5

RANDOMIZED EXPERIMENTS AND QUASI-EXPERIMENTAL DESIGNS IN EDUCATIONAL RESEARCH

Peter M. Steiner, Angela Wroblewski, and Thomas D. Cook

Introduction

Since the 1960s, nearly all highly industrialized societies have sought to improve the performance of school systems. Measures taken to support that goal are many and diverse, including evaluating what these educational reform efforts have achieved. So causal investigations are central to educational evaluation, and the main issue is: What form should these evaluations take? Are randomized experiments still the gold standard in causal inference and are quasi-experimental designs as good as randomized experiments? This chapter deals mainly with the use of randomized experiments to assess causal efficacy and effectiveness (Flay, 1986), but also considers some of the strongest quasi-experimental designs (Shadish, Cook, & Campbell, 2002). Although quasi-experimental designs are often recommended for educational evaluations, their empirical justification is inferior to that of the experiment. Within-study comparisons have shown that quasi-experiments regularly fail to reproduce experimental results (Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Myers, 2003) unless the assignment mechanism into treatment is completely known (regression discontinuity design) or extensively and reliably measured. Examples of quasi-experiments meeting this standard include nonequivalent control group designs with plausible theories of selection into treatment versus control states and extensive and reliable measurements of this selection process (Shadish, Clark, & Steiner, 2008; Steiner, Cook, Shadish, & Clark, under review). But even in these cases, randomized experiments are still more efficient and rely on fewer and clearer assumptions than quasi-experimental methods. Policymakers and evaluators in the fields of education are well advised to stick to experiments whenever possible. However, if experiments cannot be conducted, strong quasi-experimental designs are still possible, and we outline the best warranted of them. A slight trend toward quasi-experimental methods has recently been observed. The American Educational Research Association (AERA) has edited a book outlining strong quasi-experiments (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007), and even the Institute for Educational Sciences now supports more use of regression discontinuity and matched group designs. This also holds for the European Union (EU; European Commission, 2004). Finally, the increasing international coordination in educational planning and policy, which has proved so effective in descriptive monitoring like in PISA and TIMMS, will likely be extended to cover summative causal outcome evaluation. This will probably mean closer coordination of experimental and quasi-experimental practices across nations.

The chapter is organized as follows. After a brief description of the underlying concept of causation, the justifications of randomized experiments as well as quasi-experimental designs are discussed. Then we focus on the reasons that experimental designs are relatively rare in educational evaluation. First, we consider the role of experimentation in the context of the evaluation tradition, particularly in the United States and the EU. Then reservations about using experiments in educational evaluations are discussed in some detail, thereby revealing the strengths and limitations of randomized experiments.

Causation

In modern sciences, the notion of causality has been strongly affected by David Hume and John Stuart Mill. Hume discussed three main conditions for causation: (a) The cause and effect have to be in spatial and temporal contiguity, (b) the cause should occur prior to the effect, and (c) the cause and effect are constantly conjoined (i.e., they are perfectly correlated). The dependence of causation on counterfactuals also goes back to Hume, who asked: What would have happened if the cause had not been there—for instance, how student performance might have been without a specific intervention? The implication is that a causal effect can only be claimed with reference to some kind of a control condition. Mill took up some of the same ideas as Hume, but in a way that pointed more explicitly toward the necessity for linking cause to variation in what happens when a cause is present or absent (hence to control groups) and to the advantages of studying active intrusions into an ongoing process. According to Mill, a causal relationship may exist if (a) the cause precedes the effect, (b) the cause is related to the effect, and (c) no plausible alternative explanation for the effect exists other than the cause (i.e., all other competing causes can be ruled out). Active intrusion helps with the first and last of these conditions.

The clear identification of causal relationships is difficult because a given effect may be produced by different causes or by a complex interplay of multiple causes. It may also depend on specific conditions in the setting where the research takes place. In practice, we cannot identify all of these other causal circumstances especially as concerns how they relate to each other. Therefore, it might be more accurate to refer to the causes that educational evaluators typically study as *inus* conditions—as "an *insufficient* but *nonredundant* part of an *unnecessary* but *sufficient* condition"—(Mackie, 1974, p. 62; italics original). To envisage this, consider increasing the number of school days per year (cause) in order to improve student performance (effect). In some applications, more school days can increase performance and so is sufficient for it. But it is not necessary for a performance increase because other mechanisms can achieve this end. Even in a particular application that increases performance, to add days does not by itself cause performance to rise because the children also have to be attentive and the additional time also has to be spent in effective instruction. However, increasing the number of days is not the same as creating attention or having more effective instruction, and so it is a nonredundant part of what is sufficient for increasing performance. The implication here is that a full explanation of any causal relationship is necessarily context-dependent and that many factors are usually required for a given cause-effect relationship to occur. This renders causation a probabilistic rather than a deterministic concept. A cause (i.e., *inus* condition) does not always lead to an effect; it merely increases the probability that it will occur.

During the 20th century, statisticians—with Neyman (1923/1990), Rubin (1974, 1978, 1986), and Holland (1986) at the forefront—developed a formal model of causation that is closely related to experimentation. It is now generally known as the Rubin Causal Model (RCM). Rubin explicitly

defined the causal effect as the difference between what would have happened, for instance, to students under the treatment condition and what would have happened to these same students in the counterfactual or control situation (i.e., without intervention, but under identical circumstances). RCM can be characterized by three main characteristics (Holland, 1986). First, it refers to the *effect of a cause* and not to the cause of a given effect. Second, the effect of a cause is always *relative* to another cause—the counterfactual, so that both a treatment and contrast are required to define a cause-effect relation. Third, only *manipulable* events can be a cause ("no causation without manipulation"; Holland, 1986). In this theory, events and attributes that cannot be manipulated in practice or in theory (e.g., weather or student's age and sex) cannot be causes.

This RCM theory leads to a fundamental problem of causal inference: It is not possible to expose a student, class, or school to the treatment and control condition under exactly the same circumstances at exactly the same time. So the most appropriate counterfactual is not possible—the same unit (e.g., the same student) being simultaneously exposed to both the treatment and control condition. Nonetheless, three main research strategies are commonly considered to justify causal inference, although each requires assumptions:

1. Within-subject designs measure each experimental unit under the treatment and control condition, not simultaneously but successively. Unfortunately, within-subject designs are rarely useful in education unless they are linked to other methods for enhancing causal conclusions. This is primarily because students typically mature over time, and this maturation is confounded with treatment effects, particularly when performance is measured only at pretest and again at posttest. Moreover, it is not reasonable to expose someone to a control condition after a treatment one if there is any reason to expect that the original treatment effect will persist over time.

2. Matching designs seek to pair nearly identical units before or after assigning them to the treatment and control conditions. But it has not yet been possible to identify matching techniques that consistently re-create the same results as experiments. This is primarily because of the possibility of unmeasured differences between groups that might be correlated with the outcome (for a summary, see Glazerman et al., 2003).

3. Random assignment seeks to create treatment and control groups that, on average, are identical with respect to all measured and unmeasured variables save the treatment they receive. Randomized experiments, also known as randomized clinical or controlled trial (RCT), have a clear theoretic warrant in statistics, and they are routinely used today in many sectors other than education, in clinical or agricultural research, for instance. If certain transparent assumptions are met, we know that randomized experiments generate unbiased causal inference.

Random assignment entails using the equivalent of a fair coin toss to create two or more initially equivalent groups. The intervention under consideration is then assigned to the treatment group, whereas the control group is exposed to something else—often no explicit treatment, but sometimes a qualitatively different one. Randomization ensures that, prior to treatment, both groups will be on average equal in all measured and unmeasured variables. Consequently, if an experiment is properly implemented initially and then maintained over time, any observed group differences at

the end of a study can be reasonably attributed to the intervention, they are not likely to be due to selection, thanks to the random assignment process. But this particular counterfactual is not perfect. *Individual* causal effects cannot be estimated, only *average* ones—thanks to the random assignment equating treatment and control groups on average. In Rubin's conceptualization, the causal effect of an experiment is defined by the difference between the average outcome of the treatment and control groups.

RCM offers us a clear but restrictive formal conceptualization of cause. It focuses solely on causal description (i.e., ascertaining the average effect of a presumed cause). It does not seek to explain any of the causal mechanisms through which cause and effect are related, nor can it deal with a large number of contingency variables that limit the conditions under which a cause and an effect are related. Moreover, cause only refers to potentially manipulable events and excludes nonmanipulable ones that are central for causal explanation in some social sciences. Of course, RCM was never intended to be a model of causal explanation, nor was it intended to be so general as to apply to all the everyday life contexts where causation is invoked. Nonetheless, due to its clarity and strength, RCM is the predominant model of causal description. The following discussion of experiments and quasi-experimental designs strongly sticks to this causal model.

Causation and Experiments

Experiments are well suited for inferring causal relationships because (a) the presumed cause (treatment) is manipulated, making it easy to know that the cause precedes the effect in time; (b) covariation between the treatment and outcome can be readily observed; and (c) the treatment and control groups are treated identically in every way other than for treatment assignment, thus ruling out all alternative interpretations when certain assumptions are met. These are quite transparent assumptions because one can readily observe whether the groups were initially similar, whether there has been differential attrition from the study, and whether there is contamination between the treatment and control groups. No nonexperimental method matches the experiment on all of these characteristics that promote stable causal inference.

Some assumptions are crucial even when an experiment is done. The key ones are the following. First, randomization must be successfully implemented. For instance, if members of the administrative staff responsible for the assignment of students, classes, or schools overrule the random process, then a selection bias may emerge that corrupts causal effect estimates by confounding them with a potential selection effect. Second, because randomization equates treatment and control group "on average," detectable group differences may occur within probability limits, particularly when the number of sampled units is low and so unhappy randomization may result even from a proper randomization procedure. Third, random assignment controls for selection, but selection is only one of the many threats to internal validity on Cook and Campbell's (1979) list. So we have to add the further assumption that the treatment and control conditions are treated similarly in all ways other than treatment assignment, particularly in the ways that observation and measurement take place. Fourth, it is assumed that there is no differential attrition between the treatment and control groups. Attrition may occur when parents take their children out of the program or when students change schools. If the pattern of attrition differs by group, then a selection confound is introduced. Finally, neither the random assignment procedure nor the treatment or nontreatment other students (classes, schools) receive should affect a student's outcome. This assumption, often called the stable-unit-treatment-value-assumption (SUTVA), is violated if, for instance, some control students seek to compensate for not receiving the planned

treatment or if some intervention students do not faithfully comply with the program details. To achieve SUTVA, it is advisable to implement the experiment in a way such that students and teachers do not become aware of the specific treatment or control condition to which they were not assigned. This can often be achieved by selecting treatment and control classes from different schools and districts.

These assumptions are small in number, testable, and intuitively clear to theorists and practitioners of educational evaluation alike. Moreover, the long tradition of experimentation has led to developing strategies that protect against the violation of assumptions by learning both how to prevent them from occurring and how to deal with them if they should occur and are not extreme (see Cook & Campbell, 1979; Shadish et al., 2002).

Quasi-Experimental Designs

For ethical, practical, legal, or political reasons, randomized experiments are sometimes hard or impossible to implement. Many of the nonexperimental methods are exclusively correlational, making it difficult to know which of the correlated variables is the cause and which is the effect, given the ambiguity of temporal precedence. Moreover, correlational relationships may be due to a confounding variable correlated with both the cause and the effect—the reason for the cliché that correlation cannot prove causation. Even the strongest quasi-experimental designs—regression discontinuity designs, interrupted time series analysis, and nonequivalent control group designs, including propensity score matching and selection modeling—are less well suited for causal inference than experiments (Cook, Shadish, & Wong, 2008; Glazerman et al., 2003). These alternative designs and their linked analyses typically require more numerous and less realistic causal assumptions than the experiment, and the statistical techniques on which they depend for estimating effects require even more and even less transparent assumptions. Also a problem is that statistical tests are less efficient with non- and quasi-experiments than with experiments. However, nonrandomized designs are frequently recommended as good alternatives to randomized experiments (e.g., by AERA; Schneider et al., 2007). Here, we briefly describe the basic settings of the strongest quasi-experimental designs and discuss their warrants (for a general overview, see Shadish et al., 2002; West, Biesanz, & Pitts, 2000).

Regression Discontinuity Designs

Regression discontinuity designs have a long tradition, but have only recently experienced a renaissance (Cook, 2008), including educational evaluation (e.g., Angrist & Lavy, 1999; Barnett, Lamy, & Jung, 2005; Cohen, 2006; Gormley, Gayer, Phillips, & Dawson, 2005; Jackson et al., 2007; Jacob & Lefgren, 2004; Lockwood, Gill, Setodji, & Martorell, 2007; Van der Klaauw, 2002; Wong, Cook, Barnett, & Jung, 2008). The basic regression discontinuity design requires that participants are deterministically assigned to a treatment and control condition on the basis of a quantitative assignment variable (e.g., a student's birthday or a pretest score). There is no necessity that the assignment variable has a clear meaning or is measured without error. The crucial need is that participants with a measure below a fixed cutoff value of the assignment variable are assigned to one condition (treatment or control), whereas those above the cutoff are assigned to the other condition. After treatment, the causal effect on an outcome variable is investigated by regressing the outcome on the quantitative assignment variable and a treatment dummy variable—treated units are coded with 1, and untreated units are coded with 0. The dummy variable models the expected

discontinuity in the regression line exactly at the cutoff point and represent the causal effect. Because the treatment effect is estimated via regression, the assignment variable must be quantitative (i.e., a continuous variable). Nominal variables, such as gender or race, cannot be used because no regression line can be estimated—the discontinuity due to the treatment would be confounded with the effect of the dichotomous assignment variable.

Regression discontinuity designs are warranted by the feature that the assignment process into treatment or control conditions is completely known (Goldberger, 1972a, 1972b). For this reason, unbiased treatment effects can be estimated at the cutoff value, but only if assumptions in addition to those for the randomized experiment are met. First, the functional form of the regression equation must be correctly specified; second, there should be no interaction between the treatment and assignment variables. Frequently, a linear relationship between the outcome and assignment variable is assumed. However, curvilinear relationships should be modeled by including higher order polynomial terms of the quantitative assignment variable if substantive theory and data suggest it. Misspecifications of the functional form result in biased estimates of the treatment effect. In practice, the robustness of effect estimates should be checked by using different regression models. The assumption that there is no interaction effect between treatment and assignment variable implies that the regression lines are parallel for the treatment and control group. This means that there is only a discontinuity in the regression line at the cutoff, but no change in the slope. Interpretation complicates when there is a significant change in the slope or, more generally, in the functional form. If there is no discontinuity at the cutoff, but a change in slope, the increased or decreased slope cannot be uniquely attributed to the treatment without further assumptions. This is because the change may also reflect a nonlinear relationship between the outcome and the assignment variable. If there is a discontinuity at the cutoff in addition to the change in slope, the offset can be interpreted as a causal effect, but at the cutoff point only. The estimation of (nonconstant) treatment effects at other values than the cutoff is only possible if the change in slope was uniquely caused by treatment—alternative explanations, including nonlinear functional forms, must be ruled out.

However, if these additional assumptions hold, regression discontinuity designs are also empirically warranted alternatives to randomized experiments (Cook & Wong, in press). Indeed, at the cutoff point, regression discontinuity designs are equivalent to randomized experiments. But in contrast to randomized experiments, they have considerably lower power (given the same sample sizes). Power mainly depends on the choice of the cutoff value—cutoffs at the extreme ends of the assignment variable should be avoided—and the strength of correlation between the outcome and assignment variable. Even for well-designed regression designs, 2.75 times more observations are necessary to achieve the same power as in a corresponding randomized experiment (Goldberger, 1972a). Moreover, the use of only a single cutoff value restricts the generalization of causal effect estimates because treatment effects can only be interpreted at or close to the cutoff value. However, more complex variants of the basic design can at least partially deal with these restrictions (Judd & Kenny, 1981; Shadish et al., 2002; Trochim, 1984).

The key requirement of the regression discontinuity design is that subjects are assigned to the treatment or control condition solely on the basis of the cutoff of a quantitative assignment variable. This requirement is as strict as random assignment in a randomized experiment. If assignment does not solely take place according to the cutoff, the strength of the regression discontinuity design is corrupted, and biased effect estimates may result. This is the case if administrators or participants override the assignment rule in order to achieve or avoid treatment. For instance, teachers may override the assignment rule for students close to the cutoff value because they think that some of

these students do or do not need treatment. If students know in advance about the assignment variable (e.g., a vocabulary pretest) and the cutoff, they may try to manipulate their own pretest score by intentionally producing poor results. Because in such cases assignment is not completely controlled, selection bias may contaminate estimated treatment effects. The same holds if there are treatment crossovers (subjects assigned to treatment do not receive treatment, and subjects assigned to the control condition receive treatment) and attrition from the study.

Interrupted Time Series Designs

Interrupted time series designs are similar to regression discontinuity designs. The quantitative assignment variable is exclusively given by a time variable. The implementation of an intervention at a certain point in time separates an observed time series of the outcome under investigation into two parts: the time series before and the time series after intervention. As with regression discontinuity designs, regression analysis is used to assess potential effects of the intervention. For an effective intervention, one would expect an interruption in the pattern of the observed time series immediately after the intervention point. In the simplest case, this can be either a change in the time series' level, slope, or both.

To achieve unbiased estimates of treatment effects with interrupted time series designs, two assumptions must be met. First, time is the sole variable determining implementation of treatment. If time is not the single factor determining the assignment to the control and treatment condition, treatment effects may be biased. Second, the functional form of the outcome over time must be correctly modeled. That includes the correct specification of the long-term trend, as well as the identification of potential periodic cycles in the time series. Moreover, serial dependencies are likely because observations close together in time are likely to be not independent of each other. Although misspecifications of long-term trends or periodic cycles result in biased treatment effect estimates, the failure to adequately model serial dependencies leads to biased standard errors and, as a consequence, incorrect statistical inference. The proper specification of a time series typically requires a long time series—100 observations, as a rule of thumb (Velicer & Harrop, 1983). Otherwise long-term and cyclical trends cannot be reasonably estimated. If only short time series are available, competing models may fit the data equally well, but effect estimates may substantially differ. Hence, for short time series, more substantive knowledge or stronger assumptions about the functional form are needed to justify the results from a specific interrupted time series design.

Another problem associated with the interpretation of an interrupted time series design is that an observed change in the level, slope, or both must be uniquely attributable to the intervention. That is, alternative explanations for the interruption in the time series pattern must be ruled out. Particularly, effects of events occurring at approximately the same time as the intervention under investigation may be confounded with the treatment effect. First, other unrelated, competing, or compensating interventions influencing the outcome of interest may have been launched at about the same time. Second, the population under investigation may have changed. This happens if, immediately after the announcement, implementation, or becoming aware of the intervention, subjects start to select themselves into or out of treatment or the measurement framework. In such cases, the interruption in the time series is probably only due to an unintended change in the composition of the population covered by the pre- and postintervention time series. Third, changes in the measurement framework of the outcome (i.e., the reporting, measuring, or recording of the outcome of interest) may have changed over time, particularly simultaneously with the implementation of treatment.

An additional challenge with interrupted time series analyses occurs when interventions do not produce immediate but rather delayed effects. The reason for this may be that interventions are either not immediately and completely implemented, slowly diffuse through the population, or both. Delayed effects are more difficult to interpret unless theoretical justifications exist for explaining the observed delay. The longer the time period between treatment and the first possible effects, the more alternative interpretations are plausible. In particular, with short time series, only immediate effects can be detected. The assessment of delayed effects requires much longer time series. Further, power issues are also important for interrupted time series analysis. If time series are short and show a high amount of unexplained error, weak effects are difficult to prove.

For both the regression discontinuity and the interrupted time series analysis, the major problem is that alternative interpretations for the discontinuity or interruption must be ruled out. If alternative interpretations other than the treatment remain likely, then the observed effect may not be causally attributed to the intervention. To rule out alternative explanations, the basic design can be improved by including nonequivalent control group time series without any treatment or other nonequivalent dependent variables that are not affected by treatment, but would reveal potential threats to the interrupted time series' internal validity. Sometimes it is also possible to show the effect of an intervention not only by introducing it, but also by removing it at a later point in time. Another strategy consists of adding a switching replication of the time series, where an additional group receives treatment at a later point in time (Shadish et al., 2002).

However, if time series are long enough and if the required assumptions hold, then interrupted time series designs are among the strongest quasi-experimental designs. Nonetheless, time series designs in educational evaluations are rare and typically not long (Henry & Rubenstein, 2002; Kearney & Kim, 1990; Lin & Lawrenz, 1999; May & Supovitz, 2006; Moon, Stanley, & Shin, 2005).

Nonequivalent Control Group Designs

The currently most popular quasi-experimental alternative is probably the nonequivalent control group design. As in a randomized experiment, a group of subjects who received treatment is compared to a control group not receiving treatment. But unlike in randomized experiments, treatment is not randomly assigned to participants. Rather, they select themselves or are selected by administrators or third persons (e.g., parents) into treatment. Thus, the selection process into treatment is usually not completely known and measured. Hence, a direct comparison of the treatment and comparison groups cannot yield an unbiased estimate of the treatment effect as long as groups differ prior to treatment with respect to important background characteristics that are related to the outcome under study. Only if treatment and comparison groups can be balanced on all important covariates that are related to both treatment and outcome effect, estimates can be adjusted for pretreatment group differences. In principal, two main strategies for aligning treatment and comparison groups are possible: individual case matching and intact group matching.

Individual Case Matching

With individual case matching treatment and comparison groups are typically balanced on the basis of individual-level covariates, but also group-level covariates may be included for each case. A variety of statistical methods has been suggested to adjust treatment effects for pretreatment group differences in observed covariates (Morgan & Winship, 2007; Rosenbaum, 2002; Rubin, 2006). Among them are covariance adjustment via regression analysis (ANCOVA), econometric selection modeling, as well as stratification, weighting, and matching approaches on the basis of either the originally observed covariates or the propensity score. Propensity scores are currently frequently

used in educational evaluations (e.g., Hill, Rubin, & Thomas, 1999; Hong & Raudenbush, 2005, 2006; Morgan, 2001). Propensity scores try to model the unknown selection process and are defined as the conditional probability that subjects received treatment, given all observed background variables (Rosenbaum & Rubin, 1983, 1984). In practical applications, propensity scores are estimated using logistic regression or discriminant analysis with observed covariates as independent variables. If the propensity score model is correctly specified, estimated propensity scores are able to balance pretreatment group differences on observed covariates. Balance in groups can then be achieved by (a) including the propensity score as a predictor into the regression model for the outcome, (b) stratifying observations on the basis of the propensity score, (c) weighting observations with weights derived from the propensity score, or (d) matching individual cases of the treatment and comparison group solely on the basis of propensity scores or together with other covariates.

However, all these methods require a strong assumption to obtain unbiased treatment effects: the assumption of a strongly ignorable treatment assignment, also called selection on observables or unconfoundedness assumption (Rosenbaum & Rubin, 1983). This assumption requires that (a) all important covariates related to treatment and outcome are identified and reliably measured, and (b) sufficient overlap of treatment and comparison group on these covariates is given. The first part of the strong ignorability assumption ensures that all pretreatment group differences that also affect the outcome can be balanced. This is possible only when all these confounding covariates are measured. The second part of the assumption requires that the joint covariate distributions of the treatment and control group completely overlap. This means that for each treated subject with specific background characteristics, a corresponding untreated subject with the same or similar background characteristics should have been observed. If no subjects in the control group share similar characteristics, a lack of overlap is given and treatment effects for this part of the covariate distribution cannot be estimated. The assumption of sufficient overlap can be checked by plotting the treatment and control group's univariate distributions of observed covariates and propensity scores. Unfortunately, an empirical test of the first part of the strong ignorability assumption is not possible. Only substantive theory on determining factors of the actual selection process and their relation to the outcome under investigation may help to justify the assumption. If there are reasonable doubts about whether all covariates related to both treatment selection and outcome have been measured, strong ignorability may not hold, and estimates of the treatment effect may remain considerably biased. For instance, if the most important covariates explaining treatment selection—that is, pretest measures on the same scale as the outcome and motivational factors for choosing or avoiding the treatment under consideration—are not measured, strong ignorability can hardly be assumed (Steiner et al., under review). It is important to note that groups must also be balanced with regard to different maturation rates, which is a major issue in educational evaluations. This can be achieved by considering changes in pretest measures as additional covariates. Even if all covariates required for establishing a strongly ignorability treatment assignment are observed, regression models for the outcome or the propensity score must be correctly specified in order to obtain unbiased effect estimates.

Intact Group Matching

In retrospective studies, which rely on existing databases, not all important covariates related to treatment assignment may be available. Consequently, a strongly ignorable treatment assignment cannot be reasonably assumed, and effect estimates based on individual case matching may be plagued by hidden bias. This is particularly true when national datasets are used to construct

matched pairs for a locally implemented intervention. National datasets are not designed to represent the complex selection models operating in local settings with a specific intervention. Within-study comparisons have shown that such retrospective nonequivalent group designs using propensity scores nearly always fail to approximate the results of their experimental benchmarks when treatment and comparison populations are initially very different (Cook et al., 2008; Glazerman et al., 2003). This emphasizes the importance of locally and focally similar comparison groups—groups that come from data samples in the same locale with the same substantive characteristics as the treatment group. Then, even without individual selection modeling, the comparison of well-matched, intact groups can result in pretest means and slopes that are similar for experimentally and nonexperimentally constructed comparisons (Aiken, West, Schwalm, Caroll, & Hsuing, 1998; Bloom, Michaelopoulos, & Hill, 2005). Even if complete bias reduction cannot be achieved with an intact group matching, the greater initial overlap relative to other possible nonequivalent populations is likely to improve bias reduction with adjustment methods such as individual case propensity score matching.

In cases where classrooms or schools are the unit of analysis, one can also take advantage of the fact that school achievement data from prior years are often available. Then multiple comparison schools can be selected by matching on school-level pretest means and slopes over several years.

Experimentation and Evaluation in the United States and Europe

Quasi-experimental designs, although not the strongest ones as described earlier, are dominant in educational evaluation, whereas experimental designs are often an exemption, particularly in Europe. Two main reasons for the marginal role of experimentation in evaluating educational programs can be identified: the historical emergence and tradition of experimentation (described in this section), and reservations about using experiments in educational evaluation (described in the next section).

In the United States, experiments are common for assessing the effectiveness of a program or intervention, although they are relatively rare in education (Cook & Gorard, 2007). In the European countries, experiments are rare except in medicine, psychology and agriculture. In general, this has to do with the different roles that evaluation plays in the United States and Europe. In the United States, evaluation has a longer tradition and is more institutionalized. According to Rossi, Freeman, & Lipsey (1999), evaluation of social programs has its roots in the United States of the 1930s. A real boom started in the 1960s, when various social welfare programs were launched and their effects had to be assessed. The demand for evaluation exceeded the capacity of the U.S. General Accounting Office, and so evaluation opened up new employment possibilities for the dramatically increasing number of social science doctorates. Two professional evaluation societies also began in the 1970s (Evaluation Research Society and Evaluation Network), and professional journals and standards were not long behind.

In the 1980s, some countries within the Anglo-Saxon tradition started introducing public sector reforms (known as New Public Management). Here, the UK, Australia, and New Zealand were foremost, and some Northern European continental countries (e.g., Sweden) followed. According to Stame (2003), these countries also participated in the Anglo-Saxon debate concerning evaluation methods and techniques. The situation in most European countries has lagged behind this development. The process of professionalism and institutionalization started during the 1990s, thanks to an external push coming from the EU. The EU has developed “a complex system of multi-level governance of which a specific architecture of evaluation is a crucial element” (Stame,

2003, p. 39). This system of multilevel governance is characterized by the following process: The EU establishes general goals and allocates money to the states, the member states establish specific goals and allocate money to regions, and the lower levels decide on programs and interventions. As a consequence, an evaluation hierarchy corresponding to these levels has been institutionalized. Important steps toward professionalism of evaluation were the foundation of the European Evaluation Society (founded in 1994) or the German Association for Evaluation (founded in 1997). The development of methods is characterized by a combination of orientation toward the established standards in the Anglo-Saxon countries and the development of original approaches building on their own cultural traditions.

Although in the United States evaluation emerged from substantive evaluation theory and the general development of social science methods, in Europe evaluation was strongly associated with standards and procedures of accounting. Auditors are much more central in evaluation than social scientists trained in evaluation theory. Not surprisingly, auditors and social scientists have quite different views about evaluation (Cook & Wittmann, 1998). Social scientists translate government programs into theoretical statement about the relationship between inputs and outputs. They are interested in the causal conditions and program factors that lead to a more or less successful implementation of the program. They focus on different levels (e.g., students, classes, schools, regions) and include unintended side effects into their investigations. They use different methods for revealing the truth about the program, including experimental and quasi-experimental designs, econometric models, and qualitative techniques. They are concerned about ethics and values, fearing the limitation to a single perspective, particularly that of a powerful government. In contrast, auditors are more engaged with auditing standards, the monitoring of program implementation according to these standards, and the cost effectiveness of public funds spent for governmental programs. They are less interested in the causal relation between inputs and outputs, and they do not care much about how an effect came about. Likewise, unintended side effects and policy and ethical considerations are of minor relevance to them. Auditing standards mainly focus on the judgment of how well a program is implemented relative to its goals. In the United States, evaluation is strongly associated with empirically based decision making. In Europe, it is seen as a supportive management tool for an efficient allocation of resources, for further development of programs, and for helping politicians when they argue for or against a program. Hence, European evaluators are generally more sensitive about the political consequences of their work.

The priority of evaluation questions also differs between the United States and European countries. In the United States, evaluation mainly refers to generating and using information on the actual performance of implemented programs. In Europe, evaluation also comprises ex-ante investigations for planning intended programs and assessing effects to be expected in the future. Hence, this kind of ex-ante evaluation strongly depends on the validity of the substantive model and the underlying assumptions, whereas ex-post evaluation—as typical for the United States—relies on demonstrated performance—that is, what happened and not what might happen.

Due to the different perspectives of evaluation, experiments are of higher significance in the United States than in Europe (Cook & Wittmann, 1998). During the 1960s, quantitative methods dominated qualitative techniques in the United States. To investigate the causal effects of programs, experimental methods were clearly preferred to nonexperimental ones, design controls were preferred to statistical controls, and qualitative techniques were downplayed because they are not able to rule out competing causal interpretations. However, in the 1970s and 1980s, qualitative methods became more and more important in education as experience led commentators to believe that large causal effects are rare. In addition, an increasing number of social scientists and scholars

thought that quantitative methods—especially experiments—are epistemologically too restricted. Because the development of evaluation in Europe lagged behind the one of the United States, the booming phase of experimentation was basically missed. Instead, European evaluators were more strongly committed to both auditing and qualitative versus quantitative methods. In the latest evaluation guidelines of the European Commission (Directorate-General (DG) Budget, 2004), experimental designs are not even mentioned, although the DG for Employment concluded that quasi-experimental designs constitute the most important way to assess effects of intervention "since perfect experimental comparisons do not exist" (European Commission, 1999, p. 14). For educational evaluation, the EU has not yet formulated explicit evaluation method preferences. However, the number of experiments is low in educational evaluation in both the United States and Europe, and educational evaluators in each setting generally use the same arguments for rejecting experiments in favor of quasi- or nonexperimental investigations. Their reservations to randomized experiments are analyzed in some detail in the remainder of this chapter.

The Validity of Reservations About Using Experiments in Educational Evaluation

The superiority of random assignment for drawing inferences about the consequences of planned interventions is routinely acknowledged in philosophy, medicine, public health, agriculture, statistics, microeconomics, psychology, criminology, prevention research, early childhood education, and marketing. Furthermore, it is also acknowledged in those parts of political science and sociology concerned with improving opinion surveys, as well as in all the elementary education method textbooks we have consulted. However, random assignment is relatively rare in educational evaluation, especially for assessing the impact of educational interventions of obvious policy relevance. Random assignment is also rare in sociology, political science, macroeconomics, and management. Yet causal statements are routinely made in these fields, usually through a process that links substantive theory to various qualitative or quantitative nonexperimental practices.

We do not argue that correct causal conclusions come only from experiments. We argue that experiments provide a better warrant for such conclusions than any quasi-experimental method (see also the series of discussions in the Point/Counterpoint Section of the Journal of Policy Analysis and Management 27(2), 27(3), and 28(1), with the opening statements in Nathan, 2008). So, if experiments can be conducted in schools, they should be. Not to use them requires a strong justification.

Over the last 30 years, self-ascribed educational evaluators such as Alkin, Cronbach, Eisner, Fetterman, Fullan, Guba, House, Hubermann, Lincoln, Miles, Provus, Sanders, Schwandt, Stake, Stufflebeam, and Worthen have proposed many justifications for not doing experiments (Cook, 2002). These theorists want educational evaluation to pursue goals other than describing what works in schools. Most of them want evaluation to improve the organization and management of individual districts or schools, assuming that this will routinely improve student performance. They examine ways to provide individual schools or district staff with continuous feedback about strategic planning, program implementation, and student or teacher performance monitoring. The expectation is that local officials will immediately use this feedback in their schools and that student performance will consequently improve. This model of research and its connection to organizational change is much like what we find in management consulting in the private sector. Other educational evaluators want evaluation to contribute to developing general theories, especially those that specify the often complex constellation of forces that bring about important

school effects. Engaged time on task is such a generative process and, over a broad set of circumstances, enhances academic achievement. It can be instantiated in many different ways—as more days of schooling per year, as longer school days, as more time devoted to the core curriculum, as textbooks that are engaging, as exposure to teachers who know how to motivate students, and so on. Identifying such generative causal mechanisms becomes the paramount goal of evaluation. Unfortunately, neither the management consulting nor the causal mechanism model of evaluation places the premium where experimentation does—on directly observing student change and unambiguously attributing it to a single policy-related treatment. Although the management consulting and causal mechanism approaches may deliver valuable hints and theories for causal inference, they are not able to disentangle the complex and confounded effects on empirically measured student achievements.

The objections to randomized experimentation are manifold. According to Cook (2002), who discusses several reservations to experiments and illustrates them with a lot of examples from the United States, the common arguments put forward can be divided into five main categories: (a) practical arguments, (b) arguments about undesirable trade-offs, (c) arguments that experiments are not necessary because better alternatives exist, (d) arguments that schools will not use experimental results, and (e) philosophical arguments. In the following, we only discuss the first three arguments because they are more strongly related to practical issues and alternative approaches to randomized experiments.

Practical Reasons for Not Doing Experiments

Randomized Experiments Cannot Be Mounted

Opponents of randomized experiments assert a number of reasons that experimentation cannot be implemented, particularly in school research: Many officials do not like the unequal allocation of resources generated by random assignment and fear respective negative reactions from parents and staff. Due to their complexity, educational related topics are not appropriate for experimental investigations. Therefore, other—often less effective and less esteemed—methods are generally preferred to randomized experiments (Cook, 2002). Nevertheless, it is striking that—at least in the United States—experiments in schools are to be found when the topic is not pedagogic, such as school-based programs to prevent negative behavior (tobacco, drugs, alcohol). They are also common in preschool education. One possible explanation might be the different time requirements associated with the intervention and when it is expected to achieve results. Pedagogical interventions are more likely to be multiyear; they require a change in established routines, and if they are not successful it might threaten a school's local reputation.

Furthermore, the discipline-based difference in the frequency of experiments may also be due to disciplinary culture. Random assignment is common in health sciences, where it is institutionally supported by funding agencies, integrated in graduate training programs, has a long tradition (e.g., clinical trials), and is considered in political discussions.

Cook (2000) argues that the implementation of experimental settings is easier in cases with centralized decision making (e.g., when funding of a program is bound to the use of experiments in evaluation). Furthermore, it is necessary to give an incentive to schools that participate in the control group of an experiment. A motivation for schools to participate without belonging to the treatment group could be the promise that they would be the first to offer the intervention at the study end, by when it might be improved. Last but not least, all these practical challenges indicate

that random assignment should be in independent hands and carried out by staff with experience in randomization in complex settings.

Even When Experiments Are Mounted, Many of the Planned Between-Treatment Contrasts Become Compromised

Random assignment leads to treatment and control groups that are equivalent at the pretest prior to treatment. Assuming unchanged circumstances, the treatment effect is defined as the difference in the outcome variable in a posttest. Experiments are likely to be compromised if systematic effects—making groups different—operate. Such effects are differential attrition and treatment crossovers.

Differential attrition occurs if different kinds of students drop out of various treatment groups, resulting in nonequivalent groups and consequently in effect estimates of questionable value. Attrition may be kept at a low rate if school staff can strongly be committed to participation and the acceptance of the random assignment results and if modest payments to the units experiencing less desirable treatments are provided. It is also important that treatment implementation is closely monitored, particularly in order to detect and deal with early dropout trends. However, with long-lasting treatments, some attrition—due to changes in the school management, for instance—cannot be prevented. Nevertheless, units lost to intervention should remain within the measurement framework. Although attrition may never be completely avoided, the resulting bias is likely to be less than the bias due to a complete self-selection of schools or teachers from start. Statistical selection controls are better the smaller the initial bias and the better selection have been measured (Holland, 1986).

Furthermore, experiments might be compromised by treatment crossover. Although extensive crossovers may be rare, Cook et al. (1999) showed that 3 out of 10 control schools borrowed program elements. This was mainly due to informal communication paths, which facilitated an exchange between units in the treatment and control groups concerning the intervention. Minimizing crossovers requires a well-planned experimental design, including random selection of physically separated units, innovative treatments, and the measurement of treatment fidelity.

Random Assignment Assumes Fixed Program Theory and Standard Implementation, But These Treatment-Specific Assumptions Are Not Valid for School Contexts

The interpretation of experimental results is facilitated when intervention is based on strong substantive theory, when implementation corresponds with the treatment-specific program theory, and when variation within each treatment implementation is minimal. However, in school research, these conditions are rarely met because schools are complex social organizations faced with conflicting stakeholder goals. Standardized implementation of a reform initiative or total fidelity to program theory is usually not achieved. So, the assumption of treatment homogeneity or invariance of settings in the educational contexts can hardly be justified in experimental investigations. However, random assignment does not require well-specified program theories, good management, standard implementation, or treatments that exactly correspond to program theory. Experiments primarily protect against bias in causal estimates and only secondarily against imprecision in these estimates resulting from the complexity and heterogeneity of schools. But increasing school sample sizes and measuring school-specific sources of variation to reduce their unwanted influence through statistical control can tackle this. In addition, implementation quality should be studied on its own to learn about which types of schools and teachers implement the program better. It is important to note that only a few educational interventions will be standardized once they are implemented as

formal policy. So, why standardize in an experiment? The measurement of sources of implementation variation and their inclusion in the analysis of causal effects is of greater importance than standardization.

Random Assignment Entails Undesirable Trade-offs

Increasing Internal Validity Decreases External Validity

The strength of experiments is internal validity, the focus on unbiased causal estimates, rather than external validity or causal explanation. Therefore, experiments are clearly limited in time and space. But scientists typically prefer general results—results that are at least more general than those derived from a single experiment of a specifically implemented intervention in a particular sample of schools that volunteered for experimentation. Moreover, educational evaluators seek for general causal agents whose operating mechanisms are fully understood. They place less priority on the effectiveness of a particular implementation of a program in a particular time with a particular group of respondents. This means that they are prepared to tolerate more uncertainty than other scientists who would like to know whether a program works reliably.

One possibility to overcome the limited generalizability of single experiments is to implement experiments in a way that sampling particulars permit tests of generalization across types of students, teachers, settings, and times. With random sampling of these instances, followed by random assignment to treatment, empirical robustness of effects or boundary conditions under which effects occur can hopefully be demonstrated.

However, random sampling is hardly relevant if, for instance, volunteering to be in a study is required or causal relationships may vary by historical period. Further, random sampling cannot be used to select the outcome measures and treatment variants that are used to represent general cause-and-effect constructs. So, a different generalization model is required—one that emphasizes how consistently a causal relationship replicates across multiple sources of heterogeneity (Cook, 1993): Can the same causal relationship be observed across different laboratories, time periods, regions of the country, and ways of operationalizing the cause and effect? This heterogeneity-of-replication model permits purposive instead of random sampling. Only the heterogeneous sampling plan with respect to people, settings, operational definitions, and times are of vital importance. Single experiments rarely produce definitive answers, and, in addition, they are not able to answer all ancillary questions about the contingencies on which a causal relationship depends. In this sense, causal generalization can be understood as an average effect size derived from heterogeneous studies of the same hypothesis. But it can also be seen as an identifying generative causal process. For instance, engaged time on task is presumed to stimulate achievement through activities such as more homework, summer classes, or longer school days. The methods for identifying such explanatory processes place relatively little weight on sampling; instead, they require the measurement of each variable in the presumed generative theory. Fortunately, it is easier to build these explanatory methods into individual experiments than it is to sample at random or to add populations to the sampling design. Hence, experiments could and should be designed to explain the consequences of interventions and not just to describe them. This means adding to an experiment's measurement and sampling plan and abjuring black box experiments.

Prioritizing Scientific Purity Over Utility

Critics frequently argue that experimenters focus only on uncertainty reduction about the cause in order to obtain pure effect estimates, rather than results of more general utility. Some questions

may illustrate this point of view. Why not use a more liberal level of significance, say $\alpha = .25$ instead of a conservative $\alpha = .05$? Why include schools defying treatment implementation in the treatment group for the intention-to-treat analysis? Why not investigate unplanned treatment interactions or treatment implementations? Why persist with the original research question if a more useful question has emerged during the study, even if unbiased answers to the new question are not possible? It seems that experiments are only designed for bias reduction and that other types of knowledge are secondary at best. But experiments need not be so rigid. There is no need for stringent alpha rates. One need not be restricted to the intention-to-treat analysis only, although these results should be reported. Also, interaction effects may be investigated, with substantive theory and statistical power in mind. Hence, pure effect estimates can be obtained from experimental data as well as other relevant results, especially if additional ethnographic data are used. Such data are of relevance for understanding issues on implementation, causal mediation, and unintended outcomes and improve each controlled experiment.

Random Assignment Is Not Needed Because Better Alternatives Already Exist

Intensive Case Studies Are More Flexible

Intensive qualitative case studies are often seen as superior alternatives to experiments, mainly due to their greater flexibility. Although an experiment only focuses on a narrow causal aspect, evaluators are convinced that case studies are appropriate for evaluating program theory, assessing implementation, recording program redesign, identifying intended and unplanned effects, detecting contingencies, or assessing the findings relevance for different stakeholder groups. They assert that intensive qualitative case studies are able to reduce the uncertainty about a cause to an acceptable level and sometimes—undoubtedly—even all the uncertainty about a cause. However, it will usually be difficult to know when this happened. Nonetheless, case studies do not reduce as much causal uncertainty as well-executed experiments. The absence of control groups—the causal counterfactual—makes it difficult to know how a treatment group would have changed in the absence of the intervention. If a high standard of uncertainty reduction is prioritized, randomized experiments are indispensable.

However, intensive qualitative case studies complement experiments whenever a causal question is central, but it is not clear how successful program implementation will be, why implementation shortfalls may occur, what unexpected effects are likely to emerge, what the mediating processes are, and so on. Case studies can have a central role within experiments, but are not better alternatives to experiments in causal questions.

Quasi-Experiments Are as Good as Experiments

Quasi-experiments are identical to experiments in purpose and in most structural details, the defining difference being nonrandom assignment. Quasi-experiments use design rather than statistical controls to create the best possible approximation to the missing counterfactual that random assignment would have generated. These design controls include matched comparison groups, age or sibling controls, pretest measures at several times before a treatment begins, interrupted time series assigning units based solely on a quantitative criterion, assigning the same treatment to different groups at different times, and building multiple outcome variables into studies, some of which should theoretically be influenced by a treatment and others not (Corrin & Cook, 1998; Shadish et al., 2002). Quasi-experimental designs are created through a mixing

process that tailors the research problem and the resources available to the best design that can be achieved by mixing the previous design elements.

However, strong quasi-experiments with design elements mentioned earlier are rarely found in educational evaluation. In particular, the strongest quasi-experiments—interrupted time series analysis, regression discontinuity analysis, and nonequivalent control group designs with more than one pretest measurement—started to enter educational evaluation only recently (see the section on quasi-experimental designs). Instead, weak quasi-experiments with some form of nonequivalent control groups or some pretreatment observations can be frequently found, but they run the risk of being "generally causally uninterpretable" (Campbell & Stanley, 1963; Cook & Campbell, 1979). Quasi-experiments are more likely to be biased and inefficient when compared with experimental results. In areas such as education, where few studies exist, randomized experiments are particularly needed. It will take fewer of them to arrive at what might—or might not—be the same answer, and anyway, most scholars trust the answers from experiments more than from quasi-experiments.

Theories of Change

Theories of change are used in evaluations of interventions in complex social settings such as schools or communities (Connell, Kubisch, Schorr, & Weiss, 1995). The theory of change requires a detailed explication of the substantive theory behind a reform initiative and the specification of all flow-through relationships that should occur if the intended intervention is to impact on a major distal outcome such as student achievement. To this end, highly valid measurements of each construct in the substantive theory as well as a valid analysis of multivariate explanatory processes are necessary for assessing whether the postulated relationships have actually occurred in the predicted time sequences. Without using a causal counterfactual (i.e., control or comparison groups), it is assumed that the theory under investigation is proved if the data patterns obtained are congruent with the program theory.

Without doubt, the extensive use of substantive theory to guide measurement and analysis is of great value for improving causal probes. But the issue is whether such measurement and analysis alone can completely substitute randomized experiments. There are reasons for skepticism about the validity of using theories of change to support strong causal conclusions (Cook, 2000). They comprise the difficulty with making program theory explicit and unique (competing theories may exist). They also cover problems in specifying the timelines of effects, the linearity in the flow of influence often neglecting reciprocal feedback loops or external contingencies moderating effects, or the difficulty in obtaining valid measurement. In addition, there is usually not only one unique but a set of rather heterogeneous theories of change that all fit to a single pattern of data (Glymour, Scheines, Spirtes, & Kelly 1987). The implication here is that causal modeling is more valid when multiple competing models are tested against each other, rather than when a single model is tested. As with case studies, the biggest problem with theory-of-change models is the absence of a valid counterfactual: Which models what would have happened without treatment? As a result, it is impossible to decide whether the observed data result from the intervention or would have occurred anyway. Although theories of change are not an adequate alternative to experiments if causal effects are to be analyzed, they give valuable information about why these effects occurred, as well as the mechanisms behind that. From that point of view, theories of change are—like case studies—no alternative to randomized experiments; but they are a valuable completion.

Conclusion

Within the RCM, random assignment remains the most reliable technique for justifying causal inference. It provides the logically most valid and efficient causal counterfactual. Consequently, results are more credible than those from other quasi- or nonexperimental methods. Moreover, empirical comparisons of experiments and their alternatives suggest (Bloom, Michaelopoulos, Hill, & Lei, 2002; Glazerman et al., 2003; Lipsey & Wilson, 1993) that individual experiments are less biased, and that, as studies on a topic accumulate, they are more efficient about reducing causal uncertainty than quasi-experiments. Therefore, from a pragmatic point of view, experiments have a lower risk of drawing false causal conclusions, and they are probably less expensive in the long run because fewer of them are needed for the same degree of confidence in the causal conclusion drawn.

Although the superiority of randomized experiments is generally known, experimentation is still too rare in research on the effectiveness of school-based strategies to improve student performance. However, random assignment is not at all rare in preschool education or in school research on preventing negative behaviors or feelings. One possible reason is the difference in intellectual culture. Prevention researchers and preschool teachers tend to be trained in fields, where random assignment is more esteemed and where funders and journal editors clearly prefer this technique. In contrast, training and professional rewards in educational research set no high value on experimentation. Another reason may be the difference in experiments' scale. Most school-based prevention experiments are typically shorter, implemented by researchers, rather than school staff, and research topics probably involve educators less than issues of school governance or teaching practice. It is true, experimentation is more demanding in school-based research, but it can and should be done, particularly in cooperation with evaluators trained and experienced in randomized experiments.

However, a more successful dissemination of random assignment in school-based research is restrained by the belief of most educational evaluators that experiments are of little value. They believe that the theory of causation underlying experimentation is naïve, that experiments cannot be successfully implemented, and that they require unacceptable trade-offs. They also argue that experiments deliver a kind of information that is rarely used to change policy, and that the information experiments provide can be gained using simpler and more flexible methods. Some of these beliefs are better justified than others. Beliefs about the viability of alternatives to experiments are particularly less strongly warranted because no current quasi-experimental method or other alternative provides as convincing a causal counterfactual as the randomized assignment. Educational evaluators will not be persuaded to do experiments simply by outlining their advantages and describing newer methods for implementing randomization. Most educational evaluators share some of the reservations outlined earlier. To start a dialog, advocates of experimentation will need to be more explicit about the method's limit. They will also have to take some of the critics' concerns seriously—especially about program theory, the quality of implementation, the value of qualitative data, the necessity for analysis of causal contingency, and concern to meet the information needs of school personnel as well as other stakeholders. Finally, they will have to incorporate them into experimental practice.

Strongly warranted quasi-experimental methods—regression discontinuity designs, interrupted time series designs, and nonequivalent control group designs with close matching or sophisticated pattern matching—should be used whenever randomized experiments cannot be conducted. In any case, quasi-experimental investigations can complement findings from randomized experiments

and may help in generalizing them. For successful educational planning and policymaking, we need strong causal evidence from both randomized experiments and quasi-experiments.

References

Aiken, L. S., West, S. G., Schwalm, D. E., Caroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*(4), 207–244.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effects of class size on academic achievement. *Quarterly Journal of Economics*, *114*(2), 533–576.

Barnett, W. S., Lamy, C., & Jung, K. (2005). *The effects of state prekindergarten programs on young children's school readiness in five states*. New Brunswick, NJ: National Institute for Early Education Research.

Bloom, H. S., Michaelopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), *Learning more from social experiments* (pp. 173–235). New York: Russell Sage Foundation.

Bloom, H. S., Michaelopoulos, C., Hill, C. J., & Lei, Y. (2002). *Can non-experimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs*? New York: Manpower Demonstration Research Corporation.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Cohen, J. L. (2006). *Causes and consequences of special education placement*: *Evidence from Chicago public schools*. Cambridge, MA: MIT Press.

Connell, J. P., Kubisch, A. C., Schorr, L. B., & Weiss, C. H. (Eds.). (1995). *New approaches to evaluating community initiatives: Concepts, methods and contexts.* Washington, DC: Aspen Institute.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest & A. G. Scott (Eds.), *New Directions for Program Evaluation: Understanding Causes and Generalizing about Them, 57*, 39–82. San Francisco: Jossey-Bass Publishers.

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. In L. Sechrest & A. G. Scott (Eds.), *New Directions in Evaluation*: *Challenges and Opportunities in Program Theory Evaluation*, *87*, 27–34. San Francisco: Jossey-Bass.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*: *Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cook, T. D., & Gorard, S. (2007). Where does good evidence come from? *International Journal of Research and Method in Education*, *30*(3), 307–323.

Cook, T. D., Habib, F., Phillips, J., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal*, *36*(3), 543–597.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750.

Cook, T. D., & Wittmann, W. W. (1998). Lessons learned about evaluation in the United States and some possible implications for Europe. *European Journal of Psychological Assessment*, *14*(2), 97–115.

Cook, T. D., & Wong, V. C. (in press). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*.

Corrin, W. J., & Cook, T. D. (1998). Design elements of quasi-experimentation. *Advances in Educational Productivity*, *7*, 35–57.

European Commission. (1999). *Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006*. DG Employment, Industrial Relations and Social Affairs. Retrieved July 15, 2007, from www.igfse.pt/upload/docs/aval_LP_orientacoes_processo_aval_inter_DGEmprego.pdf.

European Commission. (2004). *Evaluating EU Activities. A practical Guide for the Commission Services*. DG-Budget, Evaluation Unit, Office for Official Publications of the European Communities, Luxemburg.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15*, 451–474.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy*, *589*, 63–93.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure*: *Artificial intelligence, philosophy of science and statistical modeling*. Orlando: Academic Press.

Goldberger, A. S. (1972a). *Selection bias in evaluating treatment effects*: *Some formal illustrations*. Unpublished manuscript, Madison, WI.

Goldberger, A. S. (1972b). *Selection bias in evaluating treatment effects*: *The case of interaction*. Unpublished manuscript, Madison, WI.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*(6), 872–884.

Henry, G. T., & Rubenstein, R. (2002). Paying for grades: Impact of merit-based financial aid on educational quality. *Journal of Policy Analysis and Management*, *21*(1), 93–109.

Hill, J., Rubin, D. B., & Thomas, N. (1999). The design of the New York school choice scholarship program evaluation. In L. Bickman (Ed.), *Research designs: Donald Campbell's legacy* (pp. 155–180). London: Sage.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

Hong, G., & Raudenbush. S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205–224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*, 901–910.

Jacob, B., & Lefgren, L. (2004). Remedial education and student achievement: A regression discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226–244.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., Ross, C., Schochet, P., Swank, P., & Schmidt, S. R. (2007). National Evaluation of Early Reading First. Final Report to Congress. U.S. Department of Education, Institute of Education Sciences: Washington DC.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.

Kearney, C. P., & Kim, T. (1990). Fiscal impacts and redistributive effects of the new federalism on Michigan school districts. *Educational Evaluation and Policy Analysis*, *12*(4), 375–387.

Lin, H. S., & Lawrenz, F. (1999). Using time-series design in the assessment of teaching effectiveness. *Science Education*, *83*(9), 409–422.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181–1209.

Lockwood, J. R., Gill, B. P., Setodji, M. C., & Martorell, F. (2007). *Regression discontinuity analyses of the effects of NCLB accountability provisions on student achievement*. Paper presented at the Joint Statistical Meeting in July, Salt Lake City.

Mackie, J. L. (1974). *The cement of the universe*. Oxford, England: Oxford University Press.

May, H., & Supovitz, J. A. (2006). Capturing the cumulative effects of school reform: An 11-year study of the impacts of America's choice on student achievement. *Educational Evaluation and Policy Analysis*, *28*(3), 231–257.

Moon, S., Stanley, R. E., & Shin, J. (2005). Measuring the impact of lotteries on state per pupil expenditures for education: Assessing the national evidence. *Review of Policy Research*, *22*(2), 205–220.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, *74*(4), 341–374.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference*: *Methods and principles for social research*. Cambridge: Cambridge University Press.

Nathan, R. P. (2008). The role of random assignment in social policy research. *Journal of Policy Analysis and Management, 27*(2), 401–415.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Translated in *Statistical Science*, *5*(4), 465–480.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation. A systematic Approach* (6th ed.). Sage.

Rubin, D. B. (1974). Estimation of causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects. *The Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge: Cambridge University Press.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*, 1334–1343.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stame, N. (2003). Evaluation and the policy context: The European experience. *Evaluation Journal of Australasia*, *3*(2), 36–43.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (under review). The importance of covariate selection in controlling for selection bias in observational studies.

Trochim, W. M. K. (1984). *Research design for program evaluation*. Beverly Hills, CA: Sage Publications.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment. A regression-discontinuity approach. *International Economic Review*, *43*(4), 1249–1287.

Velicer, W. F., & Harrop, J. W. (1983). The reliability and accuracy of time series model identification. *Evaluation Review*, *7*(4), 551–560.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings. Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). Cambridge, UK: Cambridge University Press.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 1–33.