The Renaissance of Field Experimentation in Evaluating Interventions

William R. Shadish[1] and Thomas D. Cook[2]

[1]University of California, Merced, California 95344; email: wshadish@ucmerced.edu

[2]Institute for Policy Research, Northwestern University, Evanston, Illinois 60208-4100; email: t-cook@northwestern.edu

Annu. Rev. Psychol. 2009. 60:607–29

First published online as a Review in Advance on July 9, 2008

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi: 10.1146/annurev.psych.60.110707.163544

0066-4308/09/0110-0607$20.00

Key Words

experiment, quasi-experiment, regression discontinuity, propensity scores, matching, time series

Abstract

Most experiments are done in laboratories. However, there is also a theory and practice of field experimentation. It has had its successes and failures over the past four decades but is now increasingly used for answering causal questions. This is true for both randomized and—perhaps more surprisingly—nonrandomized experiments. In this article, we review the history of the use of field experiments, discuss some of the reasons for their current renaissance, and focus the bulk of the article on the particular technical developments that have made this renaissance possible across four kinds of widely used experimental and quasi-experimental designs—randomized experiments, regression discontinuity designs in which those units above a cutoff get one treatment and those below get another, short interrupted time series, and nonrandomized experiments using a nonequivalent comparison group. We focus this review on some of the key technical developments addressing problems that previously stymied accurate effect estimation, the solution of which opens the way for accurate estimation of effects under the often difficult conditions of field implementation—the estimation of treatment effects under partial treatment implementation, the prevention and analysis of attrition, analysis of nested designs, new analytic developments for both regression discontinuity designs and short interrupted time series, and propensity score analysis. We also cover the key empirical evidence showing the conditions under which some nonrandomized experiments may be able to approximate results from randomized experiments.

Contents

HISTORY IN BRIEF

The modern use of field experiments in the social sciences first became widespread in the 1960s (Cook & Shadish 1994, Riecken et al. 1974, Shadish et al. 1991). An exemplar is the New Jersey Negative Income Tax Experiment (Nathan 1989, Rossi & Lyall 1978), which was designed in the late 1960s to test whether guaranteeing working poor families an income they could spend at their own discretion might serve as an alternative to the welfare system of the day that was heavy with professional services. The fear was widespread that providing cash instead of services would undermine the work motivation of individuals and would lead to welfare as a lifestyle. Hence, the major purpose of the demonstration was to estimate how different income guarantees, combined with different marginal tax policies, affected labor force participation. This was tested in a large randomized experiment that took many years and many millions of dollars to complete.

Randomized experiment: a design that assigns units to conditions based on some chance process such as the toss of a coin

However, all did not go smoothly. The basic idea was very controversial. Some commentators believed that it promoted the wrong values, while many bureaucratic and professional groups felt that their own welfare would be compromised if cash were substituted for services. The length of the experiment meant that by the time the results were available, the policymakers who were originally interested in the idea—the Nixon administration in this case—were no longer in office, and a new set of policy questions were at the fore. Questions also arose about how to analyze the data, given that the participants did not always accept the conditions to which they were assigned, and others dropped out before completing the study. For all these reasons, many observers wondered whether the time and effort invested in this experiment were worthwhile.

This study is just one example of the large-scale field experiments conducted beginning in the 1960s and 1970s. Others include the National Institute of Mental Health Collaborative Depression Project (Elkin et al. 1985, 1989), the Head Start and Follow Through evaluations (Wholey et al. 1970), and the Manhattan Bail Bond Experiment (Botein 1965). These experiments were partly fueled by demand for data about the implementation and effects of the many social programs introduced under President Lyndon Johnson's Great Society initiatives. Such experiments were time consuming and expensive, and hopes were high that their results would be definitive and useful for policy.

However, these pioneering experiments seemed neither as definitive nor as useful as had first been hoped. Some problems were logistical. For example, some experiments were unable to recruit enough participants, and various participants were unwilling to accept random assignment. Other problems reflected an unexpectedly diverse set of social reactions to

the design of the experiment. For instance, some groups with a stake in the intervention or its outcomes began to challenge whether the treatment really reflected what participants needed and whether the outcomes measured were what the treatment was designed to accomplish. Still other problems were technical. In randomized experiments, poorly implemented random assignment procedures and differential attrition from conditions compromised the group equivalence that randomization aimed to achieve, and it was unclear what analyses might compensate. In nonrandomized experiments, selection biases at the start of the study led to similar compromises, exacerbated because no statistical method for adjusting those biases had either a clear theoretical or an empirical warrant. In all experiments, both partial treatment implementation and participant crossover from one condition to another led to confusion about what inferences about treatment effects were justified. Similarly, both randomized and nonrandomized experiments encountered problems of units nested within conditions (e.g., students nested within classrooms) that violated independence assumptions of ordinary least squares statistics in a manner that led to increased Type I errors—concluding that an intervention worked when a proper analysis might not support that conclusion.

When social experiments were finally completed and reported, years had usually passed since raising the question the experiment addressed. Often the stakeholders who first asked the question were no longer players in the policy-shaping community, having left political office, been removed or resigned from appointed positions, or moved on to other social issues. Even when the policy-shaping community retained interest in the original question, the experimental results proved less than definitive. This was because of not only the implementation and technical problems described above, but also because different stakeholders could legitimately debate the meaning and value of the observed outcomes and because decisions about adopting an intervention turned out to be far less driven by whether an intervention works than by other social, economic, and political factors.

These problems caused many scholars and practitioners to raise serious and sustained questions about whether field experiments were a viable or valuable contribution to either science or policy (e.g., Cronbach 1982, Cronbach et al. 1980). Critics questioned whether causal questions were worth asking, whether field experiments were capable of answering any questions that might be worth asking, and whether the disputed answers that experiments provided would be useful in the cases in which they were possible. Consequently, the 1970s and 1980s saw a proliferation of alternative methods for studying social interventions, ranging from nonexperimental econometric methods to qualitative and anthropological case studies, as well as a demotion of causal questions in the understanding of which questions were worth asking. Entire fields, particularly education but also some parts of economics and sociology, largely rejected experimentation in favor of these and other alternatives.

Yet interest in field experimentation remained strong in some areas, especially in medicine, public health, and a few subdisciplines. Examples include parts of labor economics where funding for large-scale experiments continued, and parts of clinical psychology where the implementation and technical problems of experimentation were never as great because researchers had high levels of control over the intervention and its setting. Slowly over four decades, the sustained attention in these areas to solving the early problems of field experimentation laid the foundations for today's renaissance.

After reviewing the reasons why field experimentation has staged such a dramatic comeback, this article then examines the key issues on which progress has been made. We organize the review around the major kinds of experiments in order of their strength for causal inference, from the strongest randomized designs through the presumed weakest designs with nonrandomized controls. In each of these designs, we show how progress had been made

Regression discontinuity design: a quasi-experimental design that uses a cutoff score on a measured variable to assign some units to treatment and others to control

in addressing all of the key problems that emerged with the design in the early years for field experimentation, and we discuss the empirical evidence suggesting the conditions under which each of the nonrandomized designs might provide effect estimates comparable to randomized designs. We conclude with a discussion of some of the issues that remain to be addressed, some technical issues, but mostly social and political issues concerning the use of results in practice and policy.

MODERN RENAISSANCE OF FIELD EXPERIMENTATION

A number of developments in the past two decades created the technical and social conditions under which field experimentation could again flourish on a wide scale.

Evidence-Based Practice

We do not know who first used the term evidence-based practice, but the term or its cognates—and more importantly the policy of using fiscal and organizational incentives to encourage practitioners and administrators to develop and adopt interventions with demonstrated empirical effectiveness—are now widespread in nearly all fields where practitioners intervene with people. Medicine and public health pioneered in developing the ideology and social infrastructure to support evidence-based practice, as best seen in the Cochrane Collaboration, an organization in the United Kingdom begun in 1993 and dedicated to improving health care globally through systematic reviews of the effectiveness of health care interventions (**www.cochrane.org**). In that regard, a nod must also go to the development of meta-analysis (Glass 1976), the methodology for creating quantitative syntheses of evidence often applied to studies about the effects of treatments. A host of scholarly and governmental organizations both adopted the ideology and contributed to the methodology in areas such as prevention research, clinical psychology, and public health. Interest in evidence-based practice was partly ideological but also economic. In clinical psychology, for example, insurance companies stressed the role that demonstrated efficacy would play in reimbursing psychotherapists.

In fields where ideology and recent practice had rejected experimentation, most notably education, the turn of the twenty-first century saw federal mandates for the use both of experiments and of interventions with experimental support, and funding priorities underwent a large shift to back this mandate. An example is the creation of the Institute of Education Sciences within the U.S. Department of Education under the leadership of R. Grover Whitehurst. Among other things, the Institute of Education Sciences gave preferential funding to randomized experiments (and sometimes high-quality nonrandomized experiments such as the regression discontinuity design), created the What Works Clearinghouse to synthesize research about effective educational interventions, and initiated a new professional association (the Society for Research on Educational Effectiveness) to counter resistance in the traditional educational establishment to experimentation. The result was a detectable increase in the use of experimental methods in education (Constas 2007).

Renewed Interest from Economists and Statisticians

Simultaneous developments in economics and statistics also played a major role in the renaissance. The majority of economists interested in causal inference had long subordinated experimental design to statistical analyses that adjust nonexperimental data for selection bias, most notably using selection bias models like those developed by Heckman (e.g., Heckman 1979, 1992). Yet a subset of economists retained an interest in experiments, fueled in part by demonstrations that selection bias models could not reproduce gold standard experimental results (Glazerman et al. 2003, LaLonde & Maynard 1987). At the same time, a few private sector research corporations staffed mostly by

economists were solving some of the practical and technical problems in early economic experiments, most notably the Manpower Development Research Corporation (now simply MDRC) headed by economist Judith Gueron (Gueron 1985). Consequent to such developments, by the twenty-first century, many economists—especially young labor economists at the best universities—renewed their interests in strong experimental design. This included both randomized experiments and the regression discontinuity design that was revived from obscurity almost single-handedly by its widespread adoption in economics (Cook 2007).

Similarly, although statisticians had always expressed a preference for randomized experiments, they rarely contributed to the solution of practical problems with them and even more rarely addressed the problems of nonrandomized experiments. That changed with widespread interest in work by statistician Donald Rubin (e.g., Rosenbaum & Rubin 1983, Rubin 1974), himself a student of one of the few statisticians with sustained interest in nonrandomized experiments, William G. Cochran. A good summary of what is often referred to as Rubin's Causal Model, or more precisely the potential outcomes model, is found in Rubin (2004). The three key elements of Rubin's Causal Model are units, treatments, and potential outcomes. If Y is the outcome measure, define $Y(1)$ as the potential outcome that would be observed if the unit is exposed to a treatment ($W = 1$), and define $Y(0)$ as the potential outcome that would be observed if that same unit is not exposed to the treatment ($W = 0$). If so, then the (potential) effect is the difference between these two potential outcomes, $Y(1)$–$Y(0)$. This effect is defined on each unit, and the average of individual causal effects is the average causal effect. However, these are potential outcomes only until treatment begins. After treatment begins, only $Y(1)$ or $Y(0)$ can be observed, with the other being missing, so that the fundamental problem of causal inference in Rubin's Causal Model is how to estimate the missing outcome. These missing outcomes are sometimes called counterfactuals because they are not, in fact, what happened. Furthermore, after treatment begins, individual causal effects cannot actually be observed as defined [$Y(1)$–$Y(0)$], although the average causal effect over all units can be observed under conditions such as random assignment. When presented in detail (Rubin 2004; W.R. Shadish, manuscript submitted), this model provides a statistical definition of an effect, a set of assumptions that allowed the effect to be calculated, and most importantly, statistical tools for approaching the estimation of effects in nonrandomized experiments that heretofore had resisted such solution (Morgan & Winship 2007; Rubin 2004; W.R. Shadish, manuscript submitted). The model and its attendant developments gave new credibility to causal estimates from nonrandomized experiments, though the limits of this claim are still being contested today.

The Empirical Program of Experimentation

A crucial development was the increased attention to developing empirical evidence about which methodologies and statistical analyses live up to their promise to provide accurate answers to causal questions (Shadish 2000). Early examples were by economists who compared results from randomized and adjusted nonrandomized experiments (Fraker & Maynard 1987, LaLonde & Maynard 1987), but examples today are so widespread that they themselves are the subject of quantitative and narrative reviews (Cook et al. 2008, Glazerman et al. 2003, Heinsman & Shadish 1996, Shadish & Ragsdale 1996). Though consensus has yet to emerge on the results of such inquiries, the methods for doing such empirical work are improving (Shadish et al. 2008), the standards for interpreting such work are becoming more clear (Cook et al. 2008), and there is reason to believe that better understanding is emerging of some of the conditions under which both randomized and nonrandomized experiments can live up to their promise of providing good

answers to causal questions. We discuss such matters below.

RANDOMIZED EXPERIMENTS

Key to the ability to meet the demand for randomized experiments is the creation of an infrastructure to support the effort. Areas such as medicine and public health developed the infrastructure earliest, building a preference for experiments into grant funding, journal publication, and graduate education whenever causal issues were at stake. The wide availability of health-related funding provided the experiential learning necessary to identify and solve practical problems with large-scale experiments. The funding also permitted hiring biostatisticians who developed and applied statistical solutions. Some parts of psychology did likewise, so that a strong tradition of randomized field experiments developed in the evaluation of the effects of psychotherapy, of large-scale behavioral medicine interventions, and of some interventions in school settings to manage classroom behavior and achievement. To a lesser extent, this is also true of labor economics as it persisted in the task of identifying effective interventions after some early problems with experiments (Gueron 1985). In education, which provides a primary outlet for applying psychological ideas, developing an infrastructure similar to that in health or clinical psychology has only recently begun, and without that infrastructure, it will be impossible to counter claims that experiments cannot be conducted in education. A successful tradition of experimentation requires the kind of persistence that many health researchers, many psychologists, and labor economists displayed but that has been lacking in educational research heretofore. Several key methodological developments that improved the yield from experimental work have made the institutionalization task easier.

Implementing Successful Experiments

Some of the early experimental disappointments led to developing methods for dealing with the human, organizational, and political factors required to implement a randomized experiment successfully (Boruch 1997, Gueron 1985, Shadish et al. 2002). These matters, such as ways to discuss random assignment with stakeholders hostile to the idea, seem mundane but are nonetheless crucial to successful experimentation to ensure that enough qualified participants can be recruited to the study, to maximize the chances that random assignment occurs properly, to coordinate large multisite trials, to monitor the implementation of treatment and outcome protocols to ensure fidelity, to review early results for possible adverse outcomes, and to ensure the collected data are digitized rapidly and accurately for later analysis. Workers in some professions, such as labor economics or public health in school settings, stuck with the task of solving these early problems and are now lending their expertise to those in fields that gave up early, such as education, thus cutting short the time it will take the latter to develop solutions. However, professionals in education subfields will inevitably present their own contextual problems that will require at least partially novel solutions.

Partial Treatment Implementation

A key problem in randomized experiments was how to estimate effects when treatments are only partially implemented or when participants cross over from one condition to another in an uncontrolled way. The classic solution is an intent-to-treat (ITT) analysis in which the researcher measures outcomes on all participants no matter what their treatment status, and then analyzes their outcomes in the condition to which they were originally assigned. This analysis preserves the integrity of the initial random assignment and provides an unbiased estimate of the intent to treat that is often of special policy interest. After all, policymakers can rarely make policies that force citizens to accept a treatment; they are still interested in estimating causal effects if only a fraction of those assigned actually receive their assigned treatment.

Other stakeholders are interested in a different causal question: What is the effect of the treatment on the treated (TOT)? This previously was estimated (suboptimally) by comparing those who self-selected treatment with those who self-selected the control (or other treatment comparison) condition, conditioning on whatever covariates are available. This estimate is suboptimal because of error in the covariates and the need for covariates that completely account for the selection process or the outcome. However, some economists and statisticians have outlined a practical solution to this dilemma by combining an instrumental variables tradition in economics with the potential outcomes tradition in statistics (Angrist et al. 1996). An instrumental variable is related to the treatment but not to the outcome except through its relationship on the treatment. With such an instrument, one can obtain an unbiased effect of a treatment despite selection bias. Therefore, instrumental variables have been a mainstay of econometric causal inference for decades even though establishing that one has actually chosen an instrument that meets all the analytic assumptions may be as much a matter of rhetoric as it is a matter of producing evidence. Yet one instrument is undoubtedly bias-free: random assignment to conditions, which by definition can only be related to outcome through its effects on the receipt of treatment. Capitalizing on this, Angrist et al. (1996) made random assignment an instrument for receipt of treatment, thereby providing an unbiased estimate of the effects of treatment on the treated. They also outlined the assumptions needed, which are more stringent than for an ITT analysis but nonetheless often plausible.

Angrist et al. (1996) illustrated the method and its assumptions with the example of the Vietnam draft lottery. Birth dates were assigned random numbers from 1 to 365, and those below a certain number were then subject to the draft (in effect, being randomly assigned to draft eligibility). However, not all those subject to the draft actually served in the military. Suppose the question of interest is whether actually serving in the military increases mortality, the TOT estimator, rather than draft eligibility, the ITT estimator. The standard ITT analysis uses randomization to examine whether draft eligibility increases mortality. Although this yields an unbiased estimate of effects, it will not be the question of greatest interest for all stakeholders. But to compare the mortality of those who did or did not serve in the military is biased by the many unknown factors, in addition to the draft, that caused people to serve (e.g., volunteering because of a family tradition or being cajoled by peers who have enlisted). The Angrist et al. (1996) method provides an unbiased instrumental variable estimate of the TOT question when its assumptions are met.

When both treatment implementation and outcome are dichotomous, the computations require only a hand calculator. The following computations are for the draft example. Because 35.3% of those who were draft eligible actually served in the military, and 19.4% of those who were not eligible also served in the military, we can say that the lottery (random assignment) caused 15.9% to serve in the military. This is the usual ITT analysis of a randomized experiment with military service as the outcome. In addition, 2.04% of the draft eligible died, as did 1.95% of those not eligible. So being draft eligible caused 0.09% to die. This is an ITT analysis with death as the outcome. The unbiased TOT estimate of actually serving in the military on death is simple to calculate, being $0.0009/.159 = 0.0058 = 0.56\%$. So serving in the military caused the death of about one-half of 1% of those soldiers.

Variations on this method rapidly appeared for use in studies with nondichotomous measures of treatment intensity, such as the extent of drug dosage or hours of exam preparation. These variations used multivalued instrumental variables, providing bounds on estimates rather than point estimates (Angrist & Imbens 1995; Balke & Pearl 1997; Barnard et al. 1998; Efron & Feldman 1991; Fischer-Lapp & Goetghebeur 1999; Goetghebeur & Molenberghs 1996; Imbens & Rubin 1997a,b; Little & Yau 1996; Oakes et al. 1993; Robins 1998). The plausibility of assumptions decreases in some of these

applications, and new developments are so rapid that readers are advised to search the most recent literature before relying solely on the references above.

When faced with partial treatment implementation, an ITT analysis remains the method of choice because of its fewer and more transparent assumptions and its frequent policy interest. Nonetheless, the instrumental variables analysis now allows researchers to estimate and report TOT effects that are unbiased in theory and will likely be so in much research practice. Both estimates are usually worth computing.

Attrition

Large amounts of attrition from randomized experiments, that is, loss of participants to outcome measurement, can decrease power and damage the credibility of a randomized experiment. Much worse is differential attrition—when dropouts in one condition differ from dropouts in another in ways related to the outcome. Such attrition is selection bias that occurs after random assignment, and it calls into question whether the resulting effect estimate is unbiased. The problem of attrition is by no means solved, but it has been constructively addressed in two ways. First, we know more about how to prevent attrition from occurring in the first place. This knowledge was developed, by necessity, among researchers working with populations that are hard to find or do not wish to be found, such as abused women, illegal drug users, or homeless persons with severe and persistent mental problems (e.g., Ribisl et al. 1996). Successful strategies are expensive, require staff dedicated solely to the task, and of greatest significance, require researchers to leave the office, but they can also result in retention rates as high as 98% over two years with difficult-to-track populations.

Second, since some loss to outcome measurement will typically occur even with the best preventive efforts, researchers have developed sensitivity analyses about the potential effects of attrition on effect estimates (Shadish et al. 1998, Shih & Quan 1997). Results are encouraging, especially when effect sizes are not too small. Shadish et al. (1998) developed a method and computer program for examining the potential effects of attrition on effect estimates with dichotomous outcomes, varying assumptions about whether dropouts succeeded or failed for each condition. They applied the method to a set of randomized experiments on the effects of family therapy on substance abusers, a population that has high attrition from treatment. Results suggested the therapy would be successful under any plausible assumption about what happens to dropouts. Shih & Quan (1997) do similar kinds of analysis for continuous outcomes.

Because of all these developments, today we can field randomized experiments with much greater confidence that the results will be accurate despite problems that four decades ago would have spelled trouble. As a result of the policy emphasis on evidence-based practice, confidence is also higher that results will be useful in practice both conceptually and instrumentally.

Nested Designs

It is common in field experimentation for units to be nested within aggregated groups that are themselves nested within conditions. Examples include students nested within classrooms, clients nested within psychotherapy groups or psychotherapists, patients nested within physician practices, or workers nested within worksites. When aggregates (e.g., classrooms) rather than individuals (e.g., students) are assigned to conditions, the designs are often called group-randomized or cluster-randomized (Murray 1998). Regardless of whether aggregates or individuals are assigned to conditions, however, such nesting must be presumed to induce a dependency among nested units, which leads to violations of the independence assumptions of most ordinary statistics such as *t*-tests, analysis of variance, correlation, and regression. Such violations dramatically affect the Type I error rates (usually $\alpha = 0.05$), so that even the most modest dependencies could change α from 0.05

to as much as 0.20 to 0.50. Until the 1980s, no strong solution to this problem existed, so researchers tended to ignore it.

Progress has occurred in two areas. First is the invention and wide dissemination of statistical models and associated computer programs that appropriately analyze such data (Donner & Klar 1996, Goldstein 1986, Murray 1998, Raudenbush & Bryk 2002). The statistical models are variously called hierarchical linear models, multilevel models, or random coefficients models, and they are implemented in freestanding computer programs such as HLM (Raudenbush et al. 2004) and MLwiN (Rasbash et al. 2005) as well as popular computer packages such as SAS Proc Mixed (Littell 2006). Though these models have not been as widely adopted as they should have been, given the prevalence of nested data in field experimentation (e.g., Baldwin et al. 2005), they do solve the basic underlying statistical problem, yielding appropriate Type I error rates.

The second area of progress has been improving power in nested designs. The underlying problem here is that a proper analysis with a multilevel model often requires a large number of aggregate units (e.g., classrooms) to achieve the desired statistical power of 0.80. Because gathering data on, say, 50 classrooms per condition may be prohibitive for all but the best-funded research, investigators are faced with the choice between a properly analyzed study with too few aggregate units resulting in low power or a study improperly examined by an analysis that ignores nesting but that results in statistical significant results, albeit artifactually. This tradeoff encouraged many researchers simply to ignore the proper analysis. Yet it turns out that power can be improved, often dramatically, by using some simple design features such as matching or stratification, repeated measures, or most effectively, covarying a pretest measure that is highly correlated with the outcome measure. Exact results vary somewhat depending on the assumptions one makes about the intraclass correlation (the measure of how dependent units within aggregates really are) and the pretest-outcome correlation, but they suggest that as few as 10–12 classrooms per condition might yield desired power levels. The latter is logistically feasible for many more researchers.

Initially, all these developments in the analysis of nested models, partial treatment implementation, and attrition occurred independently. Today, models are being developed that incorporate more than one analysis. An example is Jo et al. (2008), who show how to combine proper nested design analysis with the Angrist et al. (1996) partial-treatment implementation analysis discussed above. Similar progress is occurring that combines these analyses with modern missing data analysis methods. Software for the latter may not yet be quite as widely available and transparent as that for the basic developments themselves, but rapid computerization seems likely.

Matching: equating two nonequivalent groups by selecting pairs of units with similar values on a variable correlated with outcome

REGRESSION DISCONTINUITY DESIGNS

Regression discontinuity designs (RDDs) are sometimes also called cutoff-based designs. They assign units to conditions based on a cutoff score on an ordered assignment variable, with units that fall on one side of the cutoff receiving treatment and those on the other side receiving the comparison condition. A regression is then fit to predict outcome from the assignment variable (minus the cutoff) and a treatment dummy variable. An effect is inferred if the regression line displays a discontinuity—a change in slope or intercept—at the cutoff between treatment and control. For example, the State of California passed legislation giving unemployment compensation to newly released prisoners, but only if they had worked more than 652 hours over the previous 12 months while in prison. Those who worked fewer hours than the cutoff value of 652 were ineligible. Berk & Rauma (1983) found that those receiving unemployment compensation had a recidivism rate 13% lower than controls.

First Goldberger (1972) and then Rubin (1977) showed that the regression discontinuity at the cutoff is an unbiased effect estimate under a proper analysis. Yet the design

was rarely used for 30 years after Thistlewaite & Campbell (1960) invented it. In the 1990s, however, economists and a few others began to use the design more often, taking advantage of many naturally occurring cases of assignment to treatment using a cutoff—for example, students to remedial writing training if they scored lower than a cutoff on a measure of writing skills (Aiken et al. 1998), or villages to receipt of social welfare assistance if they scored lower than a cutoff on a measure of village development (Buddelmeyer & Skoufias 2003). Combining the statistical developments from Rubin's potential outcomes model with standard econometric regression methods, they simultaneously made progress toward solving two problems with the RDD—modeling nonlinearities in the relationship between the assignment and outcome variables, and dealing with failures to assign to treatment by the cutoff score and nothing else. Once the design came back into vogue with these developments, researchers began to see opportunities to use it in many other naturally occurring situations where assignment is by cutoff (Cook 2007).

Modeling Nonlinearities

In RDD, the size of the effect is measured by the size of the discontinuity between treatment and control group regression lines at the cutoff. This estimate is unbiased only if the form of the relationship between the assignment and outcome variables is correctly modeled. The standard RDD analysis uses ordinary linear regression and so usually assumes a linear relationship. If the true relationship is nonlinear, though, a reliable discontinuity may be discovered that is an artifact of the wrong model rather than a true effect. This would happen, for example, if the true relationship is quadratic because those with higher assignment variable scores benefit more from treatment than those with lower scores—as when economically advantaged children learn more from Sesame Street than disadvantaged ones. Often there is no independent way to know the form of the true relationship. Standard advice to reduce the problem (e.g., Shadish et al. 2002, Trochim 1984) is (*a*) to examine the plot visually to try to identify nonlinearities; (*b*) to overfit the linear model by adding nonlinear functions of the assignment variable and interaction terms between assignment and outcome until all these terms are nonsignificant; (*c*) to compare the RDD data to preintervention data using the same variables, thus treating the preintervention regression as a control against which underlying nonlinearities can be distinguished from a treatment effect; and (*d*) to use nonparametric regression techniques that do not assume linearity. Economists particularly like the last solution (as well as the third one), which uses kernel density functions, lowess smoothers, and local linear regression techniques. These techniques are sensitive to exploring and modeling nonlinearities and allow estimating the causal impact locally at the cutoff point where an RDD estimate is most valid anyway. After all, this is where the treatment and comparison groups are most similar, sometimes differing only by measurement error in the assessment of the assignment variable. But nonparametric methods do entail additional assumptions of their own. Modern practice in the analysis of RDD emphasizes using all or most of the approaches listed above. It has become normal to present multiple estimates of the causal effect to test the sensitivity of results to the assumptions made about functional form. Robust results imply great confidence; more variable ones, less confidence. A good example of such recent practice can be found in Wong et al. (2008).

Fuzzy Regression Discontinuity

Effect estimates from RDDs are only unbiased if assignment is strictly by cutoff. This is similar to the assumption in randomized experiments that assignment adheres strictly to a chance process like the flip of a coin. The term "fuzzy regression discontinuity" refers to the case where some participants are assigned to conditions in violation of the cutoff, or where they cross over from their assigned condition to another condition. Biased estimates can then result. Just as

in a randomized experiment, the importance of such misassignment depends on its magnitude. Traditionally, the claim is that 5% fuzziness is unlikely to be much of a problem; but this is itself a fuzzy criterion, particularly when an effect is quite small. The likely severity of the problem can be diagnosed with an assessment of the amount of crossover, especially graphically around the cutoff. A precondition for diagnosis of severity is measurement of both the treatment intended for each unit and the treatment each unit actually received.

An ITT analysis should be done to assess the effects of the intended intervention irrespective of the received intervention. This unbiased ITT analysis is analogous to standard practice in a randomized experiment. However, when effects of TOT are of interest, a second unbiased analysis is possible with an RDD that takes advantage of the same instrumental variable analysis described previously for the randomized experiment (Angrist et al. 1996). As with the randomized experiment, the treatment dummy variable serves as the instrument for examining the effects of the treatment actually implemented. Examples of the analysis are in Angrist & Lavy (1999) and Wong et al. (2008), and the proof is in Hahn et al. (2001).

However, RDD must always be prospective in the sense that a cutoff is set and assignment made before the intervention is administered according to the side of the cutoff on which a unit falls. This also applies to experiments where random assignment must always precede treatment delivery. Otherwise, in both RDD and the experiment, one only has the kind of retrospective observational survey in which uncontrolled selection into conditions is the source of selection bias that is routinely much more problematic than when the assignment process is known.

Several studies have deliberately compared the causal results from randomized experiments to those from similar regression discontinuity designs (Aiken et al. 1998; D. Black, J. Galdo, & J.C. Smith, unpublished manuscript; Buddelmeyer & Skoufias 2003). The results are summarized by Cook and colleagues (Cook et al. 2008), illustrating reasonably close agreement in their estimates of the size and direction of effects. This agreement is not surprising theoretically, given the known conditions under which RDDs yield unbiased estimates. It is, however, gratifying given that RDDs, like any other design, are rarely implemented in perfect accord with the ideal and then are validated against randomized experiments that are also often not implemented as one might wish.

Still, both the analytic advances and the increased use of RDD in recent years suggest that, after a long hibernation, RDD has finally emerged as both a desirable and a feasible method for estimating causal effects (Cook 2007). It is especially useful, and now more widely used, when assignment is made on the basis of need or merit, and the latter can be measured for use as an assignment variable.

Time series: a quasi-experimental design that measures a single unit many consecutive times on the same outcome measure

INTERRUPTED TIME SERIES

A traditional interrupted time series (ITS) design has about 100 observations on one unit, during which a treatment is introduced at some known time. This large number of observations is needed in order to accurately identify the statistical model including cycles, autocorrelations, and trends. Similar to RDD, an effect is measured as a change in the slope or intercept of the time series at the point of treatment introduction. Also similar to RDD, correct identification of the functional form of the relationship between time and outcome is crucial. Analytic methods for analyzing such designs are well developed, but it is rare for researchers to have the opportunity to gather so many data points over time, although some daily diary methods do permit this. Thus, recent years have seen more interest and progress in the design and analysis of short interrupted time series, having say, 10–50 time points but all the other characteristics of traditional ITS.

A major realization has been the limited set of circumstances permitting interpretable simple ITS without a control series. Such simple designs are most interpretable when a series of

circumstances are on hand that rarely co-occur. These are basically an abrupt intervention at a known point in time, the immediate or very rapid onset of a response, and either a large effect or a pretest time series with very little random variation around a clear trend line. Reality teaches us that many interventions seep into units gradually rather than enter all of them at the same time and with high intensity, regardless of what the dissemination plan specified, and that responses can be delayed. This delay makes plausible some alternative explanations for the effect, such as other events that also occurred in the period after treatment was implemented. As a result, practice has moved toward ITS designs with a control series, whether created from nonintervention units or from nonequivalent dependent variables—those that the intervention should not affect but that other alternative causes should affect.

In some ways, the most well developed part of the short-term ITS literature concerns single case designs. These are particularly prevalent in some parts of clinical psychology (Marascuilo & Busk 1988, Morgan & Morgan 2001, Wampold & Worsham 1986), education (Phye et al. 2005), and medicine (McLeod et al. 1986, Weiss et al. 1980). They are sometimes called ABAB designs to indicate measurement of outcome over repeated alternations of baseline (A), treatment (B), withdrawal of treatment but continued assessment of the outcome (A), and reintroduction of treatment (B). The causal prediction is that two spikes occur in the response function soon after each occurrence of B is put into place.

Such designs have been well developed since the 1950s. What is new today is analytic models, especially the use of pooled time series (Hoeppner et al. 2008) and multilevel models that can often estimate treatment effects when many independent short-time series assessing the same intervention on the same outcome are available (Van den Noortgate & Onghena 2003). Such designs are more widely feasible when the research can recruit, for example, 20–40 schools that are each measured 4–8 times or 20 or 40 students each measured the same number of times. The multilevel analyses treat observations within units in the Level 1 model (e.g., within schools in one case or within persons in another), whereas unit variables (e.g., person or school characteristics) are treated at Level 2. The treatment indicator is a time-varying covariate at Level 1 because different cases can introduce and remove treatment at different times and can even have different numbers of total observations and missing data patterns. The resulting effect estimate is not known to be statistically unbiased. Instead, it relies on demonstrating that the effect is present when the treatment is present and is absent or at least reduced when the treatment is absent. To the extent that the intervention and its removal are applied at different times with different units, the possibility of a cyclical maturation pattern mimicking the theoretically expected pattern of results can be ruled out. The practical difficulty is failing to detect true causal relationships because of measurement artifacts and temporal persistence patterns. If the first B leads to a spike whose effect persists over time, it will then be difficult to detect a second spike; if removing the first intervention engenders special individual or social processes associated with its removal, then these too can affect data in the subsequent AB sequence and obscure obtaining data that correspond to the theoretically expected pattern.

Several analytic problems continue to be examined in the literature on single case studies and their aggregation. Of particular interest is how many independent series, and how many time points within those series, are needed to obtain powerful and accurate causal estimates. Some evidence suggests the analysis may be feasible with as few as 10 times series with 10–15 time points each when all trends are linear, though clearly more of both is better. Conversely, for the most complex analyses involving latent variables, hundreds of cases may be needed, though the number of time points can still be relatively small (duToit & Browne 2008). For example, the latter authors used 455 schoolchildren measured at only five points in time. The main challenge is with correct modeling of the form of trends over time, and control

time series clearly help with this in all contexts, including ABAB. A second challenge is modeling the error structure of the residuals, given that the assumption of independent errors is often violated in time series work. Most software for hierarchical linear modeling builds in some form of correction for nonindependent error, and economists frequently use a series of techniques, all predicated on a variety of assumptions. Though closure on functional forms and nonindependent errors is not yet upon us, the promise is sufficient that we expect more widespread use of short-term ITS for examining the effects of social interventions across, for example, schools over time or individuals over time What is clear, though, is that in most circumstances beyond the tightly controlled ABAB design, control ITS will be needed as well as intervention ITS. Most social interventions do not have an abrupt onset, the responses to them are not immediate, the expected effects are not large, and the preintervention slopes are too short and varied for the accurate assessment of functional form that provides the counterfactual.

NONEQUIVALENT COMPARISON GROUP EXPERIMENTS

The nonequivalent comparison group experiment design, which is characterized by both a pretest and a nonequivalent comparison group, is probably the workhorse design for causal inference in many substantive areas including education. Its low repute is due to its vulnerability to selection bias—treatment effects that are confounded with characteristics of units correlated with outcome. The advances in addressing this vulnerability have been significant in the past few decades, first in design and then in analysis.

The Importance of Good Design

Focal local controls. Decades of methodological advice stress the importance of control groups that are as similar as possible to the treatment group. Indeed, that is a key purpose of random assignment. Yet many observational studies, particularly in education and economics, fail to obtain similar control groups, for example, when a locally implemented treatment is compared with a patently nonlocal control group drawn from a national random sample (Hill et al. 2004). Borrowing a turn of phrase that Campbell (1976) once used in regard to a related issue in time series designs, the desirable control in a nonequivalent comparison group experiment is a focal local control group: in the same locale as the treatment group and focused on persons with the same kinds of characteristics as those in the treatment group, most particularly the characteristics that are most highly correlated with selection into conditions and with the outcome under investigation. Often control groups are one or the other, but not both. For instance, national random samples cannot be local for a locally conducted experiment, even if they approximate some of the desired focal characteristics. Conversely, local controls are not always focal when the units have characteristics demonstrably different from those in the treatment group, as when the quasi-experiment used by Cicirelli et al. (1969) compared local Head Start children who were disadvantaged enough to meet the eligibility criteria for Head Start with other local children who were not enrolled in Head Start because they were not as disadvantaged. A fundamental element of good quasi-experimental design is that a focal local control makes the job of estimating causal effects much easier from the start.

For example, Aiken et al. (1998) used both a randomized and a quasi-experiment to test the effects of a remedial writing program for new students registering at a university. Their control group was students who were also eligible for the program but who registered for classes too late to be in the randomized experiment. The control group's focal characteristics were plausibly similar to those eligible students who registered on time. Aiken and coworkers indeed showed that this control group did not differ significantly or substantially from the treatment group on a host of pretest measurements that

Nonequivalent comparison group experiment design: a quasi-experimental design with more than one condition and a posttest

Quasi-experiment: a design that manipulates the presumed cause and measures the presumed outcome but does not randomly assign participants to conditions

Propensity score: the conditional probability that a unit will be in the treatment or comparison condition given the available pretest covariates

were correlated with posttest. Of course, one can never be certain that the two groups are similar on all unobserved variables, but the claim here is that fewer selection biases are likely to be present than if the control were not locally or focally similar, such as students from another university or whose ACT scores made them ineligible for the program. In their study, Aiken et al. (1998) found that program effects from the randomized experiment were virtually identical to effects when the nonrandomized focal local control was substituted for the randomized control.

Matching on stable covariates. The quality of a nonequivalent control also can be improved by selecting one that is well matched to the treatment group on stable covariates, especially pretest measures of the outcome variable. Many of the early problems with matching in nonrandomized experiments were due to poor matching practices. In an early evaluation of Head Start by Cicirelli et al. (1969), groups were matched on unreliable measures of individual child achievement. High-achieving students in the Head Start group were matched with lower-achieving control group students, which was necessary because the control population was higher achieving in general. Systematically higher levels of positive measurement error in the treatment group scores and of negative measurement error in the control group scores at pretest were less positive or negative by chance at posttest, causing the two group scores to regress to different means that made it appear that Head Start hurt children. However, reanalysis correcting for unreliability in the matching variables showed that this finding was not correct (Magidson 1977).

The lesson is that it it important to match on variables that are less likely to contain much measurement error. One way to do so is to match on composite variables because adding equivalent items to a measure will increase its reliability. An example of this kind of matching is an evaluation of the Comer whole-school reform program in Detroit public schools by Millsap et al. (2000), who wanted to create controls matched on pretest student achievement levels. To increase the reliability of this matching variable, the investigators matched on achievement test scores that were aggregated across individual students to the school level and also aggregated over several years of pretest data. Such multiyear average school-level scores tend to be extremely reliable. In this study, the control schools were also focal local controls in the sense discussed above. They were from similar parts of Detroit, with similar school-level and individual student characteristics.

We stress two characteristics that are needed for stable matching. One is the reduction of measurement error described above, but the other is matching on a variable that either is a pretest measure of the outcome or is another variable that is as highly correlated as possible with that outcome. Measurement error is not the only cause of regression to the mean (Campbell & Kenny 1999). Such regression will occur any time two variables are imperfectly correlated—the person who is the tallest will not necessarily be the heaviest even though both variables are measured with extremely high reliability. In the absence of a treatment effect, school-level pretest achievement scores are likely to be very highly correlated with school-level posttest achievement scores, further reducing any potential impact of regression artifacts. Combining matching on stable covariates from a pool of focal local controls can greatly improve the chances of correctly estimating effects.

Propensity Score and Other Analyses

The previous two sections emphasize the importance of good design in nonrandomized studies. Another tradition emphasizes statistical adjustments to nonrandomized results so the resulting effects might better approximate those from randomized experiments. Such adjustments almost certainly work best when used with studies that are well designed at the start, especially using the design features described above. Currently, the most popular adjustment is propensity score analysis.

Propensity score analysis is another development from Rubin's potential outcomes model. Propensity scores are the predicted probabilities of being in treatment or control given the available covariates. They are usually constructed using logistic regression, but a host of other methods can be used. Groups that are balanced on the true propensity scores are also balanced on all the covariates used to create those propensity scores, where balance is assessed using methods described in Rubin (2001). However, because the true propensity scores are unknown and must be estimated, assessing balance on the propensity scores must be supplemented with assessing balance on the individual covariates as well. The idea is to mimic the balance achieved by random assignment between groups, where groups are equivalent on all measured and unmeasured covariates. The difference is that propensity score matching only succeeds in matching on the measured covariates, making balance a necessary but not sufficient condition for unbiased effects. Hence the use of propensity scores also requires the assumption of strong ignorability, that potential outcomes are orthogonal to treatment assignment conditional on the observed variables, or in other words, that there are no unmeasured variables that cause a hidden bias. If the strong ignorability assumption is not met, propensity score matching should not produce the same results as a randomized experiment; conversely, if the strong ignorability assumption is met, it should produce the same results. Strong ignorability requires measurement of all covariates related to both treatment and outcome. A key need is to develop measures indicating how well this assumption is met. In the meantime, when data are clearly imbalanced at the start, assessing how well available covariates predict both assignment and outcome is a weak but necessary fallback. It is weak because standards for how much prediction is enough are lacking; it is necessary because demonstrably poor prediction is a cause for concern—in such cases, propensity score adjustments can increase rather than decrease bias.

To illustrate, Shadish et al. (2008) randomly assigned participants to be in a randomized or nonrandomized experiment. Those in the randomized experiment were randomly assigned to mathematics or vocabulary training, while the remaining students got to choose their training. Participants were otherwise treated identically and simultaneously and were not aware of the manipulation of assignment method. Shadish et al. (2008) created propensity scores from a very extensive set of pretest covariates that might predict treatment choice (e.g., math anxiety and preference) and outcome (e.g., math and vocabulary skills at pretest). They were then able to reproduce closely the results from the randomized experiment using several different kinds of propensity score analyses, and even using simple linear regression of the original covariates without propensity scores. However, they also showed that propensity scores based on a weak set of predictors (i.e., gender, age, ethnicity, and marital status) failed to reduce bias much, or sometimes increased it. Subsequent reanalyses of the data using an indirect test of strong ignorability—one that was possible only because the yoked randomized experiment was present—suggested convincingly that balance was not sufficient for bias reduction but that balance plus strong ignorability was sufficient (Steiner et al. 2008).

Summaries of Empirical Evidence

A considerable amount of empirical literature compares results from nonrandomized to randomized experiments. Some reviews of that literature have questioned whether statistical adjustments can reproduce results from randomized experiments (Glazerman et al. 2003). Those results are questionable because the comparisons were commonly quite poor in many ways. Cook and colleagues (Cook et al. 2008) suggested that good comparisons of experimental and quasi-experimental estimators must meet seven criteria:

1. The studies must compare one randomly formed control group and one

nonrandomly formed control group. Without this, no comparison can be made.

2. The randomized and nonrandomized experiment should both estimate the same estimator, be it treatment on treated or intent to treat, because otherwise agreement in results would not be expected.
3. The randomized and nonrandomized groups should differ from each other only in assignment method, for otherwise any differences in estimators might be due to confounds rather than assignment method.
4. The person producing the nonrandomized estimate of effect should not know the results from the randomized experiment or else they may keep searching for a nonrandomized estimator until they find one that matches the randomized one.
5. The randomized experiment should be an exemplar of its kind, not subject to large attrition or partial treatment implementation problems that statistical adjustments such as propensity scores are not designed to fix.
6. The nonrandomized design should similarly be an exemplar of its kind, similarly without attrition or partial treatment implementation problems, with focal local controls and good pretest measurement of variables related to treatment and outcome. Otherwise, we would not expect it to match a randomized result.
7. A defensible standard for what counts as a match in randomized and nonrandomized results is used. This is difficult both because reasonable people might disagree on substantive criteria that would make a difference to policy decisions and because statistical criteria will inevitably be subject to power problems.

This is a stringent set of criteria, only fully met in one comparison of random and nonrandom assignment, but met substantially by several more experiments. Cook and colleagues (Cook et al. 2008) showed that when most or all of these seven criteria were met, results from various different kinds of nonrandomized experiments yielded a reasonable match to results from randomized experiments. This was true for regression discontinuity designs, well-designed nonrandomized experiments with focal local controls and stable matching, and statistical analyses such as propensity scores. A key conclusion is, of course, that it is essential to use good design on the front end, on the presumption that good design limits the amount of bias needing to be reduced by any such adjustment.

PATTERN MATCHING DESIGNS

Sometimes none of the preceding options is feasible, and even when the options are feasible, the researcher may wish to improve causal inference by supplementing the study design. Several design strategies have proven useful in producing plausible causal inferences. The set of those strategies is termed "pattern matching designs" to emphasize that they combine various design features so as to produce multiple probes of a causal hypothesis that inform different threats to internal validity without all sharing the same threat. When such probes all converge on the same effect, the plausibility of a causal inference is increased due to the more numerous alternative interpretations informed and the absence of shared bias.

A good example of a pattern matching design is the Reynolds & West (1987) study of the effects of Arizona's "Ask for the Sale" campaign to sell lottery tickets. Participating stores selling lottery tickets agreed to post a sign reading, "Did we ask you if you want a lottery ticket? If not, you get one free"; they also agreed to give a free ticket to those customers whom they neglected to ask the question but who then requested one. Because participation was voluntary, the resulting nonequivalent control group design was supplemented in four ways. First, the authors matched treatment to control stores from the same chain (and where possible, the same zip code) as well as on the pretest market share of ticket sales. Second, they added multiple pretest and posttest

assessments by examining mean weekly ticket sales for four weeks before and four weeks after the treatment started. They then observed that pretest sales trends were decreasing nearly identically in both the treatment and control groups, and so maturation differences could not explain increasing ticket sales. Similarly, regression to the mean was unlikely because the treatment group sales were continuously decreasing over four consecutive pretests, and because control group ticket sales continued to decrease after treatment began. Third, Reynolds & West (1987) studied treatment effects on three nonequivalent dependent variables in the treatment group, discovering that the intervention increased ticket sales but not sales of gas, cigarettes, or grocery items. Fourth, they located some stores in which the treatment was removed and then repeated, or was initiated later than in other stores, and found that the outcome tracked the introduction, removal, and reinstatement of treatment over time while sales in the matched controls remained unchanged. Nearly all of these analyses suggested that the "Ask for the Sale" intervention increased ticket sales after the program began, making it difficult to think of an alternative explanation for the effect.

Pattern matching designs counter the unfortunate notion that researchers should choose from a small and fixed set of designs. Campbell & Stanley (1963) inadvertently encouraged such oversimplifications with their table of simple designs and associated plus and minus signs indicating which validity threats are controlled—which they themselves acknowledged at the end of their text might be so simple that they mislead. Instead, pattern matching designs attend to a less often noticed piece of advice, to predict a diverse pattern of results whose strong testing might require multiple nonrandomized designs each with different presumed biases, "the more numerous and independent the ways in which the experimental effect is demonstrated, the less numerous and less plausible any singular rival invalidating hypothesis becomes" (Campbell & Stanley 1963, p. 206). The spirit of this recommendation is the same as that of R.A. Fisher, who advised to "make your theories elaborate" (cited in Rosenbaum 1984, p. 41) in order to improve causal inference from nonrandomized experiments.

Given such a pattern matching logic, statistical analyses are required that test the overall fit of all the hypothesis tests, not just the difference between adjacent means as in the simple designs. But such tests are not as well developed as those for testing the difference among a small number of means. It may be that testing effects in pattern matching designs requires an approach more resembling meta-analysis, such as combined probability tests. This is a topic needing considerable attention.

CONCLUSION

Great progress has been made in the past four decades about how to solve the implementation and technical problems associated with experiments. In this sense, it is today possible to entertain a vision in which we understand the conditions under which many different kinds of experimental methods, both randomized and nonrandomized, are known to yield accurate effect estimates. Such a vision was hard to conceive four decades ago, but its very possibility points to the extent of the advances that have been made. Scientific progress of this potential magnitude takes decades of effort, persistent research attention, and perhaps a partly fortuitous combination of technical developments.

That being said, we do not claim that the vision has been achieved in practice, on several counts. Randomized experiments remain the method of choice, all things being equal and when they are feasible and ethical, for their strong statistical warrant has no peer among the nonrandomized alternatives. The conditions under which those nonrandomized alternatives seem to produce accurate results are only tentatively understood and place demands on the researcher for good measurement and careful attention to design that cannot be routinely assumed to occur in practice. Many researchers will not know how to implement

the best approaches to analysis of nonrandomized experiments or will implement them under conditions of poor pretest measurement that probably yield poor results. And a great deal more empirical research is needed to tease out the nuances of good quasi-experimental design and analysis in a manner that further clarifies the conditions under which they work better or worse. It may take another several decades of that kind of work before we can know how much confidence we should place in nonexperimental estimators.

Also, this review primarily addresses progress in the technical aspects of experimentation. These are crucial developments in many ways. Yet we must remember that field experimentation exists in a policy context. As we said at the start of this review, one of the factors contributing to the renaissance of field experimentation is the evidence-based practice movement. The presumption is that the policy environment today is more open to influence based on information about what works. Yet this presumption badly needs empirical examination. A key lesson of the early rounds of experiments in the 1970s was indeed that policy responds surprisingly little to information about what works (Shadish et al. 1991). The current environment does indeed seem to produce information about what works, information that is of much higher quality than was the case in the 1970s. Much of that information is produced in response to mandates, however, leaving open the possibility that the information will again have little impact on practice and policy. We know of few researchers seriously investigating this possibility.

Fortunately, researchers interested in studying the use of experimental results do not have to start from scratch either conceptually or empirically. Program evaluators have studied the use of evaluation results extensively, developed conceptual systems to help us understand the kinds of uses that can occur and the conditions that facilitate each of these uses. The literature is more extensive than we can summarize here, but good starting points for those interested in such research are in Shadish et al. (1991), Shulha & Cousins (1997), Weiss (1998), and Weiss et al. (2008). The most likely findings will be that short-term instrumental use of experimental evidence to change policy are rare but do occur, especially in cases where the intervention is of interest to the practitioner and has a naturally high turnover rate so that substitution occurs easily. Nevertheless, these results will have significant impact both on how policymakers and practitioners think and on the education received by their replacements that are still in training. In the end, though, significant empirical investigation of such uses is the only way to know for sure if current investments in experimental work really do lead to evidence-based practice.

We have been involved in reviewing the state of social experimentation for more than 20 years (e.g., Cook & Shadish 1986, 1994; W.R. Shadish, manuscript submitted; Shadish et al. 1991). It is fair to say that the present review leaves us more optimistic about the state of experimentation, both technically and socially, than at any time before. We hope—indeed, anticipate—that a review that is written 10 to 20 years from now will yield even more encouraging conclusions.

SUMMARY POINTS

1. A round of early social experiments in the 1960s and 1970s led to disappointing results given technical problems and a perceived failure to use the results. Some fields stopped using experiments as a result.
2. Experimentation is experiencing a renaissance today due to the emphasis on evidence-based practice, the increased involvement of economists and statisticians, and the growing empirical literature that clarifies the effects of certain design and analytic practices.

3. Progress in randomized experiments has involved increased understanding of how to implement them well, better analytic methods for coping with partial treatment implementation, and more knowledge of how to prevent attrition and bracket its possible effects.
4. Progress in regression discontinuity designs has come from vastly increased usage of the design in the past decade by economists who have provided new analytic methods for how to model nonlinearities and how to cope with violations of assignment by cutoff score.
5. Evidence suggests that nonequivalent comparison group designs can approximate answers from randomized designs when they use focal local controls, careful matching on stable covariates, and measure a rich set of pretest predictors of treatment and outcome that can be used in statistical adjustments such as propensity score analysis.
6. When none of the other designs is feasible, researchers should assemble more than one design that predicts a pattern of causal results, for if the results match the predicted pattern, there are fewer plausible alternative explanations.
7. We have made great progress toward understanding the conditions under which a wide variety of kinds of randomized and nonrandomized designs are known to yield accurate effect estimates.

FUTURE ISSUES

1. We need much more knowledge of the conditions under which propensity score analysis can be useful. What sample sizes are needed? How can we best measure the strong ignorability assumption? What approach to missing data in the pretest covariates is best?
2. Methods for the analysis of short interrupted time series need continued development. Especially needed is work on producing estimates of effects that are in the same metric as estimates from between-group designs so that we can know how well the former approximate the latter.
3. The concept of a focal local control needs clarification so that researchers can better know when such a control is present and likely to be comparable to a nonrandomized treatment group.
4. Methods for the analysis of pattern-matching designs have received virtually no attention and are badly needed.
5. Propensity score analysis would benefit from ways to test the strong ignorability assumption.
6. The quality of studies comparing results from randomized and nonrandomized experiments needs improvement in general, and much more of this kind of research needs to be done.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by grants H324U050001-06 and R305U07003/01 from the U.S. Department of Education, Institute of Education Sciences.

LITERATURE CITED

Aiken LS, West SG, Schwalm DE, Carroll JL, Hsiung S. 1998. Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: efficacy of a university-level remedial writing program. *Eval. Rev.* 22:207–44

Angrist JD, Imbens GW. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.* 90:431–42

Shows how to obtain an unbiased estimate of treatment on the treated in the presence of partial treatment implementation.

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Angrist JD, Lavy V. 1999. Using Maimonides' rule to identify the effects of class size on scholastic achievement. *Q. J. Econ.* 114:533–75

Baldwin SA, Murray DM, Shadish WR. 2005. Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. *J. Consult. Clin. Psychol.* 73:924–35

Balke A, Pearl J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92:1171–76

Barnard J, Du J, Hill JL, Rubin DR. 1998. A broader template for analyzing broken randomized experiments. *Sociol. Methods Res.* 27:285–317

Berk RA, Rauma D. 1983. Capitalizing on nonrandom assignment to treatment: a regression discontinuity evaluation of a crime control program. *J. Am. Stat. Assoc.* 78:21–27

Boruch RF. 1997. *Randomized Field Experiments for Planning and Evaluation: A Practical Guide*. Thousand Oaks, CA: Sage

Botein B. 1965. The Manhattan Bail Project: its impact on criminology and the criminal law process. *Texas Law Rev.* 43:319–31

Buddelmeyer H, Skoufias E. 2003. *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA*. Bonn, Germany: IZA

Campbell DT. 1976. Focal local indicators for social program evaluation. *Soc. Indicators Res.* 3:237–56

Campbell DT, Kenny DA. 1999. *A Primer on Regression Artifacts*. New York: Guilford

Campbell DT, Stanley JC. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally

Summarizes the history, repeated reinvention, and current revival of the regression discontinuity design.

Cicirelli VG. 1969. *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development. Vol. 1 & 2. A Report to the Office of Economic Opportunity*. Athens: Ohio Univ. Westinghouse Learning Corp.

Constas MA. 2007. Reshaping the methodological identity of education research: early signs of the impact of Federal policy. *Eval. Rev.* 31:391–400

Cook TD. 2007. "Waiting for life to arrive": a history of the regression-discontinuity design in psychology statistics and economics. *J. Econometrics* 142:636–54

Cook TD, Shadish WR. 1986. Program evaluation: the worldly science. *Annu. Rev. Psychol.* 37:193–232

Cook TD, Shadish WR. 1994. Social experiments: some developments over the past 15 years. *Annu. Rev. Psychol.* 45:545–80

Summarizes evidence that well-done nonrandomized experiments can approximate results from randomized experiments.

Cook TD, Shadish WR, Wong VC. 2008. Three conditions under which experiments and observational studies often produce comparable causal estimates: new findings from within-study comparisons. J. Policy Anal. Manag. In press

Cronbach LJ. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey-Bass
Cronbach LJ, Ambron SR, Dornbusch SM, Hess RD, Hornik RC, et al. 1980. *Toward Reform of Program Evaluation*. San Francisco, CA: Jossey-Bass
Donner A, Klar N. 1996. Statistical considerations in the design and analysis of community intervention trials. *J. Clin. Epidemiol.* 49:435–39
duToit SHC, Browne MW. 2008. Structural equation modeling of multivariate time series. *Multivariate Behav. Res.* 42:67–101
Efron B, Feldman D. 1991. Compliance as an explanatory variable in clinical trials. *J. Am. Stat. Assoc.* 86:9–26
Elkin I, Parloff MB, Hadley SW, Autry JH. 1985. NIMH Treatment of Depression Collaborative Research Program: background and research plan. *Arch. Gen. Psychiatry* 42:305–16
Elkin I, Shea T, Watkins JT, Imber SD, Sotsky SM, et al. 1989. National Institute of Mental Health Treatment of Depression Collaborative Research Program: general effectiveness of treatments. *Arch. Gen. Psychiatry* 46:971–82
Fischer-Lapp K, Goetghebeur E. 1999. Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Control. Clin. Trials* 20:531–46
Fraker T, Maynard R. 1987. Evaluating comparison group designs with employment-related programs. *J. Hum. Resour.* 22:194–227
Glass GV. 1976. Primary, secondary and meta-analysis. *Educ. Res.* 5:3–8
Glazerman S, Levy DM, Myers D. 2003. Nonexperimental versus experimental estimates of earnings impacts. *Ann. Am. Acad. Pol. Soc. Sci.* 589:63–93
Goetghebeur E, Molenberghs G. 1996. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *J. Am. Stat. Assoc.* 91:928–34
Goldberger AS. 1972. *Selection bias in evaluating treatment effects: some formal illustrations*. Discuss. pap. #123, Inst. Res. Poverty, Univ. Wisc., Madison
Goldstein H. 1986. Multilevel mixed linear model analysis using generalized least squares. *Biometrika* 73:43–56
Gueron JM. 1985. The demonstration of state work/welfare initiatives. In *Randomization and Field Experimentation*, ed. RF Boruch, W Wothke, pp. 5–13. San Francisco, CA: Jossey-Bass
Hahn J, Todd P, Van Der Klaauw W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69:201–9
Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61
Heckman JJ. 1992. Randomization and social policy evaluation. In *Evaluating Welfare and Training Programs*, ed. CF Manski, I Garfinkel, pp. 201–30. Cambridge, MA: Harvard Univ. Press
Heinsman DT, Shadish WR. 1996. Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychol. Methods* 1:154–69
Hill JL, Reiter JP, Zanutto EL. 2004. A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*, ed. A Gelman, X-L Meng, pp. 51–60. New York: Wiley
Hoeppner BB, Goodwin MS, Velicer WF, Heltshe J. 2008. An example of pooled time series analysis: cardiovascular reactivity to stressors in children with autism. *Multivariate Behav. Res.* 42:707–27
Imbens GW, Rubin DB. 1997a. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Stat.* 25:305–27
Imbens GW, Rubin DB. 1997b. Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* 64:555–74
Jo B, Asparouhov T, Muthén BO, Ialongo NS, Brown CH. 2008. Cluster randomized trials with treatment noncompliance. *Psychol. Methods* 13:1–18
LaLonde R, Maynard R. 1987. How precise are evaluations of employment and training experiments: evidence from a field experiment. *Eval. Rev.* 11:428–51
Littell RC. 2006. *SAS for Mixed Models*. Cary, NC: SAS Publ.
Little RJ, Yau L. 1996. Intent-to-treat analysis for longitudinal studies with dropouts. *Biometrics* 52:1324–33
Magidson J. 1977. Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation. *Eval. Q.* 1:399–420
Marascuilo LA, Busk PL. 1988. Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behav. Assess.* 10:1–28

Provides methods for estimating effects in regression discontinuity designs in the face of violations of the cutoff.

McLeod RS, Taylor DW, Cohen A, Cullen JB. 1986. Single patient randomized clinical trial: its use in determining optimal treatment for patient with inflammation of a Kock continent ileostomy reservoir. *Lancet* 1(March 29):726–28

Millsap MA, Chase A, Obeidallah D, Perez-Smith AP, Brigham N, Johnston K. 2000. *Evaluation of Detroit's Comer Schools and Families Initiative: Final Report*. Cambridge, MA: Abt Assoc.

Morgan DL, Morgan RK. 2001. Single participant research design: bringing science to managed care. *Am. Psychol.* 56:119–27

Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference*. London: Cambridge Univ. Press

Murray DM. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford Univ. Press

Nathan RP. 1989. *Social Science in Government: Uses and Misuses*. New York: Basic Books

Oakes D, Moss AJ, Fleiss JL, Bigger JT Jr, Therneau T, et al. 1993. Use of compliance measures in an analysis of the effect of diltiazem on mortality and reinfarction after myocardial infarction. *J. Am. Stat. Assoc.* 88:44–49

Phye GD, Robinson DH, Levin J, eds. 2005. *Empirical Methods for Evaluating Educational Interventions*. New York: Academic

Rasbash J, Steele F, Browne WJ, Prosser B. 2005. *A User's Guide to MLwiN Version 2.0*. Bristol, UK: Univ. Bristol

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage

Raudenbush SW, Bryk A, Cheong YF, Congdon R. 2004. *HLM6: Hierarchical Linear and Nonlinear Modeling*. Chicago: Sci. Software Intl.

Reynolds KD, West SG. 1987. A multiplist strategy for strengthening nonequivalent control group designs. *Eval. Rev.* 11:691–714

Presents the best available summary of methods for preventing attrition.

Ribisl KM, Walton MA, Mowbray CT, Luke DA, Davidson WS, Bootsmiller BJ. 1996. Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: review and recommendations. *Eval. Program Planning* 19:1–25

Riecken HW, Boruch RF, Campbell DT, Caplan N, Glennan TK, et al. 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. New York: Academic

Robins JM. 1998. Correction for noncompliance in equivalence trials. *Stat. Med.* 17:269–302

Rosenbaum PR. 1984. From association to causation in observational studies: the role of tests of strongly ignorable treatment assumptions. *J. Am. Stat. Assoc.* 79:41–48

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:141–55

Rossi PH, Lyall KC. 1978. An overview evaluation of the NIT experiment. In *Evaluation Studies Review Annual*, ed. TD Cook, ML DelRosario, KM Hennigan, MM Mark, WMK Trochim, 3:412–28. Newbury Park, CA: Sage

Provides an exemplar of how to assess balance using propensity scores in observational studies.

Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701

Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2:1–26

Rubin DB. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* 2:169–88

Furnishes a good general introduction to Rubin's potential outcomes model.

Rubin DB. 2004. Teaching statistical inference for causal effects in experiments and observational studies. *J. Educ. Behav. Stat.* 29:343–67

Shadish WR. 2000. The empirical program of quasi-experimentation. In *Validity and Social Experimentation: Donald Campbell's Legacy*, ed. L Bickman, pp. 13–35. Thousand Oaks, CA: Sage

Shadish WR. 2008. *Bandwidth versus fidelity: working through commensurability of Campbell and Rubin for causal inference*. Manuscr. submitted

Shadish WR, Clark MH, Steiner PM. 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *J. Am. Stat. Assoc.* In press

Serves as a good general reference to many aspects of experimental and quasi-experimental design.

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin

Shadish WR, Cook TD, Leviton LC. 1991. *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, CA: Sage

Shadish WR, Hu X, Glaser RR, Kownacki RJ, Wong T. 1998. A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychol. Methods* 3:3–22

Shadish WR, Ragsdale K. 1996. Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *J. Consult. Clin. Psychol.* 64:1290–305

Shih WJ, Quan H. 1997. Testing for treatment differences with dropouts present in clinical trials—a composite approach. *Stat. Med.* 16:1225–39

Shulha LM, Cousins JB. 1997. Evaluation use: theory, research and practice since 1986. *Am. J. Eval.* 18:195–208

Steiner PM, Cook TD, Shadish WR, Clark MH. 2008. *The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies.* Manuscr. submitted

Thistlewaite DL, Campbell DT. 1960. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.* 51:309–17

Trochim WMK. 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Newbury Park, CA: Sage

Van Den Noortgate W, Onghena P. 2003. Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behav. Res. Methods Instrum. Comput.* 351:1–10

Wampold BE, Worsham NL. 1986. Randomization tests for multiple-baseline designs. *Behav. Assess.* 8:135–43

Weiss B, Williams JH, Margen S, Abrams B, Caan B, et al. 1980. Behavioral responses to artificial food colors. *Science* 207:1487–89

Weiss CH. 1998. Have we learned anything new about the use of evaluation? *Am. J. Eval.* 19:21–33

Weiss CH, Murphy-Graham E, Petrosino A, Gandhi AG. 2008. The Fairy Godmother—and her warts: making the dream of evidence-based policy come true. *Am. J. Eval.* 29:29–47

Wholey JS, Scanlon JW, Duffy HG, Fukumoto JS, Vogt LM. 1970. *Federal Evaluation Policy: Analyzing the Effects of Public Programs*. Washington, DC: Urban Inst.

Wong VC, Cook TD, Barnett WS, Jung K. 2008. An effectiveness-based evaluation of five state prekindergarten programs. *J. Policy Anal. Manag.* 27:122–54

Annual Review of Psychology

Volume 60, 2009

Contents