average family in low-poverty neighborhoods might be better educated or have different attitudes about criminal activity than the average family in high-poverty neighborhoods, which might affect children's chances of getting into trouble with the law. A decade earlier, your critics probably would have insisted, as did Baron de Keverberg, that you separate the data into every possible combination of characteristics that might be related to juvenile crime and show that only the type of neighborhood can explain the difference in crime rates. But thanks to Yule, you can now perform a multivariate regression analysis using as the dependent variable the likelihood that a particular child has committed a crime, and as the explanatory variables factors thought to affect juvenile crime such as whether the family lives in a high-poverty neighborhood, the family's income, the parents' education level, and the child's age. The estimated coefficient for the factor specifying whether a family lives in a high-poverty neighborhood gives you an estimate of the effect of the type of neighborhood on juvenile crime, independent of the other characteristics included in the regression.

In parallel with the advances in statistical theory sketched here, the shift toward statistical inference gained impetus when university researchers in medicine began calling for systematic scientific review of medical interventions such as drugs. The prevalent belief at the time that physicians could tell helpful drugs from harmful ones solely on the basis of their own observation and intuition allowed drug manufacturers to aggressively market products that had not been clinically proven (Marks 1997). Unfortunately, the rigorous assessment of drugs was not the norm in 1938, when 106 people died because a drug manufacturer unwittingly used a known toxin to buffer a new medicine. Although the U.S. Food and Drug Administration (FDA) was granted the authority to regulate drugs for effectiveness and safety in 1938, it still was unclear how benefits and hazards were to be measured. For example, when the FDA was considering whether to approve sulfapyridine as a treatment for pneumonia, although it knew there was evidence of negative side effects as well as benefits, it had no rigorous information on the size of either. The need for more—and more rigorous—evidence fueled the use of statistical inference.

In social policy as in medicine, the use of statistical inference grew with the need to make decisions about government policy. Because the social programs created as part of the New Deal of the 1930s were never evaluated, their descendants in the War on Poverty of the 1960s were designed without hard evidence about the earlier programs' effectiveness (Rossi and Williams 1972). The legislation that launched the War on Poverty mandated and funded evaluations of the policies it established. Thus, evaluation research took off during the twentieth century at least partly because, as publicly financed programs expanded, the government demanded more and better information as to whether it was investing its billions of dollars wisely.

Despite these theoretical advances and the desire for more rigorous evidence, social researchers soon came to realize that the effects of an intervention—estimated by comparing the differences in outcomes between two or more groups—reflect the influence not only of the intervention under study but also of other factors that could affect group membership. Failure to take account of these other factors in the statistical model leaves impact estimates at risk of deviating from true impacts because of selection bias—systematic differences between the groups in the study other than that arising from the intervention. A controversial analysis of the effects of parental alcoholism on children conducted by Karl Pearson in 1910 illustrates the problem. On the basis of his finding that, on average, children of alcoholic parents had about the same levels of intelligence and health as other children, Pearson concluded that parental alcoholism does not harm children (Stigler 1999). In support of this conclusion, he argued that the alcoholic and nonalcoholic parents in his analysis were comparable because the two groups' average hourly wages were similar. But prominent economists of the day challenged Pearson's study, arguing that the two groups probably differed in other ways that could not be ruled out. To them, the fact that the two groups earned the same average amount suggested that the alcoholic parents actually had higher average ability that had been suppressed by alcoholism, or else they would not have been able to "keep up" with the nonalcoholic parents. Because Pearson's critics believed that ability was partially inherited, they concluded that the children of alcoholic parents in the study probably would have performed at a higher level but for their parents' alcoholism.

Over time, researchers developed increasingly sophisticated methods to reduce selection bias. In social policy, most of these methods involve comparing a group that is exposed to the treatment under study, referred to in this book as the program group (it is sometimes called the experimental group or the treatment group), with a comparison group that is not. The simplest approach to reducing selection bias relies on multivariate regression methods to control for as many observed differences between the program group and the comparison group as possible. No matter how many such factors are included, however, this strategy is not likely by itself to overcome selection bias arising from unobserved factors.

An alternative approach to reducing selection bias focuses on choosing a comparison group that matches the program group with respect to as many observed characteristics as possible. If two groups have the same education, income, and employment history, it is hoped, they might also have the same motivation, attitudes, and other attributes that are difficult to measure. For example, to understand the effects of housing conditions on a range of outcomes, one study compared three hundred families that were placed in improved public housing with three hundred families with similar social and demographic characteristics that were chosen from a waiting list (Wilner et al. 1962, as described in Gilbert, Light, and Mosteller 1975).

Special cases of this method use comparison groups composed of people who applied for a program but subsequently withdrew (withdrawals), were accepted into a program but did not participate in it (no-shows), or were screened out of a program for failing to meet eligibility criteria (screen-outs). Unfortunately, the strategy of choosing a comparison group on the basis of observed characteristics stumbles on the same problem as multivariate regression: Groups that are similar with respect to observed characteristics might be dissimilar with respect to unobserved characteristics that could influence both whether the treatment is received and subsequent outcomes.

Another method of reducing selection bias relies on comparing the program group to itself at two or more points in time. For example, medical researchers first realized that insulin might be the key to fighting diabetes by observing dramatic improvements in the health of diabetics who took insulin (Marks 1997). Similarly, evaluators have sometimes measured the effects of job training services by examining changes in trainees' earnings over time. Michael E. Borus (1964) used this method to study a 1960s federal program that trained workers who had lost their jobs because of technological change. Comparing the program group with itself holds constant all unobserved characteristics that do not change over time, but it opens the door to other biases. For example, William Shadish, Thomas D. Cook, and Donald T. Campbell (2002) pointed to potential biases from maturation and history. Maturation bias arises when outcomes would gradually change even without intervention, such as when ill people either get better or die. History bias arises when other policy or societal changes occur during the period of observation. For example, the economic growth of the 1990s made it difficult to estimate the effects of the federal welfare reform legislation passed in 1996.

Rather than focusing on how to select or construct a comparison group, some researchers devised statistical means to control for differences in unobserved characteristics between the program and comparison groups that give rise to selection bias (Barnow, Cain, and Goldberger 1980). Given some assumptions, this bias can be eliminated if the researcher has all relevant information regarding how a person was selected to receive a treatment (Cain 1975). Suppose, for example, that applicants to an employment training program are accepted solely on the basis of their earnings in the prior month. Although applicants with earnings just above the threshold receive no training from the program and applicants with earnings just below the threshold are accepted into the program, they should otherwise be quite similar. If the below-threshold applicants have very different earnings after they go through the program than the above-threshold applicants, it is plausible to conclude that the program affected earnings. This approach, first proposed by Donald L. Thistlewaite and Donald T. Campbell (1960), is sometimes called regression discontinuity. Unfortunately, participation in many programs is determined by factors that cannot be observed by researchers.

Selection bias can also be eliminated through the use of statistical assumptions about how observed and unobserved factors affect which treatment a person receives. The method is robust, however, only if a factor that is related to the selection of treatment but not to unobserved determinants of the outcome is identified, which can be very difficult.[3]

Despite their ingenuity, these efforts to overcome selection bias in studies of training, housing assistance, education, youth programs, and health care have met with considerable skepticism because even in the largest studies with the most sophisticated designs, different evaluators have come to different conclusions depending on the comparison groups and statistical methods selected (Barnow 1987; Kennedy 1988; Gilbert, Light, and Mosteller 1975; McDill, McDill, and Sprehe 1972; Glenman 1972). The essential problem is that selection bias can arise from variation in unobserved characteristics, making the bias difficult to detect and eliminate.

Random Assignment

> Let us take out of the Hospitals, out of the Camps, or from elsewhere 200, or 500, poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one halfe of them may fall to my share, and the other to yours; . . . we shall see how many Funerals both of us shall have.
>
> —John Baptista van Helmont (1662)

> Random assignment designs are a bit like the nectar of the gods; once you've had a taste of the pure stuff it is hard to settle for the flawed alternatives.
>
> —Robinson G. Hollister and Jennifer Hill (1995, 134)

There is one method that, when well implemented, is guaranteed to eliminate selection bias: random assignment of individuals or groups to treatments. In a typical experiment, half the study members are randomly assigned to the program group, and the other half are randomly assigned to the control group, a term reserved for the special case in which membership in the comparison group is randomly determined. Because random assignment ensures that unobserved factors such as motivation and ability are distributed about equally in the two groups, any differences that later emerge between them can be confidently attributed to differences in the groups to which they were assigned rather than to preexisting differences.

Widespread use of random assignment is relatively new in social policy research, but the idea has been around for much longer. The quotation at the beginning of this section shows that an understanding of the logic behind randomization—if not sensitivity to the rights of the sick and the poor—dates back at least to the seventeenth century. In an early application of the method, Charles S. Peirce and Joseph Jastrow (1884/1980) drew playing cards to determine the order of stimulus presentation in a study of Fechner's just noticeable difference.

Much of the foundation of the modern use of random assignment, however, can be traced back to Ronald A. Fisher's work in the early twentieth century. On becoming chief statistician at Rothamsted Experimental Station in England in 1919, Fisher found that data on crop yields had been collected for more than sixty years at the station but had not been rigorously analyzed (Gigerenzer et al. 1989). The Rothamsted researchers faced two problems discussed in the previous section: They did not know how to separate the effects of fertilizers from the effects of other factors influencing crop production, and they did not know how to assess the likelihood that fertilizers made a difference because crop yields varied even when the same fertilizer was used. Fisher realized that random assignment provided a way to deal with both problems. By randomly choosing which fields would receive fertilizers, he was able to convincingly argue that systematic differences in crop production between fertilized and unfertilized fields were due to the fertilizer. By repeating random assignment across many fields, he was able to compare the variation in crop production within each group (the unfertilized fields and the fertilized fields) to the difference in production between the two groups and assess the likelihood that the difference between groups was due to random variation—and, by extension, the likelihood that the choice of fertilizer made a real difference in production. Fisher (1925) helped other researchers understand how to use random assignment by providing them with information on how to implement random-assignment studies and how to conduct formal statistical inference with respect to their results.

Random assignment rules out alternative explanations for observed results so effectively that it is often called the "gold standard" in evaluation research, but this was not understood when Fisher first developed the method and began to promote it to his colleagues in agricultural research. Indeed, it was not until well after World War II that use of the approach became widespread in medical research (Marks 1997), and several more decades were to pass before random assignment would be used with any frequency in social policy research.

By one estimate, more than 350,000 randomized experiments have been conducted in medical science over the last fifty years (Cochrane Collaboration 2002). According to *The Digest of Social Experiments*, more than 800,000 people were randomly assigned in two hundred twenty studies of new or existing social policies between 1962 and 1997 (Greenberg and Shroder 1997). A number of other studies have involved the random assignment of larger entities, such as schools or municipalities (Boruch and Foley 2000). And within the last three years, the Administration for Children and Families within the U.S. Department of Health and Human Services alone has launched a number of major projects encompassing several dozen random-assignment studies to investigate the effects of policies designed to affect the employment, family structure, child-care choices, and marital relationships of low-income people.

The following are a few of the many social policy areas in which experimental studies have been performed:

Child health, nutrition, and education. In 1997, the government of Mexico launched a major experiment to measure the effects of substantial payments designed to encourage poor rural parents to send their children to school more regularly and over a longer period and to improve their children's health care and nutrition (Teruel and Davis 2000). This experiment randomized 320 rural communities to the program group and 186 rural communities to the control group.

Crime and juvenile delinquency. Random-assignment studies have been used extensively to study policies aimed at reducing juvenile delinquency and crime (Lipsey 1988; Sherman 1988). One of the most successful and oft-cited studies tested a policy of requiring alleged assailants in domestic assault cases to spend a night in jail.

Early child development. Random assignment has been used to study interventions to help children. For example, a number of experiments have investigated the effects of improvements in day care (Schweinhart and Weikart 1993; Ramey, Yeates, and Short 1984; Love et al. 2002) and of Head Start (Bell et al. 2003).

Education. Until the U.S. Department of Education made the use of random assignment a priority in evaluating new education approaches (Coalition for Evidence-Based Policy 2003), experimental designs were used only infrequently in education research. Earlier applications tested the effects of vouchers allowing low-income children to attend private schools (Mayer et al. 2002); of a high school reform in which students' learning is organized around a career theme (Kemple and Snipes 2000); and of smaller class sizes at the elementary school level (Mosteller, Light, and Sachs 1996).

Electricity pricing. Faced with increasing demand and limited supply, electricity policy analysts in the 1970s recommended making electricity more expensive during peak periods to induce consumers to shift their electricity use to off-peak periods. This approach, called "time-of-use pricing," was studied in a series of randomized experiments (Aigner 1985).

Health services. To investigate the link between health-care copayments and use of health-care services, one experiment randomly assigned different people to pay different proportions of their health-care expenditures (Newhouse 1996; Orr 1999). Random assignment has also been used to study the efficacy of mental health treatments (Ciarlo and Windle 1988).

Housing assistance. In the 1970s, the effects of direct cash low-income housing assistance in Pittsburgh and Phoenix were examined in a randomized experiment (Kennedy 1988). More recently, the U.S. Department of Housing and Urban Development launched an experiment investigating the effects of housing vouchers that require users to move to low-poverty neighborhoods (Goering et al. 1999; Orr et al. 2003).

Income maintenance. Among the earliest studies in the social sciences to use random assignment were the negative income tax experiments conducted in the 1960s and 1970s in Denver, Seattle, New Jersey, and Gary, Indiana. In these studies, different families were randomly assigned to different levels of income that would be guaranteed if the family had no other sources of income and to different "tax rates" that determined how much of the guaranteed level would be lost when the family's income from other sources increased (Munnell 1987).

Job training. A recent meta-analysis of voluntary education and training programs for low-skill adult workers and adolescents cites ten studies that used random assignment, most of them conducted after 1983 (for a list, see Greenberg, Michalopoulos, and Robins 2003). One of the best-known recent studies is a large-scale multisite experiment that measured the impacts of a national program to provide job training in a residential setting to adolescents and young adults at risk of economic failure and criminal behavior (Burghardt et al. 2001).

Unemployment insurance. Because many unemployed workers return to work only after their unemployment insurance benefits have been used up, random assignment has been used to explore the effects of financial incentives to return to work (Bloom et al. 1999; Woodbury and Spiegelman 1987; Spiegelman, O'Leary, and Kline 1992).

Welfare-to-work programs. Since 1980, dozens of experiments have been conducted to test programs designed to help welfare recipients move into employment (Gueron and Pauly 1991; Greenberg and Wiseman 1992; Bloom and Michalopoulos 2001).

As already discussed, random assignment has become more common in social policy research partly because nonexperimental statistical methods do not eliminate selection bias convincingly and do not provide widely persuasive estimates of policy effects. Another reason is that the sophisticated methods devised by statisticians and econometricians to reduce selection bias are often difficult for policy practitioners to understand—and easy for critics to challenge.

The logic of random assignment, in contrast, is straightforward to grasp and explain. At its most basic, a random-assignment analysis involves making simple comparisons between the average outcomes for the program group and the average outcomes for the control group. The simplicity and rigor of randomized experimentation have made it influential in policymaking. For example, experimental studies of welfare-to-work programs in the 1980s have been credited with influencing the debate that led to the Family Support Act of 1988 (Greenberg and Wiseman 1992), which paved the way for the landmark federal welfare reforms of 1996.

Learning More from Randomized Experiments

Despite its considerable strengths, random assignment of individuals or groups to treatments cannot answer all the important questions about what works best for whom and why.[4] For this reason, researchers have devised creative ways to combine experimental data with nonexperimental statistical methods. By doing so, they can capitalize on the strengths of each approach to mitigate the weaknesses of the other. The central premise of this book is that combining experimental and nonexperimental statistical methods is the most promising way to accumulate knowledge that can inform efforts to improve social interventions. This idea, articulated three decades ago by Robert F. Boruch (1975), is only now beginning to gain greater recognition.

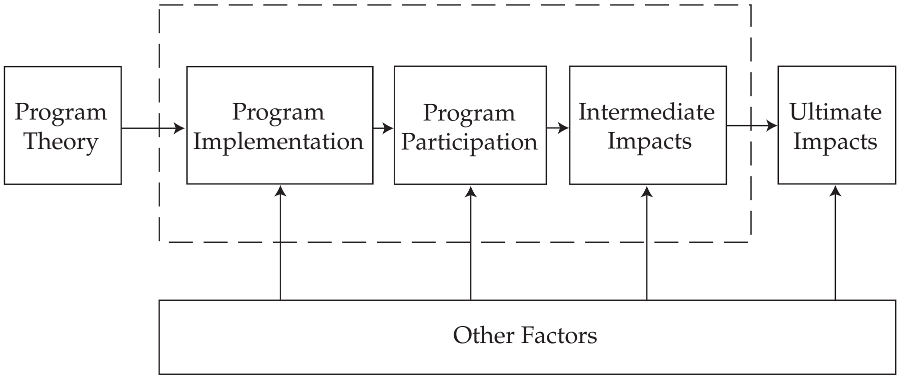
Opening the Black Box

In some cases, random assignment is too blunt an instrument to be used to test theories about what makes a program effective. Budget limitations can also preclude the use of random assignment to compare different versions of a program model. Furthermore, randomized experiments are usually unable to disentangle the contributions of specific program components to the program's overall effects. In other words, they often do not reveal what happens inside the "black box" where their effects unfold. Figure 1.1 displays a model of the black box, which focuses on how program implementation, program participation, and intermediate impacts lead to the ultimate impacts of an intervention.

"Program implementation" represents the components of the program and the way in which they are carried out. For example, a mandatory welfare-to-work program might include features such as financial assistance for education and certain required activities, and might sanction those who do not engage in them. More generally, the effectiveness of any program depends on whether staff are enthusiastic about it, whether potential participants know about and understand it, and the management practices of the people running it.

A program's overall effects will depend on "program participation": how many people participate in it and how intensively they do so. For example, in a study of an antihypertensive drug, the drug's effects can be expected to increase as the percentage of patients who comply with the prescribed treatment increases. To help families earn more, a welfare-to-work program must engage parents in job search activities.

Figure 1.1 How Programs Create Impacts



Source: Author's compilation.

Housing vouchers change the lives of more families both as the proportion of families using the vouchers increases and as the length of time that vouchers are used increases.

Participation in a program can have a variety of impacts on intermediate outcomes. For example, an antihypertensive drug should lower blood pressure. A welfare-to-work program should increase employment and reduce welfare use. A housing voucher program might benefit families not only by helping them move to lower-poverty neighborhoods per se but by bringing their children into closer contact with middle-class peers, who might be less likely to commit crimes, engage in premarital sex, and use illegal drugs.

Finally, program impacts on intermediate outcomes can lead to a range of impacts on ultimate outcomes. For example, the ultimate goal of antihypertensive medication is to help people live longer, healthier lives. Most welfare programs aim to improve the economic well-being of families and children. The ultimate objective of a housing voucher program might be to help families move into the middle class and to reap the financial and nonfinancial benefits associated with that status. In many cases, the distinction between intermediate and ultimate outcomes is not cut-and-dried. For example, increasing employment and reducing welfare use might be the ultimate objective of a policymaker rather than an intermediate objective. In other words, it might be viewed as an end in itself rather than a means to some other end.

Random assignment studies are typically designed to reveal whether a treatment affects intermediate and ultimate outcomes, not to elucidate the role of implementation and program participation or the link between intermediate and ultimate outcomes. So how can going beyond random assignment using nonexperimental statistical methods help researchers open the black box of experiments to explore these issues?

Measuring Implementation Effects

One important question that randomized experiments have difficulty addressing directly is how various components of a program contribute to its effects. For example, if a three-component program reduces teen pregnancy rates, how can one infer whether and to what extent each component contributed to the program's overall success? To answer such questions, evaluators often turn to research synthesis, which is an attempt to draw lessons by looking across studies. A frequently applied example of this approach, narrative synthesis, entails qualitatively comparing results across sites or studies. For example, the final report in the six-county evaluation of a California welfare-to-work program called Greater Avenues to Independence (GAIN) found the largest effects in Riverside County. In their effort to synthesize results from the six counties and understand how Riverside differed from the other counties, the authors looked at how more than a dozen site characteristics were correlated with the variation in program impacts across sites. The evidence suggested that the Riverside program's strong emphasis on getting participants into employment set it apart from the others (Riccio, Friedlander, and Freedman 1994).

Narrative synthesis often leads to several possible conclusions. The National Evaluation of Welfare-to-Work Strategies (NEWWS) used random assignment to study the effects of eleven programs in seven sites, including Portland, Oregon, where the program was outstandingly effective. Evaluators were left to wonder why Portland's effects stand out. Was this attributable to the program's services, the message conveyed by its staff that participants should accept only "good" jobs, the strong local economy, exemptions given to welfare recipients who staff thought would not benefit from the program, its vigorous efforts to connect welfare recipients with specific employers, or something else (Hamilton et al. 2001)? Although both GAIN and NEWWS are excellent examples of the use of random assignment, the experimental design could take the researchers only so far in understanding why the policies under study were more or less effective.

These examples illustrate that simple experiments have a limited ability to separate policy from practice. More complex designs can help. The most common approach is to randomly assign entities to two or more interventions. In the evaluations of the Minnesota Family Investment Program (MFIP) and Canada's Self-Sufficiency Project (SSP), for example, families were assigned at random to three groups: a group that received financial incentives to work, a group that received job search assistance as well as the financial incentives, and a control group (Miller et al. 2000; Michalopoulos et al. 2002). The design was especially useful in MFIP. In that study, incentives alone did not increase employment, but they increased family income because financial payments were made to families that would have worked in the absence of the incentive. In addition, children in the incentives-only group were better off after three years than were children in the control group, providing evidence that increasing family income would help children.

Although three-group random-assignment designs are useful, they provide information on only two policy components. Early social experiments were much more ambitious (Greenberg and Robins 1986). For example, the negative income tax experiments of the 1970s assigned families to fifty-eight different combinations of tax rates and guarantee levels to study how people respond to changes in those policy parameters (Munnell 1987; Greenberg and Robins 1986). Similarly, the direct cash low-income housing assistance experiment randomly assigned people to receive housing subsidies equal to 0, 30, 40, 50, or 60 percent of their rent to understand how strongly people responded to different subsidy rates (Kennedy 1988). Finally, a health-care copayment experiment randomly assigned people to pay 0, 25, 50, or 95 percent of their health-care expenditures up to an annual maximum of $1,000 to allow researchers to extrapolate to other subsidy rates (Newhouse 1996). Exploiting such variation to extrapolate from experimental results to other types of policies was one of the original motivations behind using random assignment (Heckman 1992). But testing many policies at once also has a drawback: Testing more variants of a program with a given sample results in less precise estimates of the effects of any one variation because fewer people are assigned to each variation.

In the absence of random assignment to many different treatments, meta-analysis, another form of research synthesis, can be used to explore why some programs are more effective than others. Meta-analysis is a statistical technique for synthesizing quantitative results. For example, researchers in the United Kingdom conducted a meta-analysis of twenty-four random-assignment evaluations of mandatory welfare-to-work programs in the United States and concluded that job search is associated with larger effects on earnings and welfare receipt, that vocational training is associated with smaller effects on earnings and welfare receipt, and that programs with more white recipients have larger effects than other programs (Ashworth et al. 2001).[5] Obtaining statistically precise estimates of the determinants of program effects using meta-analysis, however, requires dozens of policy experiments (Greenberg et al. 2003).

A third type of research synthesis uses multilevel statistical modeling (which accounts for the grouping of individuals in aggregate units) to explore how the natural variation in program effects across sites in a single experiment or across experiments in a set of studies co-varies with features of the program, its participants, and its environment. This approach is illustrated by Howard Bloom, Carolyn Hill, and James Riccio in chapter 2 of the present volume, "Modeling Cross-Site Experimental Differences to Find Out Why Program Effectiveness Varies." Their analysis focuses on how the earnings effects of welfare-to-work programs vary with respect to how the programs were implemented, the specific services they provided, the characteristics of their participants, and local labor market conditions. This analysis was made possible by the fact that comparable data had been collected for three large-scale multisite randomized evaluations of welfare-to-work programs: the National Evaluation of Welfare to Work Strategies (NEWWS), the California GAIN program described earlier, plus Florida's welfare-to-work program Project Independence (Kemple and Haimson 1994).

Future random-assignment studies could likewise be set up to learn more about how programs achieve their effects. For example, a recently completed study of a vaccine to prevent infection with HIV (human immunodeficiency virus) that randomly assigned people in fifty-nine locations in the United States, Canada, and the Netherlands found that the vaccine worked better for African American and Hispanic sample members than for white sample members (Pollack and Altman 2003). If procedures varied from site to site, the variation could be used to help understand why the vaccine worked better for some people than for others. In fact, procedures in different sites in an experiment could be varied intentionally to assess the effects of different procedures. Even if an experiment is not conducted in dozens of sites, collecting information about each site might help future researchers investigate the role of program practices and inputs. For example, the GAIN, Project Independence, and NEWWS evaluations were not conducted at the same time, but they were conducted by a single research organization that collected comparable data across studies.

Assessing the Role of Participation

Randomized experiments are designed to measure the average effect of a treatment among all the people who are randomly assigned to it, but it is sometimes of interest to estimate the effect among those who actually receive the treatment. Consider a random-assignment study of a new vaccine that is otherwise unavailable. The effect of the intervention among all people who are assigned to receive the vaccine might represent the expected effect of introducing the vaccine in a real-world situation in which some patients do not comply with their treatment. The effect among people who actually receive the vaccine provides an estimate of the effect of the vaccine on compliant patients.

In many cases, an intervention is unlikely to have an effect on people who are not exposed to it. For example, a vaccine against a noncommunicable disease cannot benefit someone who does not receive it. In such cases, the effect among people who receive the treatment can be calculated by dividing the overall average effect by the proportion of program group members who actually received the treatment (Bloom 1984). Suppose the average effect of a welfare-to-work program on earnings is $500 per year, including participants and nonparticipants. If only half of program group members actually participate in the program, then the average effect on earnings per participant is twice as high since they generated the entire effect. The effect of a treatment on those who receive it, sometimes referred to as the effect of the treatment on the treated, is discussed in chapter 3.

It is even possible for an intervention to have effects on people who do not participate. For example, a requirement that welfare recipients attend a job club or face sanctions might increase employment even among people who do not attend the job club if the threat of sanctions encourages them to look for work on their own or to take a job they have already been offered. When an intervention might have effects among those who are assigned to the treatment but do not receive it, determining the effect of treatment on the treated or the effect of treatment assignment on those who do not receive the treatment is not straightforward. If the effect on each of the two groups is roughly constant across individuals within them, however, then the effect for each group can be derived (Peck 2002).

Exploring Causal Pathways

Increasing program participation is rarely the ultimate goal of a treatment, but understanding the link between program participation and other outcomes can inform practitioners' efforts to implement a treatment and policymakers' efforts to design a policy. Likewise, knowing how different outcomes are related to one another, especially whether changing one outcome is likely to cause changes in other outcomes, can help shed light on the question of which intermediate outcomes or mediators should be the target of future treatments and policies.

For example, mediators can help explain the effects on children of programs targeted at parents. To take one recent example, studies of policies that increase family income have been found to benefit school-age children (Morris et al. 2001; Clark-Kauffman, Duncan, and Morris 2003). Higher income might allow parents to purchase goods and services that directly help their children, such as more nutritious food or higher-quality child care. Programs that increase income typically have many effects, among them increasing parents' earnings and hours of work. Extra income might also affect parenting by reducing stress or influencing the likelihood that the parent will marry, which under certain conditions might help their children. Understanding whether extra income per se benefits children—and if so, how much each dollar of income helps—could help policymakers decide whether investments in programs such as the federal Earned Income Tax Credit are worthwhile or whether money should be spent on other objectives, such as improving the quality of child care used by low-income families.

Nonexperimental researchers have developed statistical methods to investigate the role of mediators. An approach that has generated much recent interest is instrumental variables analysis (Angrist and Krueger 2001). A researcher who wants to understand the effects of income on children's well-being could compare the children of wealthier and poorer families, but the comparison would probably be plagued by selection bias: wealthier families might have parents with more extensive social networks or higher levels of motivation, both of which might benefit children independent of income. To explore the relationship between income and children's well-being more rigorously, the researcher could find an instrumental variable, which in this example would be a factor that is correlated with both income and children's well-being but is not correlated with unobserved factors that also might affect these two variables (such as motivation and intelligence). The researcher would then explore the effects of income on children's well-being by comparing outcomes for children whose families have different values of the instrumental variable.

The biggest obstacle to using instrumental variables is finding an instrument. Myriad factors affect not only family income but also most other outcomes of interest. Random assignment, however, can provide the needed instrument. Because assignment to the program and control groups is random, it is not related to parents' intelligence, the extensiveness of their social network, or their motivation. Indeed, it is uncorrelated with every characteristic that exists before or at the time of random assignment.

Instrumental variables analyses of the effects of mediators require at least one instrument for each hypothesized intermediate outcome. In the study of housing vouchers mentioned in the introduction to this chapter (Ludwig, Hirschfield, and Duncan 2001), two intermediate outcomes were assumed to be important: whether the family moved at all and whether the family moved to a lower-poverty neighborhood. The random assignment of families to two program groups provided the two instruments needed. Families given unconditional vouchers were more likely to move out of public housing but were not more likely to move to lower-poverty neighborhoods than did control group families. Comparing the rate of juvenile crime in these two groups therefore indicated how much moving reduced juvenile crime. Families given conditional vouchers, in contrast, were not only more likely to move out of public housing but moved to lower-poverty neighborhoods than did control group families, providing a means of determining how much moving to lower-poverty neighborhoods reduced juvenile crime. Thus, the instrumental variables technique provided the method for calculating the effects of a key mediator.

The link between welfare-to-work programs and children's well-being might be affected by many mediators, and it is unlikely that one evaluation would include enough program groups to provide such a large number of instruments. However, the necessary instruments might be obtained by using data from a number of random-assignment studies. Different studies will have larger or smaller effects on income, employment, hours of work, and other mediators. An instrumental variables analysis essentially compares the variation in the programs' effects on the mediators with the variation in the programs' effects on children's well-being to infer the effects of the mediators on children's well-being.

Although instrumental variables techniques can be extremely useful when combined with random assignment, they do require making some assumptions. One assumption is that if data are pooled across studies to achieve the needed number of instruments, the effects of the mediators on the ultimate outcomes must be the same from place to place. For example, an extra dollar of income and an extra hour of work would have to have the same effect on children in California as it does in Minnesota.

Chapter 3 of the present volume ("Constructing Instrumental Variables from Experimental Data to Explore How Treatments Produce Effects," by Lisa Gennetian, Pamela Morris, Johannes Bos, and Howard Bloom) examines the use of instrumental variables with random assignment experiments to explore the causal paths (linkages among mediating variables) by which programs produce their impacts. The authors describe how the approach works, outline the assumptions that are necessary for it to do so, and illustrate its application to numerous real-world examples. They also consider how to make the approach operational given the realities of how randomized experiments are conducted.

Studying Place-Based Interventions

Most experiments randomly assign individuals to program and control groups because they seek to study the effects of programs on the outcomes and behavior of individuals, couples, or families. There are circumstances, however, in which random assignment of larger entities such as child-care providers, firms, schools, housing developments, or entire communities might be more appropriate. The aforementioned experiment that randomized whole communities in Mexico (Teruel and Davis 2000) is an example of a place-based intervention. Sometimes called group randomization, this method is referred to in this book as cluster randomization or cluster random assignment.

Cluster randomization is appropriate where an intervention is meaningful only at the level of a larger entity or when it is impractical or unacceptable to assign individuals randomly to different treatments. For example, an experimental study of a program that trains child-care providers to teach preschool children who are receiving child-care subsidies to read should probably conduct random assignment at the level of providers rather than individual children. Randomly assigning children within a classroom to the treatment or a control group would be logistically infeasible because control-group children in classrooms with a trained provider could not be prevented from benefiting from the provider's training. And randomly assigning children to a trained provider or an untrained provider might be politically infeasible because it would restrict their families' child-care choices, violating federal regulations governing child-care subsidies. For similar reasons, the effects of public-service advertisements designed to discourage smoking have been studied by randomly assigning everyone who might come across the same advertisements to the same group (Boruch and Foley 2000). If television advertising were being used, an entire viewing area would be randomly assigned to receive the advertising or not to receive it.

Cluster random assignment has also been recommended in cases where outcomes for people exposed to an intervention have the potential to "spill over" and affect outcomes for other people (Harris 1985; Boruch and Foley 2000; Garfinkel, Manski, and Michalopoulos 1992). Vaccinating some children against a communicable disease might protect unvaccinated children from the disease as well. A study that randomly assigned individual children in the same community to receive or not to receive the vaccine would underestimate the vaccine's effects if vaccinating children in the program group indeed lessened the chance that children in the control group would contract the disease. Similarly, a large-scale training program might help trainees find better jobs, but it might also displace others who would have taken those jobs. If the displaced workers are members of the control group in a random-assignment study, the program's estimated effects will exceed its true effects.[6] Finally, a study that randomly assigns welfare recipients to a pilot welfare program or a control group might underestimate the program's effects because the amount of word-of-mouth communication and public discussion about the program, which could influence current and potential recipients' behavior, would probably be lower than if the program were implemented at full scale as welfare policy. In each of these examples, measuring the full effects of an intervention requires randomly assigning larger communities, not individuals, to the program or a control group.

In certain contexts, cluster randomization is not only theoretically sound but also practical, as demonstrated by its application in many studies (for a review of such applications, see Boruch and Foley 2000). Among the interventions that have been studied using cluster random assignment are programs to reduce drug and alcohol use among high school students, public health campaigns, crime prevention initiatives, and new hospital procedures. The most commonly randomized entities have been schools or classrooms, but random assignment has also been conducted among industrial organizations, adult literacy centers, Goodwill service providers, hospitals, police patrol areas, and entire communities. This record of successful application casts doubt on arguments that experiments cannot be conducted on aggregate entities. Chapter 4 of the present volume ("Randomizing Groups to Evaluate Place-Based Programs," by Howard Bloom) presents a detailed discussion of this approach. The chapter outlines the main reasons for randomizing groups, describes the statistical properties of the approach, considers ways to improve the approach by using background information on the randomized groups and their members, and examines the statistical properties of subgroup findings produced by the approach. The chapter also considers the conceptual, statistical, programmatic, and policy implications of the fact that individuals move in and out of randomized groups over time.

Assessing Nonexperimental Statistical Methods

If many questions can be answered only by going beyond random assignment, it is natural to wonder whether one should start with an experimental design at all. Is the problem of selection bias in nonexperimental studies more theoretical than real? Fortunately, this question can be answered by reviewing research that has compared nonexperimentally derived impact estimates with impact estimates based on well-executed random-assignment studies. Overall, the findings indicate that, even under the most favorable circumstances, nonexperimental statistical methods are often far off the mark.

One group of studies has assessed nonexperimental statistical methods by comparing experimental and nonexperimental impact estimates on the basis of data collected from individuals who were randomly assigned to a treatment or a control group. Much of this research was based on data from the National Supported Work (NSW) Demonstration. Conducted by a consortium of organizations in the mid-1970s at twelve sites across the United States, the NSW study evaluated voluntary training and assisted work programs targeted at four groups of individuals with serious barriers to employment: long-term recipients of Aid to Families with Dependent Children (AFDC), which was the federal cash welfare program until 1996; former drug addicts; former criminal offenders; and young school dropouts. To assess nonexperimental estimators of the program's effects, the results of this randomized experiment have been compared with results obtained by comparing the NSW group's outcomes with the outcomes for comparison groups drawn from two national surveys: the Current Population Survey and the Panel Study of Income Dynamics.

The earliest assessment of nonexperimental estimators using NSW data sounded an alarm, revealing that large biases can arise from using nonexperimental comparison groups (LaLonde 1986; Fraker and Maynard 1987). But the most biased nonexperimental results were later identified and excluded by means of statistical tests used to reject comparison groups whose characteristics differed markedly from those of the program group at baseline (Heckman and Hotz 1989). Another method that produced estimates close to the experimental benchmark (Dehejia and Wahba 1999) is matching, in which differences between the groups at baseline are eliminated by selecting comparison group members similar to program group members with respect to such characteristics as gender, level of education, and employment history. A recent reanalysis, however, indicates that matching worked well only for a specific subsample of the NSW data (Smith and Todd 2005).

Inspired by the early studies of date from NSW, Daniel Friedlander and Philip K. Robins (1995) compared experimentally derived estimates of program effects with nonexperimentally derived estimates using data from random-assignment studies of mandatory state welfare-to-work programs operated in the early to mid-1980s. They examined comparison groups drawn from three sources: earlier cohorts of welfare recipients from the same local welfare offices, welfare recipients from other local offices in the same state, and welfare recipients from other states. The authors found that in-state comparison groups worked better than out-of-state comparison groups, although both were problematic. Furthermore, they found that the statistical tests suggested by James Heckman and V. Joseph Hotz (1989) did not adequately distinguish between good and bad estimators with their data—which later work by Heckman, Ichimura, and Todd (1997, 629) also indicates. A more recent analysis of a mandatory welfare-to-work program in Indiana confirmed that the bias can be quite large, even when the comparison sites have labor market characteristics similar to those in the program sites (Lee 2001).

Another comparison of nonexperimental and experimental results used data from an experimental study of a set of voluntary training and subsidized work programs for recipients of AFDC piloted in seven states in the mid- to late 1980s (Bell et al. 1995). Nonexperimental comparison groups were drawn from withdrawals, screen-outs, or no-shows. The authors found that estimates based on no-shows were the most accurate, those based on screen-outs were the next most accurate, and those based on withdrawals were the least accurate. In addition, the accuracy of estimates based on screen-outs improved over time, from being only slightly better than those for withdrawals at the beginning of the follow-up period to being almost as good as those for no-shows at the end.

The most comprehensive, detailed, and technically sophisticated assessment of nonexperimental estimators to date used a data set constructed for the evaluation of employment and training programs for economically disadvantaged adults and youth funded by the Job Training and Partnership Act (JTPA) of 1982 (Heckman, Ichimura, and Todd 1997, 1998; Heckman et al. 1998). Results from this study underscored the importance of choosing an appropriate comparison group, particularly one from the same local labor market as the program group and one for which comparable measures have been collected. But even the best methods left some amount of bias arising from selection on unobserved factors.

Although limited to a single policy area—namely, employment and training programs—the methodological research that has grown out of these four sets of experiments spans a lengthy period (from the 1970s to the 1990s); many different geographic areas representing different labor market structures; voluntary and mandatory programs, the characteristics of whose participants probably reflect very different selection processes; a wide variety of comparison group sources, including national surveys and past participants in the same program; and a vast array of statistical and econometric methods for estimating program effects using nonexperimental comparison groups (for a formal meta-analysis of the results, see Glazerman, Levy, and Myers 2003).

Two recent studies using a similar approach found that nonexperimental comparisons did not yield results similar to random-assignment evaluations of education programs (Agodini and Dynarski 2001; Wilde and Hollister 2002).

Another approach to comparing experimental and nonexperimental estimates is to use meta-analysis to summarize and contrast findings from a series of both types of studies. Beginning with Mary Lee Smith, Gene V. Glass, and Thomas I. Miller (1980), meta-analyses comparing findings from experimental and nonexperimental studies have had mixed results (Heinsman and Shadish 1996).

Perhaps the most extensive such comparison is a “meta-analysis of meta-analyses” that synthesizes past research on the effectiveness of psychological, behavioral, and education treatments (Lipsey and Wilson 1993). In one part of it, the authors compared the means and standard deviations of experimental and nonexperimental estimates drawn from seventy-four meta-analyses for which findings from both types of studies were available. This comparison, which represented hundreds of primary studies, showed virtually no difference in the mean effect as estimated on the basis of experimental statistical studies, as opposed to nonexperimental. Although some of the meta-analyses reported a large difference between average experimental and nonexperimental estimates, the differences were as likely to be positive as negative, and they canceled out across the seventy-four meta-analyses.

Meta-analytic comparisons of experimental and nonexperimental estimates have also been made for voluntary employment and training programs (Greenberg, Michalopoulos, and Robins 2003, 2004). According to these comparisons, the average impact estimate for men has been substantially larger for experimentally evaluated programs than for other programs, and the difference is statistically significant. This is consistent with the results using the NSW and JTPA evaluations that implied that nonexperimental statistical methods might be subject to substantial selection bias. For women and teens, by contrast, experimental and nonexperimental evaluations of voluntary employment-training programs yielded similar estimated effects. Moreover, the two types of evaluations indicated similar changes in the effectiveness of programs over time for all three groups. It is important not to make too much of these comparisons, however, since they include only six programs. In addition, most of the nonexperimental evaluations were conducted before 1975, but most of the random-assignment studies occurred later. Thus, the two sets of evaluations might not have analyzed comparable treatments.

In chapter 5 of the present volume, “Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects,” Howard Bloom, Charles Michalopoulos, and Carolyn Hill use random-assignment studies of mandatory welfare-to-work programs to assess the utility of a variety of nonexperimental estimators of program impacts, some of which have been examined only in the context of voluntary programs. The chapter addresses the questions of which nonexperimental comparison group methods work best, under what conditions they do so, and under what conditions, if any, the best nonexperimental comparison group methods perform well enough to be used instead of random assignment.

Goal of This Book

Although the next four chapters in this volume deal with very different questions, they have two features in common: a recognition that random assignment provides valid and useful estimates of the effects of social interventions and a belief that more can be learned from experiments by combining them with nonexperimental statistical techniques. By no means a comprehensive survey of ways to integrate experimental and nonexperimental approaches, the book is meant to inspire researchers both to use randomized experiments and to go beyond experiments in their pursuit of answers to important social questions.

Notes

1. For a much more thorough treatment of the history of statistics, see Stephen M. Stigler (1986, 1989), Gerd Gigerenzer et al. (1989), and Lorenz Krüger et al. (1987). A particularly accessible and wide-ranging description of the development of statistical theory and empirical methods can be found in David Salsburg (2001).
2. Stigler's (1986) view is only one of many. As represented by Mary S. Morgan (1990), Theodore Porter (1986), for example, thought the key breakthrough in the social sciences was a philosophical shift toward interpreting individual differences in behavior as natural variation attributable to the complexity of human affairs, which in turn opened the door to the use of techniques such as correlation and regression. J. L. Klein (1986), in contrast, argued that economists began adapting and applying statistical methods used by astronomers and biometricians because it facilitated the modeling of economic time-series phenomena.
3. George S. Maddala (1983) provides a detailed discussion of the assumptions and methods of estimating the model, which is often credited to James J. Heckman (1974). The assumptions have been relaxed in statistical refinements of the model, but even the refinements demand identification of a factor that affects whether someone receives an intervention but does not influence the effect of the intervention.
4. Random assignment too has been criticized (see, for instance, Manski and Garfinkel 1992; Bloom, Cordray, and Light 1988; Hausman and Wise 1985). Among the most potent criticisms are the following: (1) The process of random assignment can affect implementation of the program under study such that the resulting estimates may not reflect the impacts of the actual program but rather of the program when it is under study; (2) being in a study affects people in ways unrelated to the treatment (often called the Hawthorne effect) by, for instance, encouraging program group members to adhere more closely to the treatment than they would if they received the treatment outside the study; (3) the program can affect the surrounding community (for example, through displacement, in which program group members take jobs that would otherwise be held by oth-

ers), causing experimental estimates of the program's impacts to systematically misestimate its true effects; and (4) some individuals do not receive the treatment to which they are randomly assigned, resulting in underestimates of the effect of the treatment. Because these concerns have been addressed in detail elsewhere and are tangential to the themes in this book, they are not discussed further here.

5. In this case, meta-analysis might have led to a misleading result. Analyses of the impacts of random assignment studies of U.S. welfare-to-work programs by subgroup show little systematic evidence that white welfare recipients fare better than black welfare recipients (Michalopoulos and Schwartz 2000; Michalopoulos 2004). This result suggests that neither race nor ethnicity per se is important but that the welfare-to-work programs operating in areas with higher proportions of white recipients happened to be more effective.
6. Although displacement of workers is often considered a potential source of bias in experimental estimates of the effects of employment programs, a number of empirical studies found little evidence for displacement resulting from such programs (Friedberg and Hunt 1995). However, a recent paper argues that displacement might be more problematic in some experiments (Lise, Seitz, and Smith 2004).

References

Agodini, Roberto, and Mark Dynarski. 2001. *Are Experiments the Only Option? A Look at Dropout Prevention Programs.* Princeton, N.J.: Mathematica Policy Research.

Aigner, Dennis J. 1985. "The Residential Time-of-Use Pricing Experiments: What Have We Learned?" In *Social Experimentation*, edited by Jerry A. Hausman and David A. Wise. Chicago: University of Chicago Press.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 50(4): 69–85.

Ashworth, Karl, Andreas Cebulla, David Greenberg, and Robert Walker. 2001. "Meta-Evaluation: Discovering What Works Best in Welfare Provision." Unpublished paper. University of Nottingham, England.

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

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies Review Annual,* edited by Ernst Stromsdorfer and George Farkas. Volume 5. San Francisco: Sage Publications.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs*. Kalamazoo, Mich.: W. E. Upjohn Institute for Employment Research.

Bell, Stephen, Michael Puma, Gary Shapiro, Ronna Cook, and Michael Lopez. 2003. "Random Assignment for Impact Analysis in a Statistically Representative Set of Sites: Issues from the National Head Start Impact Study." *Pro-*

ceedings of the August 2003 American Statistical Association Joint Statistical Meetings (CD-ROM).

Bloom, Dan, and Charles Michalopoulos. 2001. *How Welfare and Work Policies Affect Employment and Income: A Synthesis of Research*. New York: MDRC.

Bloom, Howard. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8(2): 225–46.

Bloom, Howard S., David S. Cordray, and Richard J. Light, eds. 1988. *Lessons from Selected Program and Policy Areas*. San Francisco: Jossey-Bass.

Bloom, Howard, Saul Schwartz, Susanna Lui-Gurr, and Suk-Won Lee. 1999. *Testing a Re-employment Incentive for Displaced Workers: The Earnings Supplement Project.* Ottawa, Canada: Social Research and Demonstration Corporation.

Boruch, Robert F. 1975. "Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation." *Sociological Methods and Research* 4: 31–53.

Boruch, Robert F., and Ellen Foley. 2000. "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials." In *Validity and Social Experimentation: Donald Campbell's Legacy,* edited by Leonard Bickman. Thousand Oaks, Calif.: Sage Publications.

Borus, Michael E. 1964. "A Benefit-Cost Analysis of the Economic Effectiveness of Retraining the Unemployed." *Yale Economic Essays* 4(2): 371–430.

Burghardt, John, Peter Z. Schochet, Sheena McConnell, Terry Johnson, R. Mark Gritz, Steven Glazerman, John Homrighausen, and Russell Jackson. 2001. "Does Job Corps Work? Summary of the National Job Corps Study." Princeton, N.J.: Mathematica Policy Research.

Burtless, Gary. 1987. "The Work Response to a Guaranteed Income: A Survey of Experimental Evidence." In *Lessons from the Income Maintenance Experiments*, edited by Alicia Munnell. Boston: Federal Reserve Bank of Boston.

Cain, Glen G. 1975. "Regression and Selection Models to Improve Nonexperimental Comparisons." In *Evaluation and Experiments: Some Critical Issues in Assessing Social Programs,* edited by Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press.

Ciarlo, James A., and Charles Windle. 1988. "Mental Health Evaluation and Needs Assessment." In *Lessons from Selected Program and Policy Areas*, edited by Howard S. Bloom, David S. Cordray, and Richard J. Light. San Francisco: Jossey-Bass.

Clark-Kauffman, Elizabeth, Greg J. Duncan, and Pamela Morris. 2003. "How Welfare Policies Affect Child and Adolescent Achievement." *American Economic Review: Papers and Proceedings of the American Economic Association* 93(2): 299–303.

Coalition for Evidence-Based Policy. 2003. *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User-Friendly Guide*. Washington: U.S. Department of Education, Institute of Education Sciences.

Cochrane Collaboration. 2002. "Cochrane Central Register of Controlled Trials." Database. Available at The Cochrane Library: www.cochrane.org (accessed September 14, 2004).

Coleman, William. 1987. "Experimental Psychology and Statistical Inference: The Therapeutic Trial in Nineteenth-Century Germany." In *The Probabilistic*

Revolution, volume 2: *Ideas in the Sciences*, edited by Lorenz Krüger, Gerg Gigerenzer, and Mary S. Morgan. Cambridge, Mass.: MIT Press.

Danziger, Kurt. 1987. "Statistical Methods and the Historical Development of Research Practice in American Psychology." In *The Probabilistic Revolution*, volume 2: *Ideas in the Sciences*, edited by Lorenz Krüger, Gerd Gigerenzer, and Mary S. Morgan. Cambridge, Mass.: MIT Press.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(488): 1053–62.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver & Boyd.

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

Friedberg, Rachel M., and Jennifer Hunt. 1995. "The Impact of Immigrants on Host Country Wages, Employment and Growth." *Journal of Economic Perspectives* 9(2): 23–44.

Friedlander, Daniel, and Philip K. Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review* 85(4): 923–37.

Garfinkel, Irwin, Charles F. Manski, and Charles Michalopoulos. 1992. "Micro Experiments and Macro Effects." In *Evaluating Welfare and Training Programs*, edited by Charles F. Manski. and Irwin Garfinkel. Cambridge, Mass.: Harvard University Press.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Krüger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press.

Gilbert, John P., Richard J. Light, and Frederick Mosteller. 1975. "Assessing Social Innovations: An Empirical Basis for Policy." In *Evaluation and Experiment*, edited by Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press.

Glass, Gene V. 1976. "Primary, Secondary, and Meta-Analysis of Research." *Educational Researcher* 5(10): 3–8.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science* 589: 63–93.

Glenman, Thomas K. 1972. "Evaluating Federal Manpower Programs." In *Evaluating Social Programs*, edited by Peter H. Rossi and Walter Williams. New York: Seminar Press.

Goering, John, Joan Kraft, Judith Feins, Debra McInnis, Mary Joel Holin, and Huda Elhassan. 1999. *Moving to Opportunity for Fair Housing Demonstration Program: Current Status and Initial Findings*. Washington: U.S. Department of Housing and Urban Development.

Greenberg, David H., Robert Meyer, Charles Michalopoulos, and Michael Wiseman. 2003. "Explaining Variation in the Effects of Welfare-to-Work Programs." *Evaluation Review* 27(4): 359–94.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robins. 2003. "A

Meta-Analysis of Government-Sponsored Training Programs." *Industrial and Labor Relations Review* 50(1): 31–53.

———. 2004. "What Happens to the Effects of Government-Funded Training Programs over Time?" *Journal of Human Resources* 39(1): 277–93.

Greenberg, David H., and Philip K. Robins. 1986. "Social Experiments in Policy Analysis." *Journal of Policy Analysis and Management* 5(2): 340–62.

Greenberg, David, and Mark Shroder. 1997. *The Digest of Social Experiments*. Washington, D.C.: Urban Institute Press.

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

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

Hamilton, Gayle, Stephen Freedman, Lisa A. Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo, Anna Gassman-Pines, Sharon McGroder, Martha Zaslow, Jennifer Brooks, and Surjeet Ahluwalia. 2001. *How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs*. Washington: U.S. Department of Health and Human Services and U.S. Department of Education.

Harris, Jeffrey E. 1985. "Macroexperiments and Microexperiments for Health Policy." In *Social Experimentation*, edited by Jerry A. Hausman and David A. Wise. Chicago: University of Chicago Press.

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

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42(4): 679–94.

———. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, edited by Charles F. Manski and Irwin Garfinkel. Cambridge, Mass.: Harvard University Press.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–74.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–98.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4): 605–54.

———. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2): 261–94.

Heinsman, Donna T., and William R. Shadish. 1996. "Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate Answers from Randomized Experiments?" *Psychological Methods* 1(2): 154–69.

Hollister, Robinson G., and Jennifer Hill. 1995. "Problems in the Evaluation of Community-Wide Initiatives." In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, edited by James P. Connell, Anne

C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss. Washington, D.C.: Aspen Institute.

Kemple, James, and Joshua Haimson. 1994. *Florida's Project Independence: Program Implementation, Participation Patterns, and First-Year Impacts*. New York: MDRC.

Kemple, James J., and Jason Snipes. 2000. *Career Academies: Impacts on Students' Engagement and Performance in High School*. New York: MDRC.

Kennedy, Stephen D. 1988. "Direct Cash Low-Income Housing Assistance." In *Lessons from Selected Program and Policy Areas*, edited by Howard S. Bloom, David S. Cordray, and Richard J. Light. San Francisco: Jossey-Bass.

Klein, J. L. 1986. "The Conceptual Development of Population and Variation as Foundations of Econometric Analysis." Ph.D. diss., City of London Polytechnic.

Krüger, Lorenz, Gerd Gigerenzer, and Mary S. Morgan, eds. 1987. *The Probabilistic Revolution*, volume 2: *Ideas in the Sciences*. Cambridge, Mass.: MIT Press.

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

Lee, Wang S. 2001. "Propensity Score Matching on Commonly Available Nonexperimental Comparison Groups." Unpublished paper. Bethesda, Md.: Abt Associates.

Lipsey, Mark W. 1988. "Juvenile Delinquency Intervention." In *Lessons from Selected Program and Policy Areas*, edited by Howard S. Bloom, David S. Cordray, and Richard J. Light. San Francisco: Jossey-Bass.

Lipsey, Mark W., and David B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment." *American Psychologist* 48(12): 1181–1209.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2004. "Equilibrium Policy Experiments and the Evaluation of Social Programs." NBER working paper 10283. Cambridge, Mass.: National Bureau of Economic Research.

Love, John M., Ellen Eliason Kisker, Christine M. Ross, Peter Z. Schochet, Jeanne Brooks-Gunn, Dianne Paulsell, Kimberly Boller, Jill Constantine, Cheri Vogel, Allison Sidle Fuligni, and Christy Brady-Smith. 2002. *Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start*. Washington: U.S. Department of Health and Human Services.

Ludwig, Jens, Paul Hirschfield, and Greg J. Duncan. 2001. "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment." *Quarterly Journal of Economics* 116(2): 655–79.

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

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

Marks, Harry M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge: Cambridge University Press.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell. 2002. *School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program*. Princeton, N.J.: Mathematica Policy Research.

McDill, Edward L., Mary S. McDill, and J. Timothy Sprehe. 1972. "Evaluation in Practice: Compensatory Education." In *Evaluating Social Programs*, edited by Peter H. Rossi and Walter Williams. New York: Seminar Press.

Michalopoulos, Charles. 2004. "What Works Best for Whom: The Effects of Welfare and Work Policies by Race and Ethnicity." *Eastern Economic Journal* 30: 53–79.

Michalopoulos, Charles, and Christine Schwartz. 2000. *What Works Best for Whom: Impacts of 20 Welfare-to-Work Programs by Subgroup*. Washington: U.S. Department of Health and Human Services and U.S. Department of Education.

Michalopoulos, Charles, Douglas Tattrie, Cynthia Miller, Philip K. Robins, Pamela Morris, David Gyarmati, Cindy Redcross, Kelly Foley, and Reuben Ford. 2002. *Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients*. Ottawa, Canada: Social Research and Demonstration Corporation.

Miller, Cynthia, Virginia Knox, Lisa A. Gennetian, Martey Dodoo, JoAnna Hunter, and Cindy Redcross. 2000. *Reforming Welfare and Rewarding Work: Final Report on the Minnesota Family Investment Program.* New York. MDRC.

Morgan, Mary S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morris, Pamela A., Aletha C. Huston, Greg J. Duncan, Danielle A. Crosby, and Johannes M. Bos. 2001. *How Welfare and Work Policies Affect Children: A Synthesis of Research*. New York: MDRC.

Mosteller, Frederick, Richard J. Light, and Jason A. Sachs. 1996. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size." *Harvard Educational Review* 66(4): 797–842.

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

Newhouse, Joseph P. 1996. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, Mass.: Harvard University Press.

Orr, Larry L. 1999. *Social Experiments*. Thousand Oaks, Calif.: Sage Publications.

Orr, Larry, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbomatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling. 2003. *Moving to Opportunity: Interim Impacts Evaluation*. Washington: U.S. Department of Housing and Urban Development.

Peck, Laura. 2002. "Subgroup Analyses in Social Experiments." Unpublished paper. Arizona State University, School of Public Affairs.

Peirce, Charles S., and Joseph Jastrow. 1884/1980. "On Small Differences of Sensation." Reprinted in *American Contributions to Mathematical Statistics in the Nineteenth Century,* volume 2, edited by Stephen M. Stigler. New York: Arno Press.

Pollack, Andrew, and Lawrence K. Altman. 2003. "Large Trial Finds AIDS Vaccine Fails to Stop Infection." *New York Times*, February 24, 2003, p. A1, 1.

Porter, Theodore M. 1986. *The Rise of Statistical Thinking, 1820–1900*. Princeton: Princeton University Press.

Ramey, Craig T., Keith Owen Yeates, and Elizabeth J. Short. 1984. "The Plasticity of Intellectual Development: Insights from Preventive Intervention." *Child Development* 55: 1913–25.

Riccio, James, Daniel Friedlander, and Stephen Freedman. 1994. *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program*. New York: MDRC.

Rossi, Peter H., and Howard E. Freeman. 1993. *Evaluation: A Systematic Approach*. Thousand Oaks, Calif.: Sage Publications.

Rossi, Peter H., and Walter Williams. 1972. *Evaluating Social Programs*. New York: Seminar Press.

Salsburg, David. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* New York: Henry Holt.

Schweinhart, Lawrence J., and David P. Weikart. 1993. "Success by Empowerment: The High/Scope Perry Preschool Study through Age 27." *Young Children* 49(1): 54–58.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Shadish, William R., Kevin Ragsdale, Benita R. Glaser, and Linda M. Montgomery. 1995. "The Efficacy and Effectiveness of Marital and Family Therapy: A Perspective from Meta-Analysis." *Journal of Marital and Family Therapy* 21: 345–60.

Sherman, Lawrence W. 1988. "Randomized Experiments in Criminal Sanctions." In *Lessons from Selected Program and Policy Areas*, edited by Howard S. Bloom, David S. Cordray, and Richard J. Light. San Francisco: Jossey-Bass.

Smith, Jeffrey, and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1–2): 305–53.

Smith, Mary Lee, Gene V. Glass, and Thomas I. Miller. 1980. *The Benefits of Psychotherapy*. Baltimore: Johns Hopkins University Press.

Smith, Vernon. 1994. "Economics in the Laboratory." *Journal of Economic Perspectives* 8(1): 113–31.

Spiegelman, Robert G., Christopher J. O'Leary, and Kenneth J. Kline. 1992. *The Washington Reemployment Bonus Experiment: Final Report*. Unemployment Insurance occasional paper 92–6. Washington: U.S. Department of Labor.

Stephan, A. S. 1935. "Prospects and Possibilities: The New Deal and the New Social Research." *Social Forces* 13(4): 515–21.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass.: Belknap Press.

———. 1999. *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, Mass.: Harvard University Press.

Teruel, Graciela M., and Benjamin Davis. 2000. *Final Report: An Evaluation of the Impact of PROGRESA Cash Payments on Private Inter-Household Transfers.* Washington, D.C.: International Food Policy Research Institute.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. "Regression Discontinuity Analysis: An Alternative to Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6): 309–17.

van Helmont, John Baptista. 1662. *Oriatrik or, Physick Refined: The Common Errors Therein Refuted and the Whole Art Reformed and Rectified.* London: Lodowick-Lloyd. Available at the James Lind Library web site: www.jameslindlibrary.org/trial_records/17th_18th_Century/van_helmont/van_helmont_kp.html (accessed January 3, 2005).

Wilde, Elizabeth Ty, and Robinson Hollister. 2002. “How Close Is Close Enough? Testing Nonexperimental Estimates of Impact Against Experimental Estimates of Impact with Education Test Scores as Outcomes.” Discussion paper 1242–02. Madison, Wis.: Institute for Research on Poverty. Available at the University of Wisconsin—Madison web site: www.ssc.wisc.edu/irp/pubs/dp124202.pdf (accessed January 3, 2005).

Wilner, Daniel M., Rosabelle P. Walkley, Thomas C. Pinkerton, and Mathew Tayback. 1962. *The Housing Environment and Family Life*. Baltimore: Johns Hopkins University Press.

Woodbury, Stephen A. and Robert G. Spiegelman. 1987. “Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois.” *American Economic Review* 77(4): 513–30.