

2. Causality: Forming an Evidential Base

RESEARCH DESIGNS ARE DEFINED by the types of questions asked. In the case of randomized controlled experiments, the question is: What is the effect of a specific program or intervention? An intervention, such as a curricular innovation, can be viewed as the cause of an effect, such as improved student learning. “A *cause* is that which makes any other thing, either simple idea, substance, or mode, begin to be; and an *effect* is that which had its beginning from some other thing” (Locke, 1690/1975, p. 325). As Shadish, Cook, and Campbell (2002) observe, however, we rarely know all of the causes of observed effects or how they relate to one another. Holland (1986) points out that a true cause cannot be determined unequivocally; rather, we seek the probability that an effect will occur. Estimating the likelihood that an effect will occur allows the researcher the opportunity to explore why certain effects occur in some situations but not in others. For example, a given tutorial technique may be shown to help some students perform better on an achievement test; however, when this technique is used with a different population, by a different teacher, it may not be as effective. When estimating an effect, the analyst

is not measuring the true relationship between a cause and an effect, but the likelihood that the cause created the effect.

The Logic of Causal Inference

In an analysis of causal effects, it is helpful to distinguish between the inference model used to specify the relationship between a cause and an effect and the statistical procedures used to determine the strength of that relationship. Hedges (2006) notes that “the inference model . . . specifies precisely the parameters we wish to estimate or test. . . . This is conceptually distinct from the *statistical analysis procedure*, which defines the mathematical procedure that will be used to test hypotheses about the treatment effect” (p. 3). For example, a researcher may be interested in determining whether a new curricular program is more effective than an existing program in increasing student learning outcomes. In this case, the effect to be estimated is how much better, on average, a population of students might do with the program than without the program. The goal of the analysis is to draw a causal inference or conclusion about the effect of the new program, relative to the existing program, on some outcome of interest. Once an inference model is specified, a set of statistical procedures can be used to test a hypothesis about the treatment effect (e.g., that the students in the new program score significantly higher on some measure of learning than students in the existing program).

The focus in the example above is on identifying the effect of a cause rather than the cause of an effect. This is the approach taken by Donald Rubin and his colleagues in statistics (see, e.g., Holland, 1986, 1988; Holland & Rubin, 1983; Imbens & Rubin, 1997; Rubin, 1974, 1978, 1980), and it has the advantage of being able to specify the cause and effect in question. For example, if a researcher is interested in knowing whether an innovative year-long mathematics program is

more effective in increasing the mathematics achievement of first graders than a conventional mathematics program, then an experiment can be designed in which the effects of the two mathematics programs are compared using some appropriate post-treatment measure of mathematics achievement. If children in the innovative mathematics program score higher, on average, on the mathematics assessment than do those in the conventional program, and if the students in the two groups are equivalent in all respects other than program assignment, the researcher can conclude that the higher mathematics scores are the result of the innovative mathematics program rather than of initial differences in mathematics ability. When correctly implemented, the randomized controlled experiment is the most powerful design for detecting treatment effects. The random assignment of participants to treatment conditions assures that treatment group assignment is independent of the pretreatment characteristics of group members; thus differences between groups can be attributed to treatment effects rather than to the pretreatment characteristics. Randomized experiments, however, indicate only whether there are treatment effects and the magnitude of those effects; they do not identify the mechanisms (i.e., the specific aspects of the treatments in question or of the settings in which they are implemented) that may be contributing to such effects.⁶

Designs are not developed in a vacuum; they are guided by questions that are derived from both theory and prior research. Research questions suggest boundaries for developing or selecting appropriate methods of investigation. When treatment groups can be clearly identified and there is reason to believe that one treatment may be more effective than another, an experimental approach is warranted for detecting treatment effects. Although randomized controlled experiments are designed to detect average differences in the effects of different treatments on outcomes of interest such as student achievement, researchers need to recognize that there

are a series of steps that precede the design and fielding of an experiment. In the example above, the first and most important step is to specify a theory about how students learn and what conditions contribute to student learning outcomes.

There are instances where experiments are not warranted, however. For example, if we had valid evidence in favor of a new treatment, it would be unethical to administer the old treatment.⁷ In other cases we may not have sufficient evidence to suggest that one treatment is more effective than another. In these instances, exploratory descriptive analyses of pedagogical techniques that are associated with student learning outcomes for certain populations may be a more appropriate first step. Even if there is evidence to suggest that an existing program is more effective than another, it may not be logistically, financially, or ethically feasible to conduct an experiment to test this assumption. In such instances it is sometimes possible to use large-scale datasets to approximate a randomized experiment using statistical techniques. Such quasi-experiments can be used to draw causal inferences about treatment effects based on observational data.⁸

There is a long tradition in public health that builds the case for using exploratory descriptive analyses somewhat differently, and this tradition has value for the social and education sciences as well (see Kellam & Langevin, 2003). For example, hypotheses can be generated by analyses of both cross-sectional and longitudinal data. Theory is then brought in to refine the hypotheses, which are then tested in small-scale experiments, often under highly controlled situations (i.e., high internal validity, termed *efficacy trials*). If one or more efficacy trials suggest the viability of the hypothesis, then the experiment is conducted under more “real world” conditions, in what are termed effectiveness trials. These are the clinical trials that we are familiar with.⁹ What this example shows is that there is also a place for non-experimental methods in the development of experiments.

The Formal Specification of the Causal Inference Model

Ideally, we would like to know what would have happened if an individual exposed to one treatment condition had instead been exposed to a different treatment condition. In practice this is not possible; for example, a student who completes one mathematics program cannot go back in time and complete a different program so that we can compare the two outcomes. However, Rubin and his colleagues use this hypothetical situation as the starting point for their conceptualization of causal effects.¹⁰ Rubin (1974, 1977, 1978, 1980) defined a causal effect as the difference between what would have happened to the participant in the treatment condition and what would have happened to the same participant if he or she had instead been exposed to the control condition. This conceptualization is often referred to as the counterfactual account of causality. This hypothetical causal effect is defined as

$$\delta_u = Y_{t_u} - Y_{c_u},$$

where δ_u is the difference in the effects of the conditions on unit (person) u , t refers to the treatment condition, c refers to the control condition, and Y is the observed response outcome. While this definition provides a clear theoretical formulation of what a causal effect is, it cannot be tested empirically because if we have observed Y_{t_u} we cannot also observe Y_{c_u} . This is often referred to as the fundamental problem of causal inference.

Expanding on Rubin's formulation, Holland (1986) identifies two general approaches to solving this problem, which he refers to as the *scientific solution* and the *statistical solution*. The scientific solution makes certain assumptions about the objects or units of study which are often reasonable when those objects are physical entities. In one application of the scientific solution, an object or objects are first exposed to treatment 1 and the outcome of interest is measured; the

object is then exposed to treatment 2 and the outcome is measured. The causal effect in this case is defined as the difference between the outcome that unit \mathbf{u} displayed at time 1 under the treatment condition and the outcome that same unit displayed at time 2 under the control condition: $\delta_{\mathbf{u}} = Y_{t1}(\mathbf{u}) - Y_{c2}(\mathbf{u})$. Two assumptions are made in this case. The first is *temporal stability*, which means that there is a constancy of response over time. The second is *causal transience*, which means that the effect of the first treatment is transient and does not affect the object's response to the second treatment.

A second way of applying the scientific solution is to assume that the objects under study are identical in all respects. It therefore makes no difference which unit receives the treatment. This is the assumption of *unit homogeneity*. Under this assumption, the causal effect can be determined by calculating the difference between $Y_t(\mathbf{u}_1)$ and $Y_c(\mathbf{u}_2)$, where $Y_t(\mathbf{u}_1)$ is the outcome of unit 1 under the treatment condition and $Y_c(\mathbf{u}_2)$ is the outcome of unit 2 under the control condition. The assumption of unit homogeneity is often made in the physical sciences and engineering, where the objects of study have a high degree of uniformity.

When human beings are the focus of study, these assumptions are usually much less plausible. For example, a participant's response to a treatment may vary according to the time at which the treatment is delivered, invalidating the assumption of temporal stability. Similarly, a participant's response to one treatment condition may affect his or her response to a second treatment condition, invalidating the assumption of causal transience. Even if participants in an experiment are identical twins and are known to have identical genes, they may differ in other ways that may affect their responses (e.g., knowledge, experience, motivation); the assumption of unit homogeneity is rarely plausible when the unit of analysis is the person.

The statistical solution to the fundamental problem of causal inference takes a different approach. Rather than focusing on specific units, the statistical approach estimates an *average* causal effect for a population of units (i.e., participants). The population average causal effect thus becomes

$$\delta = \mathbf{E}(Y_t - Y_c),$$

where Y_t is the average outcome for participants in the treatment group, and Y_c is the average outcome for participants in the control group.¹¹ For this solution to work, however, individuals or organizational elements (e.g., classrooms or schools) in the treatment and control groups should differ only in terms of treatment group assignment, not on any other characteristic or prior experience that might potentially affect their responses. For example, if the outcome of interest is mathematics achievement, and only high-achieving students are assigned to the treatment condition (e.g., an innovative mathematics program) while low-achieving students are assigned to the control condition (a conventional mathematics program), higher average mathematics scores for students in the treatment group could be due to the higher initial achievement of these students rather than to the program of instruction. However, if students are randomly assigned to the treatment and control conditions, one could expect that treatment group assignment would, *on average*, over repeated trials, be independent of any measured or unmeasured pre-treatment characteristic. Because random assignment assures, in expectation, equivalence between groups on pretreatment characteristics, if students in the treatment group score higher on a post-treatment assessment of mathematics achievement, the researcher can conclude, at least in large samples, that this effect is due to differences in the program of instruction rather than to differences in the characteristics of students in the two groups.

This example represents the ideal case and assumes that the innovative program is implemented with fidelity, that students do not move between treatment and control groups, and that they remain in their assigned groups for the longevity of the study. In practice, problems in implementing experiments can present substantial threats to their validity and need to be addressed. Some of these problems and proposed solutions to them are discussed in the next section.

The statistical solution to the fundamental problem of causality relies on the assumption of independence between pretreatment characteristics and treatment group assignment. This independence is difficult to achieve in nonrandomized studies. Statistical models typically are used to adjust for potentially confounding variables (i.e., characteristics of students, classrooms, or schools that predict treatment group assignment and also predict outcomes) when outcomes for different groups are compared. However, as Raudenbush (2005) points out, “No matter how many potential confounders [analysts] identify and control, the burden of proof is always on the [analysts] to argue that no important confounders have been omitted” (p. 28). Because randomized assignment to treatment groups takes into account observed and unobserved characteristics, such controls are not necessary. This is why randomized field trials are often considered the “gold standard” for making causal inferences.

Criteria for Making Causal Inferences

In elaborating on Rubin’s causal model, Holland (1986) identifies four criteria for making causal inferences. He relies on examples from controlled experiments to illustrate these criteria. “It is not that an experiment is the *only* proper setting for discovering causality,” he writes, “but I do feel that an experiment is the *simplest* such setting” (p. 946).

Causal Relativity

The effect of a cause must always be evaluated relative to another cause. In a controlled experiment, for example, the outcomes for a given treatment or intervention (one cause) are always defined relative to an alternative treatment or control condition (a second cause). Thus, in evaluating whether an innovative mathematics program is effective in increasing mathematics achievement, the outcomes of the program must be compared with the outcomes from some existing program. The question is not simply whether a program is effective but whether it is more effective than some other program.

Causal Manipulation

Each participant must be *potentially* exposable to the causes under consideration (i.e., the treatment and control conditions). For example, the instruction a student receives can be said to be a cause of the student's performance on a test, in the sense used by Holland, whereas the student's race or gender may not. Race and gender are attributes of the student that cannot typically be altered or manipulated and thus cannot be said to be causes of differences in mathematics achievement. In contrast, a student can potentially be exposed to different types of instruction.

Temporal Ordering

Exposure to a cause must occur at a specific time or within a specific time period. In determining whether students who participate in an innovative mathematics program earn higher scores on a mathematics assessment than those who participate in an existing mathematics program, the researcher must obtain students' mathematics scores after their exposure to either the treatment or control condition. In this instance, the

outcome variable (post-exposure mathematics scores) serves as a measure of the effect of the treatment. Variables thus divide into two classes: pre-exposure—those whose values are determined prior to exposure to the cause (the treatment or control condition)—and post-exposure—those whose values are determined after exposure to the cause.

Elimination of Alternative Explanations

The researcher must be able to rule out alternative explanations for the relationship between a possible cause or treatment and an effect (as measured by an outcome of interest). In controlled experiments, this is accomplished in part through the random assignment of participants to treatment and control groups. Although there may be difficulties in implementing randomization (an issue addressed later), in the ideal situation, when randomization is effective, treatment and control groups are essentially equivalent with respect to pretreatment characteristics. Any differences in the outcomes of the two groups can thus be attributed to differences in treatment assignment rather than to other causes such as pretreatment differences in ability, achievement, learning experiences, or other characteristics.

Issues in the Design and Fielding of Randomized Experiments

Sampling Imbalances

Complete equivalence on all pretreatment characteristics is rarely achieved even when random assignment is used. As Raudenbush (2005) notes, random assignment does not necessarily ensure that there will be no differences between treatment and control groups: “It is true, by chance, differences will exist among randomly formed groups; and these

differences may in fact, be quite large in small samples. But such chance differences are fully accounted for by well-known and comparatively simple methods of statistical inference” (p. 27). Typically, however, researchers compare treatment and control groups on key variables (e.g., demographics such as gender, race, socioeconomic status [SES], and so on) to make sure that randomization has been effective (see, e.g., Krueger, 1999; Nye, Konstantopoulos, & Hedges, 2000). Another way in which this issue is addressed is through replication of studies and cross-study comparisons. The comparison of results across randomized controlled experiments allows researchers to obtain more accurate estimates of causal effects and it increases the confidence that the result is real, not due to sampling fluctuations.

Specific Versus Average Effects

Because the statistical solution to the fundamental problem of causal inference estimates an average effect for a population of participants or units, it tells us nothing about the causal effect for specific participants or subgroups of participants. Holland (1986) observes that this average effect “may be of interest for its own sake in certain types of studies. It would be of interest to a state education director who wanted to know what reading program would be the best to give to all of the first graders in his state. The average causal effect of the best program would be reflected in increases in statewide average reading scores” (p. 949). But, in other cases, researchers might be interested in knowing whether certain programs would help to close achievement gaps between particular groups of students.¹² In such cases, researchers would be less interested in knowing whether the treatment produces a constant effect (one relevant to every participant in the study) and more interested in knowing whether treatment effects vary across subgroups of students. Holland notes that the assumption of a constant

effect can be checked by dividing the sample into subpopulations; an average causal effect can then be estimated for each subgroup.

Atypical Responses

Rubin (1986) observes that two additional assumptions must be valid for randomization to yield unbiased estimates of causal effects. These are ideal criteria that are frequently not met in educational and other social science research.¹³ However, they are important because they help to guide researchers in the design of their studies.

First, the mechanism for assigning participants to treatment and control groups should not affect their responses. In many studies this assumption may not be valid. For example, if disadvantaged high school students are told that they have been chosen at random to participate in a program designed to encourage college attendance, they may respond differently than if they are told that they were selected on the basis of their academic achievement. Those who believe they were selected on the basis of merit may be more motivated to participate in the program and more likely to apply to college. If the goal is to determine whether the program is effective in increasing college-going rates for disadvantaged students, then students' knowledge of the assignment mechanism may affect the outcome of interest.

Second, the responses of participants should not be affected by the treatment received by other participants. For example, if participants in the control group know that those in the treatment group are participating in a promising new program, they may become demoralized because they are not receiving the program. Alternatively, they may respond competitively and do better than they might have otherwise. Estimates of treatment effects would be biased upward in the first instance and downward in the second.

Researchers have developed several strategies for minimizing atypical responses. If participants are only aware of the condition in which they participate, their responses to the treatment or control condition will be unaffected by the use of random assignment. In practice, however, this solution may not be feasible, particularly if informed consent procedures require that participants be told that they will be randomly assigned to different treatment conditions. Another strategy for minimizing atypical responses is the use of masking or blinding procedures: Participants are not told whether they have been assigned to the treatment or control group. The experimenter is also, in many cases, unaware of which group participants are assigned to, a procedure known as double-blinding. In triple-blinding, not even the data analyst knows which participants were assigned to the treatment and control conditions. However, masking procedures often are not feasible in real-world situations, where participants may need to know that they are receiving a particular treatment or benefit for the experiment to work (e.g., financial assistance). In other cases, participants may be able to identify their treatment group assignment despite masking procedures. A third strategy that is sometimes used in randomized studies is to offer participants in the control group a program that is equally as attractive as the treatment condition but has no relation to the response of interest.¹⁴

Implementing Randomized Assignment

Implementing experiments with randomized assignment can also present problems for researchers, such as breakdowns in randomization, treatment noncompliance, and attrition.¹⁵ Researchers who use randomized designs are familiar with these potential problems, and considerable strides have been made to overcome them (see Shadish, Cook, & Campbell, 2002). The value of conducting experiments in education and

an assessment of the objections to doing them are discussed by Cook (2002, 2007).

Problems in conducting experiments are also common in other types of research such as large-scale surveys. For example, when random sample of schools are drawn, some schools may choose not to participate, some may drop out during data collection, and some may fail to comply with survey procedures and administration. Methodologists have developed a number of procedures for addressing such problems, although such solutions are not always adequate. Next, we review some of these problems and ways in which they have been addressed in randomized field trials.

Breakdowns in randomization. There is sometimes resistance to randomization, particularly when a promising new treatment is being tested. For example, parents may lobby to have their children included in a promising new program. Such problems can be avoided by monitoring both the randomization process and the actual treatment received by each participant following randomization. Another strategy to minimize breakdowns in randomization is to isolate the units under study. For example, when different treatments are given to different schools (high isolation of units), it is less likely that breakdowns in randomization will occur than when different treatments are given to different classrooms within the same school (low isolation of units).¹⁶

Treatment noncompliance. Individuals who are randomly assigned to treatment and control conditions may never actually receive treatment. Some may simply fail to show up for the particular program to which they have been assigned. For example, randomly assigning students (families) to receive a Catholic school voucher does not mean that they will use the voucher (e.g., because of family beliefs about public education, proximity to alternative schools, or other reasons). There are several practical ways to encourage participation, such

as providing incentives, removing obstacles (e.g., providing transportation), and including only those who are willing to participate. Even when such steps are taken, however, some of those selected for participation in a study may still fail to participate.

Three statistical strategies have been used in cases where there is participant noncompliance. In the first approach, known as the *intention to treat analysis*, the mean responses of those assigned to the treatment condition (regardless of whether they actually received treatment) are compared with the mean responses of those assigned to the control condition. Since noncompliers do not receive treatment, the mean for the treatment group is typically lower than it would be if all individuals assigned to the treatment condition had actually received treatment, assuming that the treatment has positive effects. As a result, this analysis usually yields conservative estimates of treatment effects. The second approach eliminates individuals assigned to the treatment condition who do not actually receive the treatment. Unless it can be shown that those who drop out of the treatment condition are a random sample of the participants in that condition, this analysis will yield a biased estimate of the treatment effect.

The third strategy focuses on estimating the intention to treat effect for the subset of participants who are “true compliers.” True compliers are those who will take the treatment when assigned it and will take the control when assigned it. Noncompliers are those who will not take what they are assigned, whether it is the treatment or the control condition (Angrist, Imbens, & Rubin, 1996; Bloom, 1984; Little & Yau, 1998). Noncompliers are of three possible types: never-takers, who will never take treatment no matter what condition they are assigned to; always-takers, who will always take treatment no matter what condition they are assigned to; and defiers, who will always do the opposite of what they are assigned (these people are often assumed not to exist or to be few in number).

Because only the true compliers can be observed both taking and not taking treatment, they are the only subgroup for which we can learn about the effect of taking treatment versus being in the control group.

An additional assumption yields the *instrumental variable estimate* for the noncompliers: There is no effect of the assignment on what would be observed.¹⁷ That is, the “exclusion restriction” says that if the assignment to treatment versus control cannot affect which condition a participant will take (i.e., the noncompliers will do what they want regardless of the condition to which they are assigned), it cannot affect the participants’ outcome. Extensions of this approach that weaken various assumptions and deal with complications, such as missing data, also exist (e.g., Imbens & Rubin, 1997; Rubin, 1998; Frangakis & Rubin, 1999; Hirano, Imbens, Rider, & Rubin, 2001).

Attrition. In many cases, individuals selected for study initially participate but later drop out. It is not always possible to maintain contact with all participants, and those who are contacted may refuse to continue their participation. Researchers have been aware of this issue for some time (see, e.g., Jurs & Glass, 1971) and have developed strategies for estimating the effect of attrition on the outcomes of interest.

Little and Rubin (2002) review several techniques for dealing with missing data, including data missing due to attrition. They also identify mechanisms that lead to missing data. Identifying such mechanisms is important in selecting an appropriate method for handling missing data. Little and Rubin identify three categories of missing-data mechanisms: missing completely at random, missing at random, and not missing at random. Data are said to missing completely at random (MCAR) if the probability of having missing data on an outcome variable Y is not dependent on Y or on any of the variables included in analysis. If data are missing completely at

random, estimates of treatment outcomes are unbiased. Data are said to be missing at random (MAR) if the likelihood of having missing data is related to the observed values of other variables included in the analysis. In this case, the missing data are unrelated to Y after controlling for other variables. For example, individuals who drop out of a study may have lower incomes than those who remain in the study. However, if this pattern is accounted for by relationships among observed variables, such as race and education, then data are missing at random, and estimates of treatment effects are unbiased. In cases where data are not missing at random (NMAR), the probability of having missing data is dependent on both observed and unobserved values of the outcome Y . For example, attrition may depend on values that were recorded after dropout. If only individuals with incomes below a certain level drop out of the study, and data on income are available only for those who remain in the study, then estimates of treatment effects will be biased.

As Foster and Fang (2004) note in their review of methods for handling attrition, “In any given situation, the actual missing data mechanism is unknown. However, . . . the evaluator can assess the plausibility of the alternative assumptions based on what he or she knows about the evaluation and the population included and what they reveal about how the missing data were generated” (p. 438). In cases of attrition from randomized experiments, researchers typically have information on the pretreatment characteristics of participants as well as their treatment group assignments and can conduct analyses to determine whether there are any significant differences on pretest measures between those who drop out of the study and those who remain in the study. Significant differences between leavers and stayers indicate that the characteristics of those who leave the study differ from the characteristics of those who remain in the study, suggesting that the study findings may not generalize to the population of interest.

When the characteristics of participants who drop out of the treatment group differ from the characteristics of those who drop out of the control group, the estimate of the treatment effect may be biased. In such cases, researchers should cautiously explore techniques for adjusting for potential bias (e.g., imputing missing values, modeling the effects of attrition on responses, and estimating maximum and minimum values to bracket the treatment effect).¹⁸

Detecting Treatment Effects

Statistical power. In the context of experimentation, *power* refers to the ability of a statistical test to detect a true treatment effect, that is, to detect a treatment effect when it in fact exists. Existing reviews of the literature indicate that insufficient power for making statistical judgments is a problem with studies in several fields, including medicine (see, e.g., Cuijpers, 2003; Dignam, 2003; Halpern, Karlawish, & Berlin, 2002; Rossi, 1990; West, Biesanz, & Pitts, 2000). This is a serious problem, given both the cost of conducting randomized experiments and the failure of underpowered studies to yield consistent answers. As Dignam argues with respect to randomized clinical trials:

It is imperative that [randomized experiments] be carefully designed with respect to statistical power so as not to obtain equivocal findings that fail to answer the fundamental question of a new treatment under consideration. Underpowered studies can cause delay or even abandonment of promising avenues of treatment, and even a “negative” that is adequately powered is an important finding in that energy and resources can be directed into other more promising alternatives (p. 6).

There are several methods for increasing statistical power. Increasing sample size is the most obvious, but practical considerations such as cost, available resources, and access to

populations of interest (e.g., children with learning disabilities) may restrict this option for researchers. Other approaches to increasing statistical power include using more reliable measures, minimizing participant attrition, increasing the fidelity of treatment implementation, and measuring and adjusting for characteristics related to the outcome of interest.¹⁹

Hedges (2006) observes that increasing the significance level (denoted by α) used in statistical testing is one way to increase power without increasing sample size. He notes that “statistical decision theory recognizes two kinds of errors that can be made in testing. The significance level controls the rate of Type I Errors (rejecting the null hypothesis when it is true). Setting a low significance level [such as the conventional $\alpha = .05$] to control Type I Errors [concluding there are treatment effects when there are in fact no effects] actually increases the rate of Type II Errors (failing to detect effects that are actually present)” (p. 20). He argues that when resources are limited, as is the case in many intervention studies, “selection of a significance level other than .05 (such as .10 or even .20) may be reasonable choices to balance considerations of power and protection against Type I Errors” (p. 20).

The use of stratified randomization can also increase power. In small-scale randomized studies, treatment and control groups may not be well matched on certain characteristics such as age or gender. In such cases, the use of stratified randomization can increase the balance between treatment and control groups without sacrificing the advantages of randomization. Stratified randomization is achieved by performing separate randomizations with each subset of participants (e.g., as defined by gender, age, and pretreatment assessment scores).

Software packages now available for making power calculations allow researchers to compute the sample size needed to detect a treatment effect of a given size in advance of conducting an experiment. Often, an estimate of the effect size

for a particular treatment/intervention is available from prior research, especially meta-analyses. Following Cohen (1988), many researchers also rely on general “rules of thumb” about what constitutes large, medium, and small effect sizes. Tools for computing statistical power for multilevel studies (e.g., students nested within schools) are less widely available, but there have been some advances in this area (McDonald, Keesler, Kauffman, & Schneider, 2006). Researchers have found that increasing sample sizes at higher levels (e.g., schools or sites) increases power more effectively than increasing sample sizes at lower levels (e.g., students within schools; Raudenbush & Liu, 2000). Adding another site to a study, however, may be considerably more costly than adding participants within a site.

One problem faced by education researchers has been a lack of definitive knowledge about school-level characteristics associated with academic achievement. To address this problem, Hedges and his colleagues, with support from the IERI, have begun to identify factors contributing to within- and between-school variation in academic achievement. Reanalyzing data from surveys administered to nationally representative samples of students, they are conducting analyses of variation in mathematics and reading achievement “separately (by subject matter) for different grade levels, regions of the country and urbanicity (coded as urban, suburban, or rural)” (Hedberg, Santana, & Hedges, 2004, p. 5). They have found that academic achievement varies significantly at the school as well as the individual level; achievement also varies significantly by region of the country, urbanicity, and students’ stage in the life-course. These findings, which the authors plan to compile into an almanac, should be useful to researchers in designing adequately powered studies.

Generalizability in experimental studies. Experiments provide the best evidence with respect to treatment effects;

they can, however, yield results that are local and particular. Most researchers, however, are interested in knowing whether these effects generalize to other populations and settings. They may also want to know whether such effects generalize to other outcomes and treatment implementations. Researchers often rely on a combination of approaches to maximize the generalizability of their results.

Statistically, the only formal basis for ensuring the generalization of causal effects is to randomly sample from a well-defined population (not to be confused with the random assignment of participants to treatment and control groups). This is accomplished through an enumeration of the population of interest (e.g., the U.S. population of high school students). A random sample is then drawn from this population. Although formal probability sampling is viewed as the ideal with respect to generalizing to populations and settings, it is extremely difficult to implement in practice. In many cases, the population of interest cannot be precisely enumerated (e.g., neglected children). Even when enumeration is possible (e.g., from administrative records), it may not be possible to locate all members of the population or to persuade all individuals (or schools, or districts) who have been randomly selected to participate in an experiment with random assignment (Shadish, Cook, & Campbell, 2002; West et al., 2000). Randomly selecting settings (e.g., schools), while possible, may be difficult to implement in practice due to the cost of studying more than a few sites. For these reasons, there have been few experiments where randomly selected persons and settings are, in turn, randomly assigned to treatment and control conditions.

Because of the practical difficulties of implementing random sampling, researchers often rely on study replication to generalize results from single studies to other outcomes, populations, or settings. In some cases, a single researcher or team of researchers may carry out a program of research on the same topic that systematically varies key variables from study

to study to identify limits to generalization. Multi-site experiments, where randomization of participants to treatment and control groups is carried out at several sites, is another approach to the generalization of causal effects. Raudenbush and Liu (2000) note that “the multisite trial enables a formal test of the generalizability of the treatment impact over the varied settings in which the treatment may ultimately be implemented if its early results prove promising” (p. 199).

Additional Design Issues

Even if randomized experiments are implemented with fidelity, are sufficiently powered, and are generalizable, such experiments may fail to yield useful results. The outcomes being tested may be inadequately measured, the intervention may be poorly conceptualized, or a well-thought-out intervention may not be targeted to the students who could benefit most from it. As Raudenbush (2005) argues, “The randomized experiment becomes a powerful tool for warranting causal effects [only] after a rather protracted process has identified the most promising interventions for changing the most important outcomes for target children in settings of interest” (p. 29). Given the expense of fielding large-scale randomized experiments, results of studies using a variety of methods at different scales are needed to inform their design. Raudenbush points to the importance of other relevant research that can be used to inform the design of large-scale randomized experiments, including defining relevant outcomes, identifying promising interventions, and targeting specific populations of interest.

Defining relevant outcomes. Large-scale assessments of student learning such as the National Assessment of Educational Progress (NAEP), the Program for International Student Assessment (PISA), and the Third International Mathematics and Science Study (TIMSS) are useful in identifying gaps in

student achievement. Smaller-scale studies that assess aspects of students' conceptual understanding, content knowledge, and procedural knowledge in different subject areas are also important in identifying gaps in student proficiency. Without such studies, researchers and policymakers would not know what outcomes most need to be improved and for which students. Raudenbush argues that

a failure to attend systematically to this process of creating good outcome measures [may be] the Achilles heel of evaluation research on instructional innovation. If the process is ignored, trivialized, or mismanaged, we'll be measuring the wrong outcome with high reliability, the right outcome with low reliability, or, in the worst case, we won't know what we are measuring. If we don't know what we are measuring, the causal question (Does the new intervention improve achievement?) is meaningless. If we measure the right outcome unreliably, we will likely find a new program ineffective even if it is effective. If we measure the wrong outcome reliably, we may find that the intervention "works," but we'll never know whether it works to achieve our goals. (2005, p. 29).

Identifying promising interventions. Studies that identify interventions that are promising candidates for large-scale randomized trials are another important component of research designed to improve student learning. Raudenbush notes that a variety of methods can be used to identify promising interventions that could be implemented on a large scale:

Detailed descriptions of expert practice often supply key new ideas for how to intervene. Small-scale implementation studies or even careful small-scale randomized studies can provide preliminary evidence about whether a new approach can, under ideal conditions, produce an effect for a sample that probably is not representative. Secondary analysis of large-scale data can provide important evidence of promising practice. (2005, p. 29)

Targeting populations of interest. In designing large-scale randomized experiments, information is also needed on the populations of students who are in the greatest need of educational interventions or would benefit most from new approaches to teaching and learning. A variety of methods have been used to determine where achievement gaps exist and for what populations of students, as well as what settings, organizational approaches, and instructional methods might help to reduce such gaps.

Fielding Randomized Experiments in Educational Settings

To assist the education research community in conducting randomized controlled trials, the NRC (2004b) sponsored a workshop and issued a report on the practical problems of conducting such studies. This report discusses a number of pragmatic issues that must be addressed in conducting randomized controlled trials (RCTs) in educational settings: meeting ethical and legal standards, establishing adequate sample sizes and recruiting participants, grounding the study in the relevant educational context, and securing adequate resources.²⁰ Each of these issues is important to the success of RCTs in obtaining valid evidence of treatment effects.

Researchers, including those conducting randomized controlled trials, are now required to meet rigorous legal and ethical standards for conducting research with human subjects. For example, in implementing a randomized controlled experiment with students, researchers must inform parents of the goals and nature of the research and obtain their consent for their children's participation. The researchers also must demonstrate that procedures are in place to ensure that individual information and identifying data are confidential. In some cases, researchers may have trouble obtaining approval from institutional review boards (IRBs) responsible for ensuring that studies meet legal and ethical standards, particularly if

an intervention has the potential to harm participants (e.g., an intervention involving a vigorous exercise program).

Despite such safeguards, many potential participants have ethical concerns about RCTs that have received IRB approval, particularly when randomized assignment is perceived as denying beneficial services or interventions to some students. Researchers need to be aware of and address such concerns both in designing and in implementing RCTs. One way in which researchers have dealt with this issue at the school level is to include participating schools in both the treatment and control conditions. For example, in designing and implementing *Success for All*, Slavin and his colleagues randomly assigned schools to treatment and control conditions (see, e.g., Slavin & Madden, 2001, in press; Slavin, Madden, Karweit, Dolan, & Wasik, 1992; Slavin, Madden, Dolan, Wasik, Ross, & Smith, 1994). However, the intervention was implemented in first grade in one set of schools, with first graders in the other schools serving as the control group. In schools that had served as the first grade control group, the intervention was implemented in third grade, with the first grade intervention group serving as the control. Schools in both groups thus had the opportunity to participate in an intervention that might prove beneficial to students. As Slavin and others have noted, developing close and respectful partnerships with schools and school districts is an effective way to become aware of and address such concerns.

Ensuring that samples are sufficiently large to detect effects can be particularly difficult in certain educational settings. For example, in urban settings, high rates of mobility can make it difficult for researchers to recruit and retain sufficient numbers of study participants. Obtaining consent from parents may also prove to be difficult. Given enough time, researchers can meet with parents to inform them about the study and address their concerns. Building partnerships with schools can facilitate the process of recruitment, but establishing

such partnerships can be a lengthy process, requiring that relationships be established years in advance of the implementation of a randomized controlled trial.

Grounding a study in the relevant educational setting (e.g., addressing questions of particular interest to participating schools and teachers) can help to build partnerships with schools that support the implementation of randomized experiments. Determining what questions are most pressing for particular schools and teachers requires a familiarity with the political and economic environment of schools, the schools' missions and goals, and the particular challenges they face. For example, in designing interventions to reduce drug abuse, delinquency, and school failure, Kellam and his colleagues (Kellam & Van Horn, 1997; Kellam, Ling, Merisca, Brown, & Ialongo, 1998) targeted Baltimore schools that were struggling to find solutions to these problems. This partnership with the Baltimore school system has made it possible for Kellam and his colleagues to conduct three generations of randomized controlled trials.

Questions about whether a widely used educational intervention has systematic effects on student learning outcomes are often best answered by large-scale randomized field trials. However, such studies can be costly to implement, particularly when treatments are assigned at the school level, requiring the inclusion of a sufficient number of schools to detect treatment effects. When trying to measure changes in performance, such as gains in achievement, accurately assessing growth requires that trials be conducted over a sufficient period of time, typically at least a year, which also adds to the costs of fielding the study. Given such costs, it is particularly important that these studies be well designed, have a strong theoretical grounding, and be adequately informed by prior research. In some cases, the research base may be insufficient to justify fielding an RCT. In such cases, researchers may need to conduct preliminary descriptive studies or smaller-scale

randomized studies to determine whether an intervention is sufficiently promising to warrant large-scale implementation and the development of adequate measures for the variables of interest. In other cases, RCTs may not be feasible, either because of costs or for ethical reasons, and researchers may need to approximate randomized experiments with observational data. Analyzing data from large-scale datasets can be useful in both instances by providing tentative results needed to design and implement effective large-scale randomized trials or by providing alternative methods for making valid causal inferences with observational data.

Chapter 2 Notes

- 6 Randomized experiments can be used in conjunction with other methods to examine the mechanisms that help explain causes.
- 7 In education experimental studies that involve treatment and control groups, it is nearly always the case that the “control group” means business as usual. It is rare for an experiment to withhold treatment.
- 8 Several of these techniques are described in Section 3.
- 9 We thank George Bohrnstedt for this point.
- 10 There is a long history of work in statistics that has focused on causal inference. Rubin’s model builds on this tradition, which includes early work on experimental design by Fisher (1935), Neyman (1923, 1935), Cochran and Cox (1950), Kempthorne (1952), and Cox (1958a, 1958b).
- 11 Technically, E is the expected value or long-run average of the difference on Y between the treatment and control groups.
- 12 One advantage of descriptive studies that rely on large-scale nationally representative datasets is that it is possible to examine subgroups of participants because samples are large and representative of the population.
- 13 These criteria are referred to as the stable-unit-treatment-value assumption (SUTVA).
- 14 See, for example, Higginbotham, West, and Forsyth (1988) and West, Biesanz, and Pitts (2000) for discussions of atypical reactions and strategies for dealing with them.
- 15 See West et al. (2000) for a useful review of several of these problems.
- 16 When schools or other groups are assigned to treatment conditions, randomization occurs at the group rather than the individual level (see Raudenbush, 1997, for a discussion of cluster randomization). The assumption that individual responses are independent is not valid in this situation because individuals within the same group are more likely to provide similar responses than individuals in different groups. This problem is now routinely dealt with by using hierarchical linear modeling procedures, which simultaneously provide estimates of causal effects at both the individual and group levels, while

correcting for the nonindependence of responses within groups (Bryk & Raudenbush, 2002).

- 17 Instrumental variable approaches are discussed in Section 3. We thank Donald Rubin for writing the section on estimating complier average causal effects and for offering additional explanation of this technique.
- 18 Several different software programs are available for computing missing values: SOLAS™ for Missing Data Analysis (available at <http://www.statsol.ie/solas/solas.htm>); SAS-based IVEware (available at <http://www.isr.umich.edu/src/smp/ive>); MICE (Multiple Imputation by Chain Equations, available at <http://www.multiple-imputation.com>); and NORM and related programs (available at <http://www.stat.psu.edu/%7Ejls/misoftwa.html>).
- 19 See Shadish, Cook, and Campbell (2002, pp. 46–47) for an overview of strategies for increasing power.
- 20 There is a common misconception that randomized experiments are always expensive. In the context of this report, we are discussing the costs of conducting large-scale, multi-site randomized experiments. Regardless of whether studies employ an experimental or a quasi-experimental approach, most national multi-site, longitudinal collections are expensive. We thank Thomas Cook for pointing this out.