Psychological Methods
2010, Vol. 15, No. 1, 18–37

1082-989X/10/$12.00 DOI: 10.1037/a0015917

Campbell's and Rubin's Perspectives on Causal Inference

Stephen G. West and Felix Thoemmes
Arizona State University

Donald Campbell's approach to causal inference (D. T. Campbell, 1957; W. R. Shadish, T. D. Cook, & D. T. Campbell, 2002) is widely used in psychology and education, whereas Donald Rubin's causal model (P. W. Holland, 1986; D. B. Rubin, 1974, 2005) is widely used in economics, statistics, medicine, and public health. Campbell's approach focuses on the identification of threats to validity and the inclusion of design features that may prevent those threats from occurring or render them implausible. Rubin's approach focuses on the precise specification of both the possible outcomes for each participant and assumptions that are mathematically sufficient to estimate the causal effect. In this article, the authors compare the perspectives provided by the 2 approaches on randomized experiments, broken randomized experiments in which treatment nonadherence or attrition occurs, and observational studies in which participants are assigned to treatments on an unknown basis. The authors highlight dimensions on which the 2 approaches have different emphases, including the roles of constructs versus operations, threats to validity versus assumptions, methods of addressing threats to internal validity and violations of assumptions, direction versus magnitude of causal effects, role of measurement, and causal generalization. The authors conclude that investigators can benefit from drawing on the strengths of both approaches in designing research.

Keywords: randomized experiment, quasi-experiment, observational study, research design, causal inference

In this article, we provide an introduction to Donald Campbell's (Campbell, 1957; Shadish, Cook, & Campbell, 2002) and Donald Rubin's (Holland, 1986; Rubin, 1974, 2005) perspectives on causal inference. Campbell's perspective has dominated thinking about causal inference in psychology, education, and some other behavioral sciences. Rubin's causal model (a.k.a., the potential outcomes model; e.g., Rubin, 1974, 1978, 2005, 2006b) has become an important perspective on causal inference in economics, medicine, public health, and statistics. The two perspectives share many foundational ideas. Both perspectives attempt to understand the effect of a treatment of interest relative to a comparison treatment on an outcome (posttest). Both perspectives have attempted to eliminate the possible effects of other potential influences on the outcome so that the causal effect of the treatment could be isolated. Both perspectives have basic philosophical underpinnings in Hume's (Hume 1748/2007; Lewis, 1973) counterfactual model.[1] However, in their development, the two perspectives have also developed unique and complementary emphases. Campbell's perspective has emphasized the identification of potential threats to the validity of inferences prior to conducting a study and the addition of features to the basic design that can prevent those threats from occurring or can rule out those threats as alternative explanations of the findings. However, the use of these design elements can sometimes introduce ambiguity about the magnitude of the causal effect, as we describe below. In these cases, Campbell's perspective has historically focused more on determining the *direction* of the causal effect, for example, is $\mu_T > \mu_C$, where μ_T and μ_C are the population means of the treatment and comparison groups, respectively, after the effect of other potential influences have been ruled out. In contrast, the emphasis within Rubin's perspective has been on estimating the exact *magnitude* of the causal effect. Rubin's perspective has emphasized making specific, ideally verifiable assumptions that are mathematically sufficient to permit the researcher to make proper analytic adjustments for common issues that occur in research so that the causal effect of interest can be precisely estimated. On the other hand, the precision demanded by Rubin's perspective may sometimes limit its applicability and the ability of researchers to generalize their findings.

We begin by briefly reviewing the two perspectives. We restrict our focus to those ideas that are presented in the major writings of the two perspectives; we do not consider ideas that may be implicit but which have not been clearly articulated to date. Throughout the article we will initially introduce the ideas from each perspective separately, followed by a comparison. Most researchers in psychology are familiar with Campbell's perspective (particularly the early developments portrayed in Campbell & Stanley, 1966). Few

Stephen G. West and Felix Thoemmes, Psychology Department, Arizona State University.

Felix Thoemmes is now at the Graduate Program in Research, Measurement, and Statistics, Department of Educational Psychology, Texas A&M University.

Earlier versions of portions of this article were presented at the symposium on causality, Altes Schloss Dornburg, Jena, Germany (July 2006 and July 2008). We thank Leona Aiken for comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Stephen G. West, Psychology Department, Arizona State University, Tempe, AZ 85287-1104. E-mail: sgwest@asu.edu

[1] Although the term *counterfactual* was used in the early writings about his approach (e.g., Holland, 1986), Rubin (2005) now prefers the term *potential outcome*, which we will follow in this article.

psychologists have had exposure to Rubin's causal model. We introduce several of its central ideas, keeping the presentation at a nontechnical level. Rubin's perspective has opened up new approaches to the analysis of difficult problems of causal inference in "broken" randomized experiments in which problems, such as attrition or treatment nonadherence, have occurred. It also provides a strong mathematical basis for the analysis of studies in which nonrandomized treatment groups are compared. We compare Campbell's and Rubin's perspectives in terms of their approaches to randomized experiments, broken randomized experiments, and nonrandomized studies. The two perspectives have different emphases and, hence, are associated with different strengths. We believe that the strengths of the two perspectives are often complementary. Consideration of both perspectives can inform the design and analysis of research in the behavioral sciences, particularly in field settings.

Campbell's Approach

Overview

Campbell (1957) developed a practical approach to causal inference that follows the approach of a working scientist. Campbell considered the full range of pre-experimental, quasi-experimental, and experimental designs used by basic and applied researchers in the behavioral sciences. Over a 50-year period, Campbell and his colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979; Reichardt, 2006; Shadish et al., 2002) have collected and refined a list of threats to validity, representing "an accumulation of our field's criticisms of each other's research" (Campbell, 1988, p. 322). A key concept is the idea of *plausible* threats to validity, factors that may potentially undermine some aspect of the causal inference process in the specific research setting:

> We took the position that there could be lots of threats to validity that were logically uncontrolled but that one should not worry about unless they were plausible. The general spirit was that any interpretation of a body of data or research should be regarded as innocent until judged guilty for plausible reasons, as determined through the scientific method of mutual criticism. (Campbell, 1988, p. 317)

The task for researchers is to identify plausible threats to validity and then to include design elements, analyses, or other features in their research that can potentially rule out those specific threats. Otherwise stated, the specific alternative explanation of the results would be logically eliminated by the incorporation of the design element and the obtained pattern of results. No proof exists that the system of threats to validity is complete, or that methods of addressing them always work, but the approach has evolved to be quite thorough and of great practical use in behavioral science research.

More recent statements of Campbell's perspective (e.g., Cook and Campbell, 1979; Shadish et al., 2002) consider four validity types: statistical conclusion validity, internal validity, construct validity, and external validity. Given our focus on causal inference, we have a strong focus on internal validity here. We designate X as an indicator of treatment (e.g., 1 = Treatment [T]; 0 = Control [C]) and Y as the outcome (dependent) variable. The central concern of internal validity is whether the relationship between the treatment and the outcome is causal in the population under study. Does the manipulation of X produce change in Y? Or, does some other influence produce change in Y? Note that internal validity does *not* address the specific aspect(s) of the treatment that produce the change nor the specific aspect(s) of the outcome in which the change is taking place—nor does it address whether the treatment effect would hold in a different setting, with a different population, or at a different time. These issues are questions of construct validity and external validity, respectively.[2]

Internal Validity

Threats to internal validity. Threats to internal validity identify specific reasons why we can be partly or completely wrong when we make a causal inference. The threats to internal validity that will be plausible in a research context depend on the design that is chosen, the obtained pattern of results, and prior research and theory. Campbell and colleagues (e.g., Campbell & Stanley, 1966; Shadish et al., 2002) considered threats to internal validity in experimental, quasi-experimental, and pre-experimental designs. We focus here on designs comparing T and C groups in which both baseline and outcome measures are taken. These designs have been extensively considered from both Campbell's and Rubin's perspectives, providing a clear basis for comparison. They also provide a foundation for our later contrasts of the randomized experiment and the observational study,[3] designs that have been widely used in education, psychology, and public health.

Given the assumptions that (a) all participants experience identical experimental procedures except for the treatment condition and (b) all participants are assessed at both baseline and outcome, all of the simple, main effect threats to internal validity—such as history, maturation, and selection—are ruled out within Campbell's framework (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002). For a threat to be a problem, it must operate differentially in the T and C groups. Of central concern are four threats that involve *interactions* of another threat with the threat of selection, the possibility that participants in the T and C groups already differ at the beginning of the experiment.[4] We illustrate these threats in the context of an investigation of the effects of two teaching methods (computer assisted teaching vs. standard classroom instruction) on gains in mathematics achievement. All students complete a baseline measure at the beginning of

[2] Some researchers have interpreted internal validity more broadly. To highlight his original intended meaning of this term, Campbell (1986) once suggested that the term *internal validity* be replaced by "local, molar, causal validity" (p. 69).

[3] In statistics, following Cochran (1965, 1983; see also Rosenbaum, 2002), *observational studies* have been characterized as investigations in which causal inference is the goal, but the treatment cannot be imposed or randomized by the experimenter. In psychology, the term *passive observational study* has sometimes been used to characterize cross-sectional or longitudinal designs without an intervention in which the researcher aspires to make causal inferences (e.g., the concomitant time series design; Cook, Dintzer, & Mark, 1980).

[4] Of course, the implementation of identical procedures except for the treatment, the measurement of all participants at baseline and outcome, or both, may fail. We return to this issue in our discussion of randomized experiments.

the school year and an outcome measure at the end of the school year, and they experience identical experimental procedures other than the treatment.

1. *Selection* × *Maturation.* Students in the *T* and *C* groups may be naturally growing at different rates even in the absence of treatment. For example, a group of gifted students would be expected to show more rapid gains in achievement than a group of average students, even given identical teaching methods.

2. *Selection* × *History.* Participants in the *T* and *C* groups may experience different historical events (a.k.a., local history). Students in one group might be given smaller class sizes following a school district mandate; the other group in a second school district would not change in class size. The smaller class size might lead to increased gains in mathematics achievement even if there were no effect of teaching method.

3. *Selection* × *Instrumentation.* Participants in the *T* and *C* groups may be administered different outcome measures, or the outcome measures may have different measurement properties in the two groups (e.g., ceiling or floor effects; different factor structures; see Embretson, 2006). One group might be switched to an easier posttest, producing scores that would appear to be higher, even in the absence of an effect of the new curriculum.

4. *Selection* × *Statistical Regression*. One of the groups may be selected because it is extreme at baseline, whereas the other is not. In this case, the amount of regression to the mean may differ between the two groups and artifactually produce the apparent treatment effect. A decision may be made to give the computer-assisted teaching to a group of students who score low on the baseline measure and to give the standard curriculum to a group of average students. Particularly when either the internal consistency or the test–retest reliability of the measure is less than perfect, the low-scoring group may show more improvement than the control group because of regression to the mean, even in the absence of any effect of the teaching method. Campbell and Kenny (1999) have presented a full discussion of regression artifacts and their sources.

Below, we briefly consider some strategies for addressing each of these threats. Regardless of the strategy that is chosen, the plausibility of each threat also depends on the specific research context and the prior research and theory in that area. The typical threats to a design may not be plausible in a specific research area.

Addressing threats to internal validity: Design elements. Once the plausible threats to internal validity have been identified in the planning of the research, the strong priority within Campbell's tradition has been on identifying procedures that will minimize the likelihood (or extent) of their occurrence and on identifying specific elements that can be added to the design that would rule out the threat. As described in more detail below, randomization of participants to treatment conditions is the most general of these design elements, but more targeted design elements can be utilized to demonstrate that the pattern of obtained results is not consistent with the operation of the threat. To cite two examples, if maturation is a plausible threat, additional pretests can be added to estimate the maturational trend in each group prior to treatment. If history is a plausible threat, using a design in which multiple cohorts are given their baseline and outcome measures during nonoverlapping time periods helps address this threat. Because the time intervals over which each cohort is studied do not overlap, a common historical event (e.g., introduction of a smaller class size) can not be adduced that would lead to identical change in each of the cohorts. This strategy of adding targeted design elements that permit the pattern of results to rule out specific threats to validity is a key feature of Campbell's approach. We present more discussion of targeted design elements below in the context of our section on observational studies (see also Shadish & Cook, 1999; Shadish et al., 2002).

Campbell's approach also considers other less preferred approaches to reducing threats to internal validity. These approaches include careful measurement and statistical adjustment for the threat and appeals to the results of previous research and theory (Cook & Campbell, 1979; Higginbotham, West, & Forsyth, 1988; Shadish et al., 2002; West, Biesanz, & Pitts, 2000). Nonetheless, the priority of design-based approaches over statistical adjustment and other approaches in Campbell's tradition is clear: "When it comes to causal inference from quasi-experiments, design rules, not statistics" (Shadish & Cook, 1999, p. 300).

Rubin's Causal Model

Overview

Rubin's causal model (a.k.a., the potential outcomes model; Holland, 1986; Little & Rubin, 2000; Neyman, 1923/1990; Rubin, 1974, 1978, 2005, 2006b) brings the strengths of a formal mathematical/statistical perspective to causal inference. As a mathematical approach, it begins with a clear definition of the causal effect of interest, specifies the precise set of assumptions that are sufficient to make a causal inference for each research design, and uses a precise notational system that permits unambiguous specification of the parameters of interest. In addition to the mathematical precision of Rubin's causal model, much of the usefulness of the approach in applications lies in the heuristic value of *potential outcomes*, a core concept of the model. This concept has proven both to be intuitive to substantive researchers and to provide a remarkably generative way of thinking about how to obtain precise estimates of the magnitude of the desired causal effects for difficult research problems.

The Basic Model

To develop a precise definition of the causal effect of interest (the *causal estimand* or the magnitude of the difference in *Y* due solely to the treatment), we take as our starting point a single unit measured without error. We emphasize below psychology's typical unit, the single human participant, recognizing that the unit could also be an animal, group, community, and so forth. Treatment *T* is given to the participant, and the outcome variable *Y* is observed. Ideally, the comparison (control) treatment *C* is given to the *same participant at the same time and in the same context*, and the response is observed. Formally, each participant's causal effect, the individual treatment effect, is defined as $Y_T(u) - Y_C(u)$, where $Y_T(u)$ represents the response *Y* of unit *u* to treatment *T*, and $Y_C(u)$ represents the response of unit *u* to treatment *C*. Comparison of these two outcomes provides the ideal design for causal inference. The model can be easily extended to

more than two treatments, but we only consider the case of two treatments here for ease of presentation.

This ideal design provides a theoretically useful definition of the causal effect and a wonderfully heuristic way to think about problems of research design. Unfortunately, this design is a Platonic ideal that can never be achieved in practice. Consider again our example of the effect of computer-assisted teaching (*T*) as compared with standard classroom instruction (*C*) on mathematics achievement. We cannot give *T* and observe its outcome, put the same participant back to the identical time and place (thereby removing all traces of learning and returning the participant's motivation to its original state), and then give *C* and now observe its outcome. Consequently, we need to choose an approximation to the ideal design. Although Holland (1986) has discussed three approximations to the ideal design,[5] the most commonly used one in psychology is the randomized experiment that yields (given additional assumptions) an unbiased estimate of the average causal effect in the population, $\mu_T - \mu_C$, as is discussed in the section on randomized experiments below.

There are three important implications of this definition of a causal effect:

1. The model makes it clear that we are comparing two treatments, both possibly effective. ". . . [T]he effect of a cause is always relative to another cause" (Holland, 1986, p. 946). Although psychologists typically refer to a treatment group and a control group (as we do here), designating one of the groups as a "control group" presumes that it is totally benign with respect to its effect on the outcome, a *very* special case. Even placebos in well-designed pharmaceutical trials may lead to pharmaceutical or cognitive effects that influence the outcome of interest. These effects must be considered part of the so called "control" treatment.

2. The model requires a precise statement of the two treatments to be compared. Often, this statement will simply be the specific operationalization of the two treatments by the researcher. This specificity contrasts with the preference of many psychologists to describe treatments on the construct level, particularly in basic research (e.g., frustration vs. no frustration in classic work on aggression; see discussion in final section, also in Shadish et al., 2002; West et al., 2000). In other cases (as we see below), the two treatments can be compared only for the subset of people who could potentially receive either treatment. In still other cases (e.g., effects of Hurricane Katrina on health outcomes of Gulf coast residents), the researcher will need to conceptualize carefully the alternative treatment that the individual could *potentially* receive (i.e., compared with what?). Of importance for researchers in substantive areas of psychology interested in stable individual differences (e.g., abnormal psychology, personality), this definition makes it difficult to investigate the causal effects of individual difference variables because we must be able to at least conceptualize the individual difference (e.g., gender) as two alternative treatments. If we cannot do this, Rubin (1986) considers the problem ill defined. Much of the research in psychology modeling cross-sectional or longitudinal data in the absence of a treatment would fail to meet this criterion.

3. The model makes it clear that we can observe two sets of participants: (a) Group A given *T* and (b) Group B given *C*. A and B may be actual pre-existing groups (e.g., two communities) or they may be sets of participants who have selected or have been assigned to receive the *T* and *C* conditions, respectively. Of key importance, we also need to conceptualize the *potential* outcomes in two hypothetical groups: (c) Group A given *C* and (d) Group B given *T*. Imagine that we would like to compare the mean outcome of the two treatments. Statistically, in terms of the ideal design what we would ideally like to have is an estimate of either:

$$\mu_T(A) - \mu_C(A), \text{ or} \tag{1A}$$

$$\mu_T(B) - \mu_C(B), \tag{1B}$$

where *A* and *B* designate the group to which the treatment was given. Both Equations 1A and 1B represent average causal effects. Of importance, note that Equations 1A and 1B may *not* represent the same average causal effect; Groups A and B may represent different populations.[6] What we have in fact is the estimate of

$$\mu_T(A) - \mu_C(B). \tag{2}$$

These ideas are illustrated in Table 1. In the columns labeled "Potential outcomes," we represent the ideal design in which the responses of each participant are observed at the same time and setting in both *T* and *C*. In the columns labeled "Observed outcomes," we can only observe the outcomes for *T* in Group A and *C* in Group B. For observed outcomes, only half of the data we would ideally like to have can be observed; the other half of the data is missing. This insight allows us to conceptualize the potential outcomes as a missing data problem and focuses attention on the process of assignment of participants to groups as a key factor in understanding problems of selection.

Rubin's conceptual analysis makes it formally clear that we need additional assumptions if the comparison between the two groups

[5] Three approximations to the ideal design include the following: *1. Within-subjects design.* In the within-subjects design, the treatment and control conditions are both given to the same group of participants. The assumptions are (a) temporal stability, in which the same outcome will be observed regardless of when the treatment is applied (i.e., no maturation or history effects), and (b) causal transience, in which the administration of the first treatment will have no effect on the response the second treatment when it is administered later (no carryover effects). *2. Unit homogeneity.* With unit homogeneity, the units are created or selected to be identical in all relevant respects (e.g., identical ball bearings might be the units in an engineering experiment). Given unit homogeneity, the units are assumed to be exchangeable so that the responses of Units A and B to a treatment will be identical. *3. Randomization.* In randomization, each participant has an equal chance of being in treatment condition X_i. As we present in detail later, this procedure equates the participants in different treatment groups, on average, on all possible background variables in large samples. Although ideas from the first two approximations are sometimes used, the assumptions underlying these approximations to the ideal design will only rarely be fully tenable in psychological research contexts.

[6] Equations 1A and 1B will not generally be equivalent in the absence of randomization. There may be an interaction between treatment and baseline status, so that the magnitude of the causal effect may differ between the two groups. To illustrate, suppose Group A consists of college-bound high school graduates, and Group B consists of non-college-bound high school graduates, all of whom are eligible for college admission. The magnitude of the treatment effect (college education) on gains in achievement would almost certainly differ between the two groups.

Table 1
Potential and Observed Outcomes

	Potential outcomes		Observed outcomes	
Participant	*T*	*C*	*T*	*C*
1	10	10	10	■
2	11	13	■	13
3	11	11	■	11
4	12	16	■	16
5	12	12	12	■
6	12	15	12	■
7	12	13	■	13
8	13	15	13	■
9	13	17	■	17
10	14	18	14	■
	True average treatment effect: 2.0		Prima facie average treatment effect: 1.8	

Note. The columns labeled "Potential outcomes" illustrate the true responses of the 10 participants under the treatment (*T*) and control (*C*) conditions. The columns labeled "Observed outcomes" illustrate the responses of the same 10 participants in a randomized experiment or an observational study. A ■ indicates that the response was not observed. Half of the potential outcomes are not observed. The prima facie average treatment effect is the simple (possibly biased) difference between the observed means in the *T* and *C* groups.

represented by Equation 2 is to stand in for the ideal comparison represented by Equation 1A or Equation 1B. In general, Rubin's perspective prefers approaches that make the smallest number of necessary assumptions, make assumptions that are likely to be consistent with the actual context of the research problem, and make assumptions that are transparent—they are directly verifiable or their effects on the outcome can at least be probed. As Little and Rubin (2000) have noted, "Nothing is wrong with making assumptions; they are the strands that join the field of statistics to scientific disciplines. The quality of these assumptions, not their existence, is the issue" (p. 123).

Randomized Experiments: Intact and Broken

The randomized experiment is widely viewed by researchers as the "gold standard" of research designs. Proper randomization of participants to treatment conditions leads to the expectation that the assigned treatment condition will be statistically independent of any covariate at baseline in the population. Statistical independence means that the distribution of any measured or unmeasured covariate will, on average, be balanced (i.e., have the same distribution) at baseline in the *T* and *C* conditions. Two implications of statistical independence are that (a) the means of the *T* and *C* conditions will, on average, be equal on all possible measured and unmeasured covariates at baseline, and (b) the covariance of treatment assignment and any measured or unmeasured baseline covariate will, on average, be 0. These expectations imply that there are no differences on any potential covariate between the *T* and *C* groups at the beginning of the experiment in the population. These expectations provide a strong foundation for causal inference. Note these are expectations that hold exactly only in large samples or as the mean of a large number of random samples from the same population. In any single experiment, there is no guarantee that the mean pretest levels of the two treatment conditions will *not* differ on a specific covariate.

Campbell's Perspective

Given that the procedures other than the manipulation are identical and there is no attrition, there are four major threats that need to be ruled out in designs comparing *T* and *C* groups that include baseline and posttest outcome measurements on all participants. These are the four interactions with selection described above. Random assignment renders selection implausible on all possible variables, measured and unmeasured, thus simultaneously ruling out all of the interactions with selection. The generality of random assignment gives it a special status among all design elements:

> Random assignment facilitates causal inference in many ways—by equating groups before treatment begins, by making alternative explanations implausible, by creating error terms that are uncorrelated with treatment variables, and by allowing valid estimates of error terms . . . [R]andomization is the only design feature that accomplishes all of these goals at once, and it does so more reliably and with better known properties than any alternatives. (Shadish et al., 2002, p. 252)

Other threats to internal validity. In the most recent major statement of Campbell's perspective, Shadish et al. (2002) have identified two additional threats to internal validity that are *not* ruled out by randomization. These are violations of the two assumptions noted on p. 19.

1. *Nonconstant features of experiments.* In some research situations, other features of the research protocol—such as the setting, the instructions, the delay between baseline and outcome measures, or the nature of the testing—may not be kept constant. For example, control participants may be given a shorter form of the outcome measure or not be given the baseline measure because they are expected to be less cooperative than participants given a desired treatment program.

2. *Treatment-related attrition.* Attrition refers to any loss of participants from posttest measurement who were *initially* randomized whether because of refusal to answer questionnaire items, failure to show up for the posttest measurement session, moving away from the study site, or the like. Particular concerns with respect to internal validity arise when attrition is related to treatment assignment. For example, in our earlier example of mathematics achievement, students with low math ability may drop out of the innovative new program, whereas students with high math ability may drop out of the standard program.

Addressing remaining threats. Campbell's perspective emphasizes the prevention of threats to internal validity rather than their correction. For example, Shadish et al. (2002, chapter 10) have devoted 10 pages to methods of preventing attrition, such as mini-

mizing any undesirable features of the treatment conditions and the measurement process, monitoring attrition throughout the experiment, and instituting procedures for retention and tracking of participants. In contrast, only half this space is devoted to statistical analyses that describe and attempt to correct for attrition once it occurs. Shadish et al. are clearly willing to incorporate statistical corrections, including corrections from Rubin's perspective, but that is not their emphasis. Similarly, Shadish et al. encourage researchers to prevent potential problems by reviewing the proposed experimental and measurement protocol for the possibility of any nonconstant features and monitoring adherence to this protocol throughout the experiment to detect changes that may creep in.

Rubin's Perspective

Additional assumptions. Even given randomization, several assumptions are needed in the context of Rubin's causal model to provide a sufficient basis for an unbiased estimate of the magnitude of the causal effect. The first three assumptions below are normally met in the typical 1-hr laboratory experiment in psychology in which each participant separately receives a highly standardized treatment independently from other participants. However, these assumptions can fail in important applied experiments in which the treatment is highly valued by participants and in both basic and applied research in which the treatment, measurement, or both are extended in time (West et al., 2000, 2008).

1. *Full treatment adherence.* All randomized participants complete the full treatment protocol (no treatment nonadherence, a.k.a., treatment noncompliance). Participants must not be able to drop out of treatment to which they were assigned or reassign themselves to another treatment condition. As an example of the failure of this assumption, in randomized trials of the effectiveness of screening mammography in providing early detection of breast cancer in women over 50 years of age, Baker (1998) has noted that about one third of the women assigned to the mammography condition did not receive screening and that many women in the no screening control condition obtained screening outside the trial.

2. *No attrition from posttest measurement.* All randomized participants complete the posttest (outcome) measurement. Commonly observed failures of this assumption include participants refusing to answer questions or to be observed, dropping out of the research prior to the post test measurement, or moving to another location in longitudinal research. In studies of aging or serious diseases, actual participant mortality may be a significant source of attrition. Given full treatment adherence, only if attrition is differential in the *T* and *C* conditions is it theoretically a problem. Under the assumption that all variables related to both treatment assignment and attrition status (1 = complete; 0 = attrited) are included in the data set, it is possible to obtain an unbiased estimate of the causal effect using methods that adjust for missing data (Little & Rubin, 2002; discussed below). Whether this assumption is in fact met in practice is unknown, except for special planned missingness designs (see Graham, Taylor, Olchowski, & Cumsille, 2006).

3. *Stable-unit-treatment value assumption (SUTVA).* According to Rubin (1986),

> SUTVA is the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive. (p. 961)

The purpose of the SUTVA assumption is to guarantee that a *single* value of the response will be observed for the participant under *T* and a *single* value of the response (possibly [hopefully] different) will be observed[7] under *C*. As one example, participants in an HIV prevention experiment may have a lowered risk of contracting HIV if their sex partners are also individuals in the (effective) treatment condition rather than participants in the (ineffective) control condition. Another instance is our earlier educational research example, in which some participants in the standard classroom instruction condition may be exposed to computer-based instructional material by friends in classrooms receiving the computer assisted instruction. Violations of SUTVA can also occur if there are hidden variants of an ostensibly well-defined treatment condition (e.g., different variants of the treatment are implemented at different sites in a multisite experiment) and the participants have different responses to the treatment variants. Replications of the experiment become problematic because different mixtures of the treatment variants or different amounts of participant contact can be expected to produce different estimates of the magnitude of the average causal effect, even in the absence of any sampling error. Rosenbaum (2007) and Sobel (2006) present extensive discussions of interference between units.

4. *Constant treatment effect (optional).* Given randomization and that the above three assumptions have been met, Rubin's causal model provides an unbiased estimate of the *average* causal effect. Randomization yields an unbiased estimate of the average causal effect, $\mu_T - \mu_C$, an effect that characterizes the population of participants that could be randomized. The average causal effect has been strongly emphasized in Rubin's work. However, this causal effect cannot be particularized to any single participant, $Y_T(u) - Y_C(u)$, the individual causal effect if the idealized experiment could be performed on the same participant *u* at the same time and in the same context. If the estimate of the average causal effect in a randomized experiment were a 10-point decrease in depression, one possible outcome would be that half of the participants do not benefit from the treatment, half of the participants have a 20-point decrease, and no participants would show the 10-point decrease implied by the average causal effect. To achieve an unbiased estimate of *individual* causal effects in the context of a (between-subjects) randomized experiment, a further necessary assumption must be made that the treatment effect is constant so that a single value characterizes all participants. Achieving an unbiased estimate of the average causal effect is typically sufficient in areas such as experimental and social psychology; however, unbiased estimates of individual participant change may be the goal in many clinical and health contexts.

In summary, one of the strengths of Rubin's causal model is that it makes transparent the assumptions necessary to make a causal inference. The randomized experiment makes the fewest assump-

[7] Other authors (e.g., Neyman, 1923/1990; Steyer, 2005; Steyer, Partchev, Kröhne, Nagengast, & Fiege, 2009) have alternatively assumed that there is a single "true" response. As in classical test theory, random measurement error is always present. Theoretically, the true score is the mean of a very large number of observed responses by the same individual under identical conditions.

tions, notably (1) treatment adherence, (2) no attrition, and (3) SUTVA. Violations of these assumptions are potentially observable and can be monitored by researchers. If these assumptions are met, an unbiased estimate of the average causal effect can be computed. If inferences are to be made about individual rather than average causal effects of a treatment, the fourth assumption (that the causal effect is constant across participants) must be added.

What if the assumptions are not met? Addressing violations. One of the strengths of the potential outcomes idea within Rubin's causal model is that it permits researchers to conceptualize appropriate comparisons when the assumptions of the randomized experiment are not met, sometimes termed the *broken randomized experiment* (Barnard, Du, Hill, & Rubin, 1998; Barnard, Frangakis, Hill, & Rubin, 2003). Considerable work has been done within this framework to identify approaches and sets of additional assumptions that are sufficient to yield unbiased estimation of causal effects in broken randomized experiments and in other designs. Here, we provide a brief introduction to an approach to the problems of binary (all or none) treatment nonadherence and attrition for the special case of a randomized experiment with a *T* group and a *C* group.

Treatment nonadherence. Angrist, Imbens, and Rubin (1996) considered the special case in which participants could choose to take or not take the randomly assigned treatment.[8] Traditionally, such situations have been analyzed by intent to treat (ITT) analysis in which participants' responses have been analyzed following Sir Ronald Fisher's maxim of "analyze them as you've randomized them" (as cited in Boruch, 1977, p. 199). This strategy lumps together participants in the *T* group who did and who did not receive treatment, estimating the causal effect of *treatment assignment* rather than treatment per se. Even for the full sample of participants, the estimate of the effect of treatment assignment in ITT analysis may also be biased if there are missing data in addition to treatment nonadherence (Frangakis & Rubin, 1999; Hirano, Imbens, Rubin, & Zhou, 2000).

Applying the idea of potential outcomes, Angrist et al. (1996) identified four theoretically possible subgroups of participants with respect to treatment adherence. Table 2 illustrates the four possible subgroups of participants. The actual treatment received is listed in the body of the table.

Adherers take the treatment to which they are assigned.[9] Always takers will always receive treatment regardless of the treatment group to which they are assigned (e.g., women who would undergo mammography screening if assigned to the *T* group but who would also pay to get mammography screening if assigned to the *C* group). Never takers would always refuse the treatment (women who decline mammography screening regardless of the treatment condition to which they are assigned). Defiers, a theoretically possible group that is rare in most research contexts, would take the treatment if assigned to the *C* group and would take the control if assigned to the *T* group. Causal effects of treatment cannot be defined for the never taker and always taker groups because they are only willing to accept one of the treatment conditions—there is no alternative potential outcome. Sheiner and Rubin (1995) clearly note that from the potential outcomes perspective, the comparison of central interest is between an adherer assigned to the treatment condition and an adherer assigned to the control condition, termed the *local average treatment effect* (LATE; Angrist et al., 1996; a.k.a., the complier average causal effect; Little & Yau, 1998).

The estimation of the LATE begins with the assumptions of the randomized experiment—random assignment, no attrition from posttreatment, and SUTVA—but relaxes the assumption of no treatment adherence. However, additional assumptions are now needed to produce an unbiased estimate of the causal effect.

1. *Monotoncity.* There are no defiers. The theoretical existence of defiers keeps us from being able to statistically identify the model so that it can be estimated.

2. *Exclusion restriction.* Randomization must operate only through the treatment and not have other direct effects on the outcome. This assumption eliminates any possibility that the effects of treatment *assignment* on the always takers or never takers would affect the outcome. Hirano et al. (2000) have provided an illustration of the violation of this assumption in a randomized experiment in which physicians were encouraged (computer generated reminder) or not encouraged to give at risk patients inoculations for influenza. The physicians in the *T* group apparently gave the always takers, predominately the highest risk patients (e.g., chronic obstructive pulmonary disease), special treatment—such as scheduling their inoculations early—thereby decreasing their potential exposure—or providing advice about other methods about avoiding potential exposure to the virus. In this case, there would be a direct path from treatment *assignment* to the outcome in addition to the path through the treatment.[10] This effect would lead to bias in the estimate of the causal effect of the immunization against influenza. This is the key assumption underlying the LATE estimates, and it requires careful scrutiny (e.g., Hirano et al., 2000). Masking of the treatment condition to both the participant and treatment provider, assuming that it can be successfully achieved, provides the strongest assurance that the exclusion criterion is met in randomized experiments with noncompliance.

3. *Nonzero average causal effect on compliance behavior.* Mathematically, there must be at least some compliers to be able to compute an effect.

These three additional assumptions are sufficient to yield an unbiased estimate of LATE. For the special (but common) situation in which there are also no always takers (e.g., the treatment only exists in the context of the randomized experiment), the LATE estimate can be simply computed as

$$\hat{LATE} = \frac{(\bar{Y}_T - \bar{Y}_C)}{\hat{\pi}_T},$$

where $\hat{\pi}_T$ is the estimate of the proportion of compliers in the treatment condition, $\bar{Y}_T$ is the mean of the *T* group, and $\bar{Y}_C$ is the mean of the *C* group from the ITT analysis (i.e., including all participants as randomized). The above expression makes clear the

[8] Angrist et al.'s (1996) approach requires that participants who are assigned to the treatment condition, but who do not take the treatment, receive the identical treatment to participants in the control condition.

[9] We have followed current psychological terminology in referring to individuals who complete the assigned treatment as adherers. These individuals are often termed compliers in the statistical literature.

[10] From Campbell's perspective, this effect primarily represents a problem of the construct validity of the independent variable, properly describing the constructs represented by the full treatment package. The present illustration is a complicated special case because the magnitude of the effect is not uniform across different subgroups within the treatment group.

Table 2
Four Possible Subgroups: Types of Treatment Adherence and Nonadherence

Type of participant	Actual treatment received when assigned to the treatment group	Actual treatment received when assigned to the control group
Adherer (complier)	Treatment	Control
Always taker	Treatment	Treatment
Never taker	Control	Control
Defier	Control	Treatment

important role of randomization: The proportion of compliers is expected to be, on average, the same in the *T* and *C* groups. Conceptually, the never takers, $(1 - \hat{\pi}_T)$ of the participants, receive the identical *C* treatment regardless of their treatment assignment. For this subgroup, the effect size of treatment assignment is 0.

The derivation of the standard error of the LATE estimate is more challenging (see Little & Yau, 1998, for a derivation), but LATE effects can now be tested for statistical significance using standard statistical packages (e.g., SAS: Little & Yau, 1998; Mplus: Jo, 2002). Although the LATE estimate is the estimand that many researchers would like to have, researchers are often surprised by the lack of efficiency and, hence, low statistical power of tests of significance of LATE estimates. This lack of efficiency is realistic, as it reflects the uncertainty associated with which participants would be in the hypothetical group of control group members who would comply if offered treatment—information about participants' adherence status in the *C* group is missing. Typically, efficiency can be improved in randomized experiments by including covariates that predict the outcome and reduce error variance. For LATE estimates, efficiency can also be potentially improved, sometimes dramatically, by including other covariates that reduce uncertainly associated with membership in the complier group (Jo, 2002).

Attrition. Consider now the case in which proper randomization has occurred and the assumptions of SUTVA and full treatment adherence are met, but the assumption of no attrition from posttest measurement is relaxed. Now we have another form of missing data problem, here some or all of the dependent measures for some of the participants may be missing. Of most concern with respect to causal inference is the issue of treatment-related attrition in which the rate of attrition or the characteristics of participants who attrit differ in the *T* and *C* groups.

Rubin (1976; Little & Rubin, 2002; see also Schafer & Graham, 2002) has been at the forefront of the development of modern missing data theory. He proposed that there are three types of missing data. Data missing completely at random imply that the probability that the data are missing is unrelated to any measured or unmeasured variables that are related to the outcome. The estimate of the causal effect is unbiased. Data missing at random imply that all systematic sources of missingness are contained in measured study variables. Typically, the estimate of the causal effect must be adjusted to account for those measured variables that are associated with both missingness and the outcome. Finally, data may be missing not at random when missingness is related to the participant's level on the unobserved variables. For example, participants in a school-based alcohol prevention program may be more likely to miss an in-class measurement following a prior evening of binge drinking. Here, adjustment is needed, but the exact nature of the adjustment cannot be fully known.

Given two additional assumptions, proper estimates of the causal effect can be achieved in randomized experiments.

1. *Data are missing at random.* Data missing at random imply that all systematic sources of missingness are contained in measured study variables. The estimate of the causal effect must be adjusted to account for any differences in the measured variables that are associated with both missingness and the outcome. This assumption is more likely to be approximated if the researcher includes comprehensive baseline measurement of all variables believed to be related to both attrition and the outcome. Whether the researcher has, in fact, met the assumption cannot be verified.

2. *Distributional assumptions regarding measured variables.* Standard modern approaches to missing data with continuous variables, including full information maximum likelihood and multiple imputation, assumed that the measured variables had a multivariate normal distribution. Commonly used statistical software for missing data (e.g., SAS PROC MI, NORM, Mplus) rely on this assumption in computation. Fortunately, the standard approaches do not appear to be sensitive to small to moderate violations of this assumption, and original data may be transformed so that the assumption is more closely approximated. New robust methods for addressing other forms of missing data are being developed (Little & Rubin, 2002; Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; Schafer, 1997).

Given that the assumptions are met, full information maximum likelihood and multiple imputation provide unbiased estimates of the causal effect. The key assumption here is that data are missing at random. This assumption cannot be verified. Researchers are encouraged to conduct sensitivity analyses to probe the amount of change that could potentially occur given that an *unmeasured* variable is associated with both attrition and the outcome variable. Sensitivity analyses help bracket the magnitude of the causal effect, reflecting the actual uncertainty that is associated with its estimation.

Comparing the Two Approaches: Randomized Experiments

In the context of the randomized experiment, the two approaches are largely in agreement. Both perspectives agree that randomization potentially permits an unbiased estimate of the direction and magnitude of causal effect. Campbell's perspective emphasizes practical methods of preventing the remaining threats to internal validity, whereas Rubin's perspective focuses on developing the additional assumptions sufficient to make proper analytic adjustments for problems, so that the precise magnitude of the causal effect of interest can be estimated. Campbell's perspective, as reflected in Shadish et al. (2002), largely endorses Rubin's statistical adjustment procedures,

and the writings of Rubin contain no admonitions about the use of Campbell's practical methods of preventing threats to internal validity. The two perspectives largely provide complementary areas of emphasis.

Nonetheless, we wish to note two potential differences between Campbell's and Rubin's perspectives, the greater emphasis on treatment adherence and the SUTVA assumption in Rubin's relative to Campbell's perspective. Working within the potential outcomes framework, Angrist et al. (1996) strongly emphasized the LATE estimate, the comparison of the outcome of a participant in the *T* condition who actually took the treatment to the outcome of a participant in the *C* group who would have taken the treatment if given the opportunity to do so. Because no effect of treatment on the outcome can be defined for always or never takers, these individuals are excluded from the estimation of the LATE causal effect. This approach leads to the estimate of a clearly defined treatment effect within the potential outcomes framework. In contrast, Campbell's approach does not formally list treatment adherence as a threat to internal validity, but rather Shadish et al. (2002) have identified it as a problem with treatment implementation that may interfere with getting "a good estimate of a treatment effect" (p. 315). Shadish et al. appear to be more satisfied than Rubin with the results of an ITT analysis that provides an unbiased estimate of the effect of treatment *assignment* on the outcome when there is full treatment adherence. However, Frangakis and Rubin (1999) and Hirano et al. (2000) showed that the ITT estimate can be biased if both nonadherence and attrition occur, and West and Sagarin (2000) raised concerns about replicating the treatment adherence process in different investigations.[11]

Second, Rubin has emphasized the importance of SUTVA as a sufficient condition to produce unbiased estimates of the magnitude of the causal effect. Cook and Campbell (1979) originally treated issues related to interference between subjects (e.g., resentful demoralization), but Shadish et al. (2002, p. 72) now consider these issues to be threats to construct validity. Shadish et al. have not discussed reasons for this change. Threats to construct validity do not impugn causal inference.

Presumably, from Shadish et al.'s (2002) perspective, any sources of interference are added to the package that is now considered to be the treatment. From Rubin's perspective, this decision potentially muddies the clear definition of the treatment and the potential outcomes. Although rare exceptions can be cited, problems of interference can be expected to typically affect the magnitude, but not the direction of the treatment effect in randomized experiments with complete data and full treatment adherence. Sobel (2006) analytically considers more complex real world experiments with treatment nonadherence and attrition, identifying conditions under which even the inference of the direction of the estimate of the causal effect may be wrong.

The differences between Campbell's and Rubin's traditions with respect to full treatment adherence and SUTVA appear to be related to two differences in emphasis between the theories. First, Rubin relies on precise operations to define treatments, whereas for Campbell, the operation is only the exemplar of the treatment-related construct used in the particular experiment. Second, Rubin relies on the precision of the potential outcomes framework, whereas Campbell's tradition appears to use counterfactuals in a more general way. These differences appear to reflect the emphasis within Rubin's framework of defining conditions that are sufficient to obtain precise estimates of the *magnitude* of the causal effect. Historically, Campbell's framework has appeared to be satisfied with clear conclusions about the *direction* of the causal effect.

Observational Studies

Following Cochran (1965), we characterize observational studies as investigations in which causal inference is the goal, but the treatment cannot be imposed or randomized by the experimenter. We limit our consideration to designs with a *T* group and a *C* group in which baseline covariates, often including a pretest on the outcome variable, have been measured. This design was originally termed the *nonequivalent control group design* by Campbell and Stanley (1966). Such designs have often been used to study important basic and applied issues, such as the effects of sudden death of a spouse on long-term mental health of the surviving spouse (Lehman, Wortman, & Williams, 1987), the effects of secondhand cigarette smoking on cardiovascular problems (Barnoya & Glantz, 2005), and the effects of retention in grade on school children's subsequent achievement and psychosocial development (Jimerson, 2001; Wu, West, & Hughes, 2008a). Observational studies inherit all of the issues identified by Campbell's and Rubin's traditions for the randomized experiment. The central additional problem in the observational study is that the rule by which people are assigned to treatment and control conditions is unknown and must be presumed to be nonrandom. As a result, the participants in the *T* and *C* groups may differ at baseline on measured and unmeasured covariates. These covariates provide a potential alternative explanation (selection-related confounding) for any observed treatment effect.

Campbell's Approach

Overview. Campbell's approach focuses on the four prominent threats to internal validity described earlier—Selection × Maturation, Selection × History, Selection × Instrumentation, and Selection × Statistical Regression. The threats that will be of most concern depend on the specific research context. For example, Selection × Maturation would be an important concern in the evaluation of a new school program to improve reading in young children in light of children's natural growth in reading skills in the absence of treatment. Selection × Maturation would be of far less concern in the evaluation of a program to decrease adult stuttering, as stuttering behavior in adults is known to be highly stable in the absence of intervention. The task of the researcher is to bring current scientific knowledge to bear to identify the most important threats to internal validity and to add targeted design elements to address those specific threats.

Shadish and Cook (1999), Shadish et al. (2002), and Rosenbaum (1999) have presented extensive lists of targeted design elements that can potentially be used to address specific threats to validity.

[11] The ITT estimate is a weighted combination of the LATE estimate for the compliers and a 0 effect for the never takers. For this estimate to replicate, the proportion of individuals who comply must not change, and the nature of individuals who comply must not change if there is a nonconstant treatment effect.

Three example design elements are given below followed by an illustration in which they were used to strengthen causal inference.

1. *Matching.* If participants in the *T* and *C* groups can be successfully matched on all important covariates that relate to both treatment assignment and outcome, then selection bias can be eliminated (see next section for more extensive discussion). This strategy will often, but not always, reduce the viability of threats related to interactions with selection.

2. *Nonequivalent dependent variable.* A nonequivalent dependent variable is defined as one that would be expected to be affected by the same threats as the outcome variable but not by the treatment. If the outcome variable shows an effect of treatment but the nonequivalent dependent variable does not, confidence in the causal interpretation of the treatment effect increases. In our earlier example, if computer-assisted teaching led to gains in achievement in mathematics, but not in other subjects, relative to the standard classroom instruction treatment, this pattern of results would help rule out many forms of the Selection × History threat.

3. *Repeated pretests over time.* If pretests can be taken at several time points before the intervention, the pattern of maturation can be estimated in the *T* and *C* groups prior to treatment and possibly extrapolated to the posttreatment period, addressing the Selection × Maturation threat.

Illustration. Reynolds and West (1987) conducted an evaluation of a program to increase sales of state lottery tickets by convenience stores (see Figure 1). The store managers refused to be randomly assigned to program and control conditions. In the observational study, 44 stores agreed to participate in the program. Applying the first design element above, 44 program stores were matched 1:1 with 44 control stores from the same chain in the same zip code (geographical location) on the basis of previous lottery ticket sales. Increases in lottery ticket sales occurred in the program but not the control stores (see Panel A). Applying Design Element 2, substantial increases were observed within the program stores on lottery ticket sales but not on other product categories (see Panel B). Applying Design Element 3, there were no differences in sales between program and control stores during the 4 weeks before the implementation of the program, but the program stores consistently sold more lottery tickets during the 4 weeks after the program (see Panel C).

Summary. In summary, within Campbell's approach, a variety of additional targeted design elements are added to the basic design to try to rule out threats that arise in the specific research context, making it difficult to identify any potential threats to internal validity (confounding factors) that might be responsible for the observed pattern of results. The goal is to rule out all potential threats to internal validity, ideally through the use of design elements. Nonetheless, each of the design elements is subject to criticism. For example, the nonequivalent dependent variables could potentially be less reliable or less subject to confounding influences than the primary outcome—or the population that buys lottery tickets may be different than the population that buys other products at convenience stores. The strength of the present design illustration is that the results based on multiple design elements are coherent, leading to a strong causal inference of a directional effect: The sales campaign has led to an increase in lottery ticket sales.

Rubin's Perspective

Matching and its justification. Throughout his writings (see especially Rubin, 2006a), Rubin has emphasized matching[12] as the strongly preferred approach to the observational study. Conceptually, to the extent that a control participant can be identified that is identical to the treated participant at baseline (i.e., they are exchangeable), an unbiased estimate of the individual causal effect can be obtained (see Footnote 5 above describing Holland's, 1986, Approximation 2). However, in practice, when we estimate the average causal effect comparing a *T* group with a *C* group, we cannot be sure that the two groups do not differ in the population, perhaps substantially, on key variables at baseline. As noted above, observational studies have unknown rules for assignment of participants to *T* and *C* groups. Unknown biases in the selection of participants may produce baseline differences between the participants in the *T* and *C* groups on measured or, potentially more troublesome, unmeasured background variables. From Equation 2 above, $\bar{Y}_T(A) - \bar{Y}_C(B)$ is our (naïve) prima facie estimate of the causal effect. From the perspective of the potential outcomes model, what we would like to estimate is a weighted combination of Equations 1A and 1B, $\pi\,[\mu_T(A) - \mu_C(A)] + (1 - \pi)[\mu_T(B) - \mu_C(B)]$, where *A* represents the group receiving the treatment, *B* represents the group receiving the control, and π is the proportion of the population that is in the treatment group. Given that estimates of the potential outcomes cannot be observed and that the true causal effect of the treatment relative to the control may differ when applied to Groups A and B (see Footnote 6), the prima facie causal effect may be biased. Otherwise stated, to the extent that some of these baseline variables are also related to the outcome, the results are confounded. All or part of the observed "treatment effect" might be due to differences between the groups on background variables rather than the treatment.

In special cases (Rubin, 1977), a single *measured* key covariate (*COV*) may be known to be fully responsible for treatment assignment. If we can exactly match each participant in the *T* group with a participant in the *C* group on the baseline covariate, then $Y_{T_i}|COV_i - Y_{C_i}|COV_i$ represents an unbiased estimate of the individual causal effect. More generally, if *COV* has an identical distribution in the *T* and *C* groups, it cannot be the cause of the observed treatment effect. Achieving this balance of *COV* across the *T* and *C* groups forces it to be orthogonal to (not associated with) treatment assignment. This procedure reproduces the balance achieved on the specific covariate through randomization in large samples.

[12] We emphasize matched pairs here because it is conceptually the simplest procedure. Rubin (2006a) has also extensively discussed creating homogeneous strata and parametric and nonparametric adjustments for covariates (e.g., analysis of covariance). Ming and Rosenbaum (2000) have argued for variable many to one matching, particularly because of its greater statistical power. Matching and stratification are typically preferred to analysis of covariance because they do not need to specify the functional form of the relationship between the matching variables and the outcome, and they reduce extrapolation of treatment effects beyond the data. The key element of each of these procedures is the estimation of the difference between the treatment and control means, conditioned on the probability that each participant would be in the treatment group.

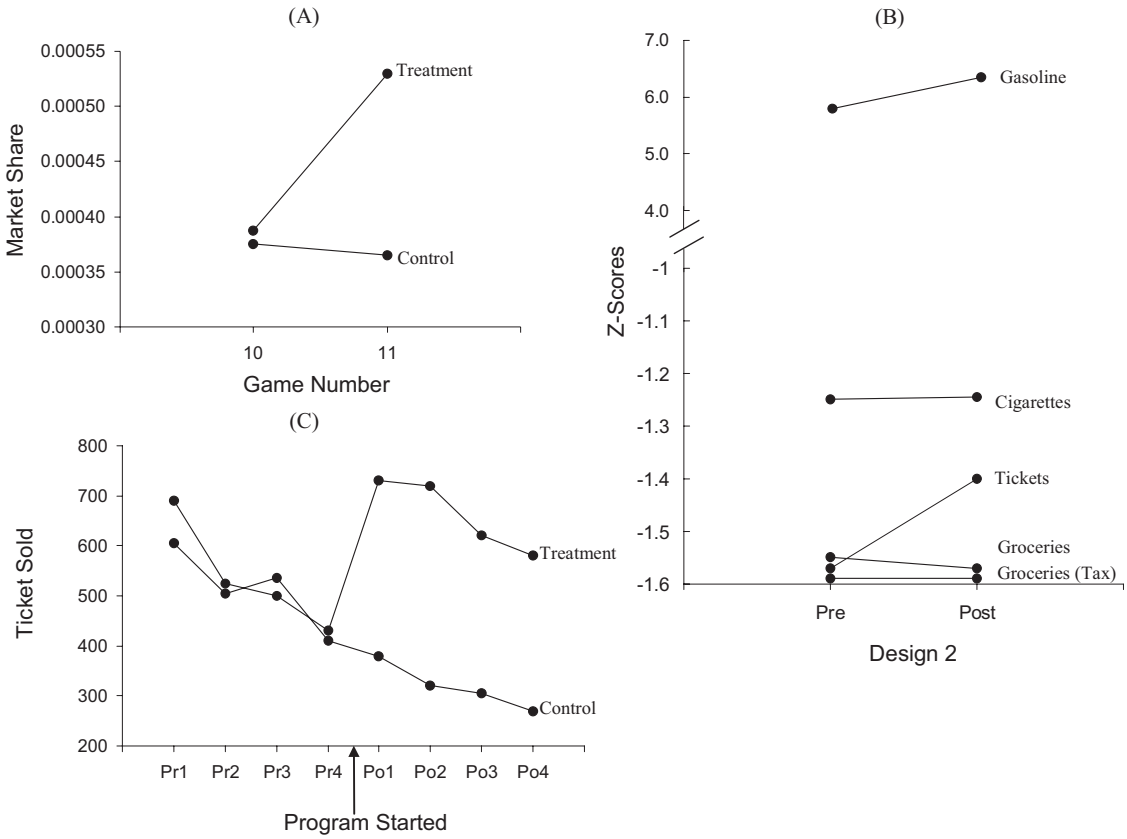

Figure 1. Adding design elements to strengthen causal inferences in observational studies. (A) *Matching*. Treatment and control stores are selected from the same chain, are in the same geographical location, and are comparable in sales during baseline (Lottery Game 10). Introduction of the treatment at the beginning of Lottery Game 11 yields an increase in sales only in the treatment stores. (B) *Nonequivalent dependent variables*. Within the treatment stores, sales of lottery tickets increase substantially following the introduction of treatment. Sales of other major categories (gasoline, cigarettes, groceries [nontaxable], and groceries [taxable]) that would be expected to be affected by confounding factors, but not treatment, do not show appreciable change. (C) *Repeated pre- and posttest measurements*. Treatment and control stores sales show comparable trends in sales during the 4 weeks prior to and following the introduction of the treatment. The level of sales in the treatment and control scores is similar prior to the introduction of treatment, but it differs substantially beginning immediately after treatment is introduced. (Adapted from "A Multiplist Strategy for Strengthening Nonequivalent Control Group Designs," K. D. Reynolds & S. G. West, 1987, *Evaluation Review*, *11*, pp. 691–714.)

Historically, a major impediment to successful matching was the need to match on multiple covariates. Even in large samples, it was difficult to identify good matches when multiple variables were involved. The development of propensity score methods (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984) provided elegant statistical theory that extended the benefits of matching from a few to many covariates. We initially provide a conceptual presentation of propensity scores followed by a research example.

A propensity score represents the probability that a given subject will be assigned to the T group, $0 < P(T) < 1$. The propensity score is estimated on the basis of the covariates measured at baseline, for example through logistic regression. Rosenbaum and Rubin (1983) have shown that if matching the T and C groups in fact achieves balance on the propensity score, then the T and C groups are expected to be balanced on *any measured covariate* from which the propensity score is constructed. Otherwise stated, if all important covariates related to both treatment assignment and outcome have been measured and all propensity scores fall within the bounds of $0 < P(T) < 1$, then it is possible to achieve an unbiased estimate of the causal effect of T. The difference between the means of the T and C groups matched on the true propensity scores will represent the average causal effect, $E[(\bar{Y}_T - \bar{Y}_C)|P(T)]$. However, in practice in studies with small to moderate sample sizes, there are several challenges to achieving the ideal results derived from statistical theory.

First and most important, it is assumed that all important confounders that are related to both treatment assignment and the outcome have been assessed. Any omitted confounders potentially lead to biased estimates of the causal effect. Given that the basis for selection into the T and C groups is unknown, this condition can *at best* only be approximated. This approximation will be better when subject-matter experts identify the critical covariates

and an extensive battery of reliably measured covariates is used. Such procedures typically provide substantial reduction in bias in the estimate of the causal effect relative to unadjusted estimates. In contrast, observational studies in which a few convenient demographic variables are used to construct the propensity score are unlikely to provide substantial bias reduction.

Second, it is important that a matched pair be identified for the maximum possible number of participants in the smaller group (typically the *T* group). If the distributions of the propensity scores only partially overlap in the two groups, it will often be impossible to find adequate matches. This overlap of distributions is referred to as the common support region. A small common support region is problematic for two reasons. (1) If matches are formed that are outside the support region, imbalances on the covariates will exist even after matching so that the estimates of the causal effects may be biased. (2) If participants are excluded who are not successfully matched, *an* average causal effect may be estimated. However, exclusion of participants may highlight limitations of the data: Generalization of the average causal effect is only possible to the range of propensity scores for which adequate matches can be achieved. Often only the average causal effect corresponding to a subgroup of the treated (or control) participants may be computed. For example, Haviland, Nagin, and Rosenbaum (2007) studied the effect of joining versus not joining a gang at 14 years of age on subsequent violent delinquent acts. The group of boys with the highest level of consistent violence prior to 14 years of age all joined gangs; there were no comparable nonjoiner boys (controls) with whom they could be matched. For participants for whom the propensity score equals 0 or 1, no potential outcome is defined. Therefore, for the most violent boys prior to 14 years of age who may be of most interest to the researchers, no causal effect of joining a gang can be defined or estimated. Matching on propensity scores appropriately highlights the limits on generalization of the causal effect, a feature that does not characterize traditional approaches, such as multivariate analysis of covariance. Little, An, Johanns, and Giordani (2000) and Stuart and Rubin (2008) have presented more detailed discussions of matching and analysis of covariance as adjustment techniques.

Third, the propensity score has to be correctly estimated and account for the relationship between the covariates and treatment selection in the sample. There are three issues here. (1) As already discussed, the central issue is the omission of important covariates associated with both treatment assignment and the outcome (termed hidden bias by Rosenbaum, 2002). (2) Some covariates may not be reliably measured, potentially producing bias in the estimation of the propensity scores. (3) There is the possibility that the function relating the covariates to treatment assignment is not correct. Traditionally, logistic regression that assumes that there is a linear relationship between the covariates, and the logit of the treatment assignment indicator is used to estimate propensity scores. Rosenbaum and Rubin (1984) and Dehejia and Wahba (1999) also included selected quadratic effects of covariates and interactions between pairs of covariates. McCaffrey, Ridgeway, and Morral (2004) used a nonparametric approach based on regression trees. The goal of all these approaches is to achieve balance between the treatment and control participants on the propensity score as well as between each of the separate baseline covariates. This balance can be checked, and the propensity score model can be re-estimated (e.g., by adding interactions), and/or supplemental adjustments can be added for important covariates, particularly those for which good balance has not been achieved (e.g., Rubin & Thomas, 2000). Nonetheless, some degree of uncertainty will remain in any real data set, especially with the modest sample sizes typically used in psychology, about how well the *T* and *C* groups have been equated on measured and unmeasured baseline covariates.

Illustration. Wu, West, and Hughes (2008a, 2008b) sought to investigate the effects of retention in first grade on children's trajectories of achievement in math and reading in later grades. In an attempt to identify children at risk of retention, 784 children who scored below school district medians on reading readiness at entry to first grade were identified as potential participants. On the basis of consultation with experts in school retention, 72 covariates (demographic, achievement, and psychosocial variables) were measured in first grade (prior to the retention decision) that might possibly be related to the retention decision, achievement outcomes, or ideally both. On the basis of the complete sample, logistic regression was used to compute a propensity score (the probability of retention on the basis of baseline covariates) for each child. Retained children were matched 1:1 with a promoted child using optimal matching with a defined caliper width of .2 standard deviation to achieve high-quality matches (for a discussion of computer algorithms that attempt to optimally match groups, see Rosenbaum, 2002, section 10.4). A total of 97 pairs of children could be closely matched on their propensity scores. Figure 2 displays the distributions of the propensity scores prior to matching (see Panel A) and following matching (see Panel B). Of note, following matching, the distributions of the retained and promoted children were very similar and ranged from .003 to .918, representing nearly the full potential 0–1 range of the propensity scores.

Checks on the balance of each of the individual covariates were performed using *t*-tests for continuous and chi-square tests for categorical outcomes. Of the 72 covariates, only one achieved statistical significance at $\alpha = .05$. The standardized effect size for this largest difference was $d = 0.33$, slightly larger than small according to Cohen's norms. These results suggest that the estimation of the propensity scores was probably adequate across the two groups.

We then estimated the effect of retention for each quintile for the propensity score matched pairs of children. These estimates were then weighted by the number of cases in the original sample to provide an estimate of the treatment effect that approximates the magnitude of the effect in at risk population that was studied. These estimates are "same age" comparisons that focus on children's raw reading math achievement; alternative analyses showing "same grade" comparisons that compare the achievement of the promoted and retained children to norms for their current grade level are reported in Wu et al. (2008a). As shown in Figure 3, the promoted children showed greater gains on the Woodcock–Johnson broad reading scores (weighted mean difference = 17.66) and math scores (weighted mean difference = 6.49) than retained children during Year 2 when the retained children were repeating first grade. By Year 4 of the study (fourth grade for promoted; third grade for retained children), this difference was substantially reduced for reading (weighted mean difference = 3.63) but not math achievement (weighted mean difference = 6.37). Inclusion of the significant baseline variable as a covariate did not alter the pattern of results.

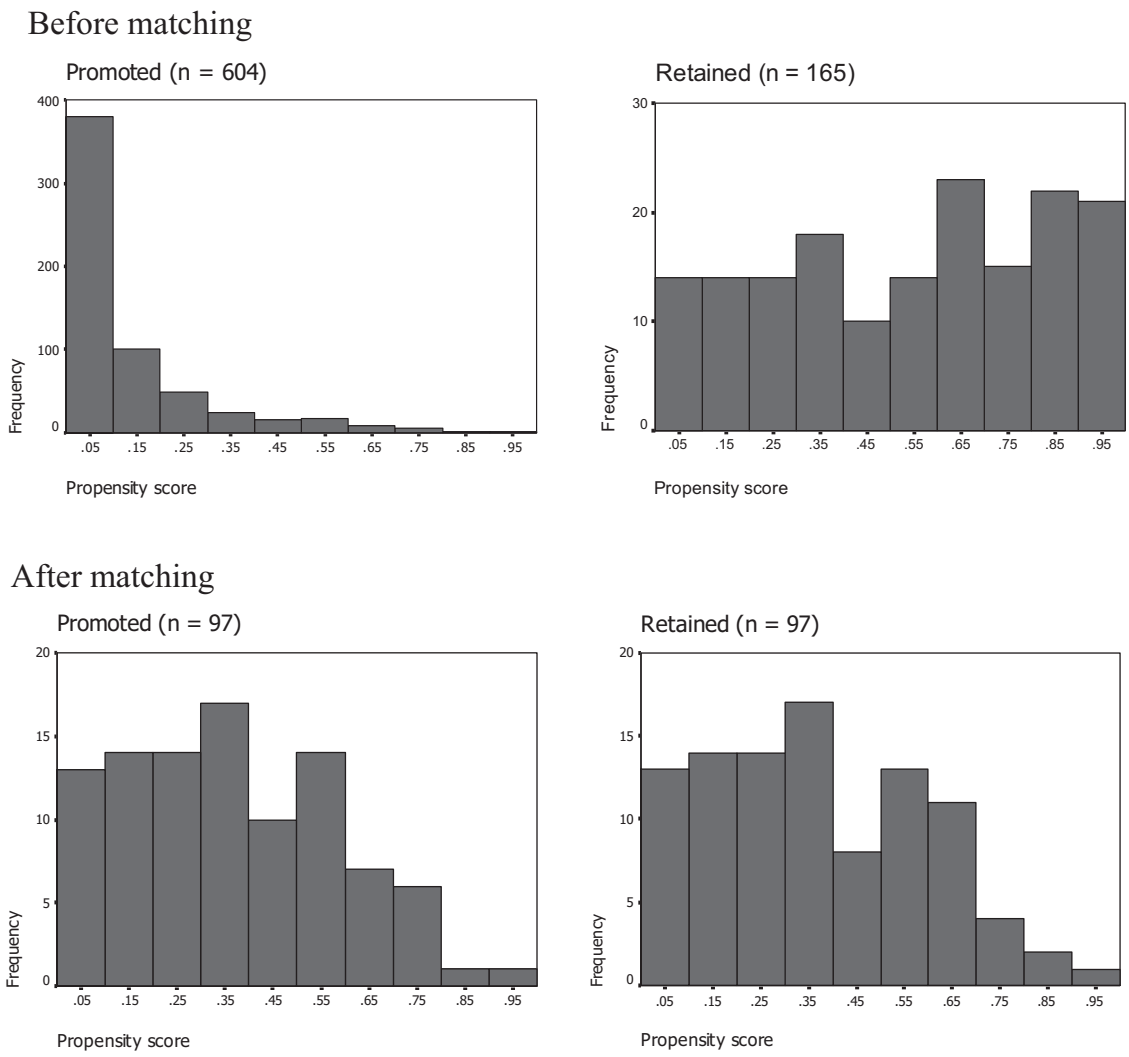

Figure 2. The distributions of the propensity score for promoted and retained groups before and after matching. The scale of the *y*-axis (Frequency) differs for the promoted and retained groups before matching, but it is the same after matching. (Adapted from "Short-Term Effects of Grade Retention on Growth Rate of Woodcock–Johnson III Broad Math and Reading Scores," W. Wu, S. G. West, & J. N. Hughes, 2008b, *Journal of School Psychology*, *46*, pp. 85–105.)

The strength of the present illustration largely hinges on the success of the matching procedure. Following Rubin (2006a), the matching variables were all collected prior to the treatment (retention in first grade), and propensity scores were estimated without knowledge of the outcome measures. The covariates selected for matching were extensive and based on input from subject-matter experts, although this knowledge is potentially fallible. Checks were performed on the balance of each of the covariates. The initial selection of a sample of children "at risk" for retention helped to yield nearly complete overlap between the range of the propensity scores in the treatment and control conditions. Nonetheless, there is no certainty that other hidden covariates may not have accounted for at least part of the observed treatment effect—nor is there certainty that even if the proper covariates were selected the estimated propensity scores closely tracked the true propensity scores in this small sample.

Comparing the Two Approaches: Observational Studies

Campbell's approach has emphasized the identification of threats to validity and the use of multiple, targeted design features to address threats to validity that characterize a specific research area. This strategy builds strongly upon current scientific knowledge. It develops complex sets of hypotheses about what the pattern of results should look like across multiple sets of design elements and compares the obtained results to the expected pattern. The research context determines the specific threats to validity and the targeted design elements that will be considered. With the exception of randomization, no design element is universally preferred to others. To the extent that the results match the hypothesized pattern, the design elements rule out confounding variables

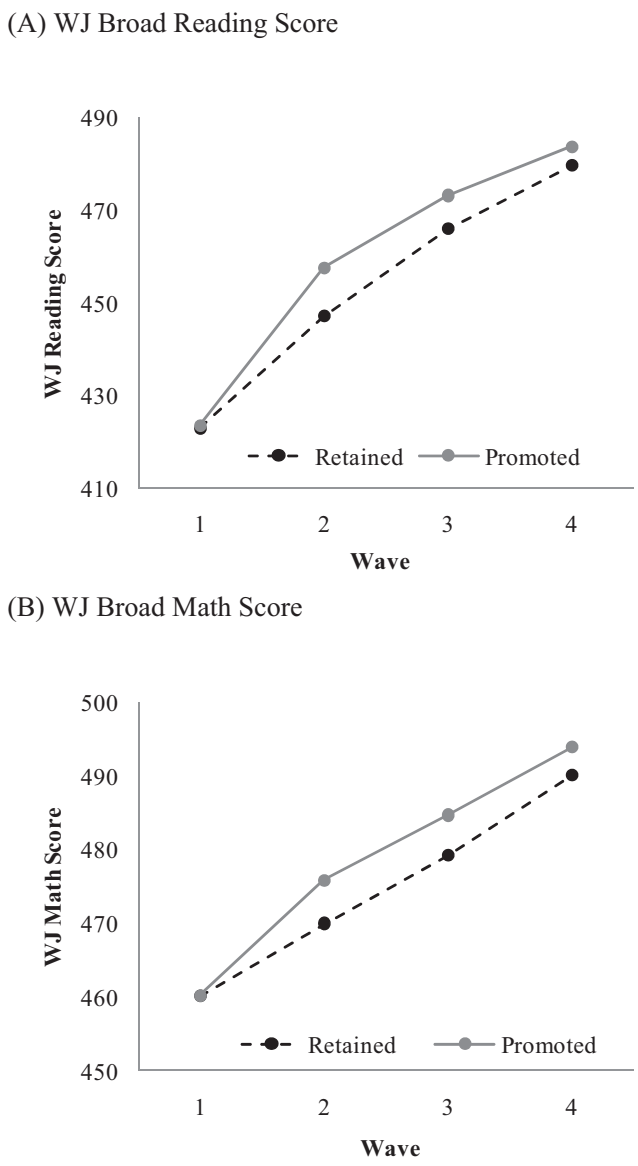

Figure 3. Means at each measurement wave for matched retained and promoted children for the Woodcock–Johnson (WJ) broad math and reading scores. Waves 1–4 correspond to Years 1–4 of the study.

that potentially serve as alternative explanations of the results. Failures to match the hypothesized pattern create uncertainty about whether a true causal effect exists.

Problems may occur in Campbell's approach for three primary reasons. First, as in all scientific endeavors, current scientific knowledge is fallible and, hence, may be incorrect. There may be less than complete certainty that all threats to internal validity (potential confounders) have been identified or that a specific design element will rule out the identified threat to validity. Second, some parts of the hypothesized pattern of results represent "no difference" predictions. For example, two different control groups are expected not to differ from each other; a nonequivalent dependent variable is expected to show no effect of treatment. Extreme care must be taken to achieve an adequate sample size and to use highly reliable and sensitive measures to minimize the possibility that a true effect of material size is *not* being detected (Meehl, 1967). Third, in Campbell's tradition, threats to validity have historically been ruled out in an all or none manner, that is, fully ruled out or not ruled out. Reichardt (2000, 2006) has attempted to develop procedures that specify the magnitude of the portion of the threat that has been ruled out, but this work is still in a preliminary stage of development.

Rubin's tradition has emphasized the development of elegant mathematical theory that attempts to yield precisely matched treatment and comparison groups on a large set of measured covariates. Matching on propensity scores can yield unbiased estimates of the causal effect. Potential problems arise in practice because estimated propensity scores are used instead of true propensity scores. The extent of bias reduction will depend on the success of the estimated propensity score as a realization of the selection model in the sample. Drake (1993) and Rubin and Thomas (1996) have suggested that overfitting a true propensity score model can lead to the benefit of having even more balanced T and C groups at baseline in the sample. The most important key to the success of the propensity score model will be the quality of the fallible scientific input regarding choice of baseline covariates. Success will also depend on the quality of the measurement (e.g., reliability) of baseline covariates. It will further depend on the researcher's care in checking the balance of the estimated propensity scores, recognizing that even careful checks may not be fully adequate in small samples. Even with the best efforts, there may be omitted covariates that were not considered in the baseline measurement. For example, economists (such as Manski, 1999; Manski & Pepper, 2000) argue that participant's perceptions of the likely effectiveness of each treatment may be an important, often overlooked determinant of treatment selection. Despite these concerns, work within the tradition of Rubin's causal model has focused almost exclusively on matching techniques because of their strong mathematical justification and theoretical ability to provide an exact estimate of the magnitude of the causal effect (but see Cochran, 1965; Rosenbaum, 1999).

Campbell's tradition (e.g., Campbell & Erlebacher, 1970) has historically been skeptical about matching and has *not* given matching a privileged status relative to other design elements. Four primary concerns have been raised. First, important covariates may be omitted as noted above. Second, measures of key covariates may be measured with less than perfect reliability so that incomplete adjustment takes place. Little work has been conducted to investigate the extent to which this is a major issue when using propensity scores based on many covariates. However, from the standpoint of Rubin's perspective, unreliability may simply be viewed as another form of the omitted covariate problem. Third, in some research situations, participants in the T and C groups may be growing at different rates (Selection $\times$ Maturation), a confounding that is unlikely to be captured by covariates measured in a single baseline measurement. Within the potential outcomes framework, Haviland et al. (2007) have offered an approach that addresses the combination of differential growth rates and differential levels on baseline covariates, but their approach requires the addition of the design element of multiple pretests to estimate the pretreatment growth rates. Fourth, sufficient overlap may not exist between the T and C groups on matching variables. The development of propensity scores and the identification of a common support region in which members of both the T and C groups have some probability

of being in the T group, that is, $0 < P(T) < 1$, has largely addressed this objection.

The first three concerns above reflect omitted covariates that result in hidden bias. An important approach to hidden bias is to develop sensitivity analyses (Gastwirth, Krieger, & Rosenbaum, 1998; Marcus, 1997; Rosenbaum, 2002) that produce estimates of the maximum and minimum magnitude of the causal effect under different assumptions about the magnitude of hidden bias. Such analyses appropriately convey the uncertainty that the propensity scores have been properly estimated by providing brackets on the possible magnitude of the causal effect.

Some Dimensions of Comparison of the Two Perspectives

In this section, we highlight similarities and differences of the two approaches on several dimensions. Our comparison is largely based on the material reviewed above but also briefly introduces other key issues that could not be fully considered because of space limitations. Table 3 summarizes the dimensions described below.

Domain of Application and Research Strategy

Campbell's perspective comes from a synthesis of research criticism in basic and applied research in psychology and education as well as evaluation research. The perspective followed inductive principles and synthesized the criticism into a set of general threats to validity. Rubin's perspective developed out of the application of frequentist and Bayesian statistical principles to important research problems in medicine, public health, economics, and social research—areas that tend to have an applied focus. Following the identification of a significant problem, the work followed deductive principles in which a set of assumptions were made, and an estimate of the causal effect was mathematically derived within the potential outcomes framework.

Substantive Science and the Definition of Treatment and Outcome Variables

Both Campbell's and Rubin's perspectives place a strong emphasis on clear operationalization of the T and C treatments. Given the emphasis in psychology on basic research testing verbal theories that purport to explain the observed effect, Campbell's approach places great emphasis on construct validity (Shadish et al., 2002). For example, does frustration produce aggression as theorized or is some other construct, such as physiological arousal, responsible for the observed effect? Campbell's focus is less on the specific operation that is used to produce frustration but on the underlying construct itself. Such concerns lead to series of experiments in which distinct treatments that purportedly manipulate only the prototypic features of the frustration construct (i.e., blockage of a goal directed activity) are contrasted with various comparison treatments. This strong emphasis on the construct validity of the independent variable characterizes much basic research in psychology (e.g., social psychology; Aronson, Wilson, & Brewer, 1998). These experiments may also include diverse operational measures of the dependent variable, possibly including multimethod assessment (Campbell & Fiske, 1959; Eid & Diener, 2006). This focus on the construct validity of the dependent variable has more often been emphasized in applied than basic research in psychology, although illustrations can be found in both. The hope is that a theoretically coherent set of results will emerge that will explicate the specific constructs that are producing this and perhaps other theoretically linked causal effects. Such understanding of the causal mechanisms responsible

Table 3
Differences in Emphasis Between Campbell's Approach and Rubin's Approach

Dimension	Campbell's approach	Rubin's approach
Domain of application	Psychology, education, large scale social research	Public health, medicine, economics, large scale social research
Type of research	Basic, applied	Applied
Treatment, outcome definition	Constructs	Operations
Key feature of causal inference	Threats to validity	Precise assumptions, formal use of potential outcomes model
Approach to development of principles	Inductive, scientific: Synthesize ideas on the basis of practical research experience	Deductive, mathematical: Makes assumptions and derives model
Primary methods of strengthening causal inference given threats to validity, assumption violations	Prevention of threat, addition of design elements, pattern of results rule out threat	Checking assumptions, sensitivity analyses, alternative approach on the basis of weaker assumptions, more robust statistical analysis
Causal effect estimate	Exact magnitude if assignment rule known; otherwise direction only	Exact magnitude
Role of measurement	Strong emphasis on measurement problems	Less emphasis on measurement problems
Causal generalization (single studies)	Formal sampling model, Cook's grounded theory	Formal sampling model, limited generalization
Causal generalization (multiple studies)	Meta-analysis, Cook's grounded theory	Response surface analysis

for the effect provides one basis for generalization of the findings to new populations, treatments, outcomes, and settings (Cook, 1993).

In contrast, perhaps reflecting the more applied perspective of his fields of application, Rubin's approach has focused on careful definitions of the implemented treatments rather than the hypothetical constructs that they may represent. Learning that a specific medical device or school voucher program produces positive effects relative to a well-defined comparison treatment often provides sufficient information to answer the specific applied question. If concerns about construct validity arise within Rubin's perspective, they are far more likely to be focused on the theoretical meaning of the dependent rather than independent variables. The relative focus on construct validity represents a central difference between the two perspectives.

Threats to Validity Versus Precise Assumptions

A key feature of Campbell's perspective is a list of threats to validity distilled from the field's accumulated methodological knowledge. This list identifies known general problems that can undermine the causal inference process. Researchers then compare features of their specific research setting with these threats and identify those threats they believe to be plausible in their specific context. There is no guarantee that the list of threats is complete (although the rate of discovery of new threats has dwindled to near 0) or that researchers will make correct decisions about which threats are plausible in their research context. Campbell (1988) argued that mutual scientific criticism of important findings would identify any omitted threats in the long run.

Precise assumptions form the analogous key feature in Rubin's perspective. The mathematics of the potential outcomes model guarantees that an unbiased estimate of the precise magnitude of the causal effect will be achieved if the assumptions are met. This is why Rubin has so strongly emphasized assumptions that are potentially verifiable. For Campbell, failure to rule out a threat to validity undermines causal inference; for Rubin, failure to meet a critical assumption undermines causal inference. Many of the assumptions needed for causal inference in Rubin's approach will be also needed in Campbell's approach; they are simply far less explicit because of the lack of a formal mathematical development.

Addressing Threats to Validity and Violation of Assumptions

When plausible threats to validity cannot be ruled out by the basic design, Campbell's perspective strongly prefers to address these threats in the planning phase of the study. The strategy is one of active prevention of threats before they occur and the addition of targeted design elements that potentially make the specific threat implausible given the expected pattern of results. Design elements, such as multiple control groups and nonequivalent dependent variables, are very general and can potentially be applied by the researcher to a wide variety of designs, including many not considered here (e.g., regression discontinuity; interrupted time series).

Rubin's approach has emphasized general design approaches that maximize the plausibility that the optimal approximation to the potential outcomes ideal will be realized, which increase the efficiency of the design, or both. Rubin has not used targeted design elements to address specific threats to validity. When design approaches are not available, the emphasis has been on checking assumptions to be sure they are viable. When they are not, Rubin's perspective has emphasized the development of "statistical cures" for important problems that result from failures of assumptions. Within the potential outcomes framework, important new statistical procedures—such as the treatment noncompliance model (Angrist et al., 1996)—have already been developed, and new developments are ongoing (e.g., Frangakis & Rubin, 2002; Haviland et al., 2007). The derivation of these procedures has often involved the introduction of additional statistical assumptions. In addition, new, more robust statistical procedures have been developed that are less sensitive to the violation of critical assumptions. When assumptions cannot be directly checked (e.g., hidden variable bias in an observational study), Rubin's approach has emphasized sensitivity analyses that identify the extent to which violations of assumptions of a specified magnitude would alter the magnitude of the causal effect. One of the strengths of sensitivity analysis is that the effects of multiple assumptions can be investigated both singly and in combination (interaction). Rosenbaum (2002) has summarized several robust nonparametric methods for estimating treatment effects as well as a number of approaches to sensitivity analyses in observational studies.

Direction Versus Magnitude of Causal Effect

The historical focus of Campbell's perspective was on the direction of rather than the magnitude of the causal effect. Threats to internal validity were treated as present or absent. The recent emphasis on the magnitude of the effect in psychology provides challenges for Campbell's perspective that have not been fully resolved. Campbell's perspective provides clear estimates of the causal effect when the mechanism of assignment to treatment conditions is known, as in the randomized experiment, the regression discontinuity design, or the interrupted time series design. However, the addition of targeted design elements to address specific threats to internal validity currently only permits the inference of directional effects. What if a targeted design element only partially rules out a threat to validity? What if the pattern of results produced by different targeted design elements is not fully consistent? Reichardt (2000, 2006) and Rosenbaum (2002) have begun to address these issues. In addition, some advocates of Campbell's perspective would argue that a correct conclusion about the direction of the causal effect is often sufficient (see Jones & Tukey, 2000)—variations in the magnitude of the causal effect of one construct on another are to be expected as a function of the population, specific treatment implementation, specific measure of the outcome, and the context. This variation may typically be of only minimal scientific or policy interest, except when the reversal or the elimination of a causal effect occurs.

Rubin's perspective allows the unbiased estimation of the exact magnitude of the causal effect when the assumptions are met. When the assumptions are not met, sensitivity analyses are conducted that attempt to precisely bracket the magnitude of the causal effect as a function of the degree to which the assumptions are violated. In practice, the uncertainty appropriately attached to

the precise magnitude of causal effects estimated from Rubin's perspective when assumptions are violated may yield little more than a statement of a directional effect. Campbell's approach has also suggested bracketing the magnitude of causal effects, but, with rare exceptions (Shadish, Hu, Glaser, Kownacki, & Wong, 1998), it has not provided formal methods for attempting to specify precise brackets.

The Role of Measurement

With its strong roots in psychology and education, Campbell's approach has afforded measurement operations a strong role.[13] Great emphasis has been placed on the reliable and valid measurement of the intended construct. Adding checks or design elements to be sure problems of instrumentation can not account for the estimated causal effect has received great emphasis. Addressing potential threats arising from regression to the mean that may arise when groups are selected on the basis of baseline scores has also received considerable emphasis in both the design and the analysis of studies (Campbell & Erlebacher, 1970; Campbell & Kenny, 1999). When regression to the mean may have occurred, recommendations are typically made to correct statistical adjustments of treatment effects for unreliability in the measurement of baseline covariates, lack of test–retest reliability in measurement, or both (e.g., Cook & Campbell, 1979; Judd & Kenny, 1981). In contrast, Rubin's perspective has given measurement issues far less attention. Reflecting the tradition in statistics, Rubin has typically emphasized the results of the measurement operation rather than "true scores" associated with the intended hypothetical construct. From Rubin's perspective, true score models involve another layer of assumptions (e.g., true score is uncorrelated with error of measurement) that are nonverifiable and therefore suspect. Further, many of the standard approaches used in psychological research to address regression to the mean may be incomplete. Instead, measurement error can be treated as simply contributing to the problem of hidden bias, and its impact can be explored through sensitivity analysis.

Causal Generalization

Within both Campbell's and Rubin's perspectives, the generalization of causal effects is straightforward if the sample has been randomly sampled from a defined population (Draper, 1995; Kish, 1987). Within Campbell's perspective, both Cook (1993) and Shadish et al. (2002) have noted that the samples of participants are not selected following any formal probability sampling model in practice (for two rare exceptions, see, e.g., Schwarz & Hippler, 1995; Wolchik et al., 2000). Otherwise stated, nearly all studies in psychology and education involve samples of convenience, subject only to possible eligibility restrictions. Further, there is often a desire to generalize beyond the specific (a) units, (b) treatments, (c) outcome measures, (d) settings, and (e) times involved in the study. For example, Weisz, Weiss, and Donenberg (1992) argued that university-based randomized controlled trials of a new therapy with highly selected patients under near state of the art conditions may not be a very representative realization of the therapy as it is delivered by a community mental health centers to the patients they serve. Cook (1993) synthesized existing empirical observations into a grounded theory of causal generalization that identifies five scientific principles that facilitate the generalization of directional causal effects.

1. *Proximal similarity.* The treatments, participants, settings, response measures, and times should include most of the central features of the population of interest.

2. *Heterogeneous irrelevancies.* Aspects of treatments, participants, settings, response measures, and times that are theoretically expected to be irrelevant to the causal relationship should be made as heterogeneous as possible.

3. *Discriminant validity.* "[W]e can interpret and label the operations used in a study more accurately if we can discriminate between different constructs that usually have overlapping content" (Shadish et al., 2002, p. 364). To the extent that the treatment affects the intended outcome, but not other similar constructs, the likelihood of causal generalization is supported. To the extent that precise types of participants or settings can be identified for which the treatment effect holds, the likelihood of generalization to the specific subpopulations of persons or settings is increased.

4. *Causal explanation.* To the extent that a theoretical explanation of the causal effect can be supported, the likelihood of generalization can be supported. Causal explanation is enhanced to the degree that the construct validity of the independent and dependent variables can be established.

5. *Empirical interpolation and extrapolation.* Causal effects are far more likely to generalize within rather than beyond the range of treatments (e.g., within the range of dosages studied), persons, settings, times, and response measures that have been studied.

In Campbell's tradition, these principles should be built into the design to the extent possible during the planning of studies if causal generalization is sought. The five principles can also be applied to the meta-analysis of research literature to facilitate the understanding of the generalization of causal effects with different populations of participants, different treatment variations, different outcome measures, and different settings. There is no proof that generalization will be achieved in the new setting, but the use of these principles is expected to substantially enhance its likelihood.

In contrast, Rubin's perspective has been primarily concerned about the generalization of the causal effect defined by the specific treatments being compared, the specific outcomes, and the specific setting to a population of which the sample participants are representative. When the formal probability sampling model cannot be applied, generalization is to a hypothetical population (superpopulation) defined jointly by the sample recruitment processes of the study and the participants for whom possible outcomes can be defined. To cite two examples, generalization of the LATE effect is to a population of treatment adherers who would participate in the randomized experi-

[13] Another source of the prominence of measurement in Campbell's approach may be the frequent use of self-reports, informant reports, and standardized tests as measurement operations in psychology and education. These instruments may be especially prone to changes over time and over context that may make it difficult to claim that the same construct is being measured at baseline and outcome. Although more subtle, many physical and biological measurements share these same problems. For example, consider the problem of calibrating functional magnetic resonance imaging or cholesterol measures taken at baseline and outcome, often with different (improved) equipment or scored using different procedures or by different laboratories.

ment. When the distributions of propensity scores do not overlap in an observational study, generalization of the causal effect is limited to the common support region for which potential outcomes are defined. Careful weighting of cases is done to estimate the causal estimand of interest, for example, the causal effect for the population willing to receive treatment. However, for single studies, there is no formal mathematical basis within Rubin's approach to generalize the estimate of the precise causal effect to other populations, treatments, observations, settings, and possibly times of interest.

For multiple studies, Rubin (1992) has proposed an elegant meta-analytic approach based on response surface modeling. The effect sizes of each study serve as the outcome of interest, and the study characteristics and study quality constitute two important dimensions of predictor variables. By using information from all available studies, a response surface representing the relationships between the predictors and the effect size can be estimated. The effect sizes of scientific interest are those for the high-quality studies. The effect size for the average causal effect or the effect size at any location on the surface can be estimated. This method highlights areas of sparse empirical research and clearly identifies whether interpolation within or extrapolation beyond the range of existing studies is taking place. Theoretically, this response surface approach offers a principled method of generalizing magnitudes of treatment effects to the units, treatments, observations, settings, and possibly times of interest. In practice, quantitative scaling of the dimensions of generalization will be often challenging. Mabe and West (1982) have reported an early application of this general procedure to estimate the validity of self-reports. Shadish, Matt, Navarro, and Phillips (2000) used this procedure to estimate the magnitude of the causal effects of psychotherapy under clinically representative conditions.

Conclusion

In this article, we have provided an introduction to Campbell's and Rubin's perspectives on causal inference. We have also considered two common designs, the randomized experiment and the observational study, from each of these perspectives. We have compared the two perspectives on several dimensions. Campbell's approach strongly focuses the researcher's attention on practical issues in the planning of the research. It develops strong methods to prevent threats to internal validity from arising and encourages researchers to add design elements that can help rule out those threats should they occur. It also provides researchers with strategies for maximizing the likelihood that their results can be generalized to other applied problems of interest. It draws heavily on current scientific knowledge to inform the interpretation of patterns of results. In some cases, these features may limit conclusions to the direction, but not the magnitude, of causal effects. Rubin's work has primarily emphasized providing methods for the precise estimation of causal effects. The approach has emphasized careful definition of the treatments that are being compared and the population to which the causal estimate can be applied. It has led to design solutions to produce clear estimates of causal effects and statistical solutions when underlying assumptions cannot be met. It uses the machinery of mathematics to provide precise estimates with specified uncertainty regarding the magnitude to the causal effect. The cost of this precision is that Rubin's approach provides little basis for generalization of causal effects in single, but not multiple, studies. From our perspective, the two approaches are largely complementary, but they have different emphases. Over time, methodologists will continue to explicate the key differences between the two approaches. However, researchers can already benefit from drawing on the complementary strengths of both approaches in the design, analysis, and interpretation of their studies (see also West et al., 2008). Perhaps this will be another yet affirmation of Campbell's belief in the power of the scientific method of mutual criticism to improve scientific practice and ultimately bring us closer to the "truth" of scientific claims (Overman, 1988).

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion and rejoinder). *Journal of the American Statistical Association, 91,* 444–472.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). New York: McGraw-Hill.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or- none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association, 93,* 929–934.

Barnard, J., Du, J., Hill, J. L., & Rubin, D. B. (1998). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research, 27,* 285–317.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City (with commentary). *Journal of the American Statistical Association, 98,* 299–323.

Barnoya, J., & Glantz, S. A. (2005). Cardiovascular effects of secondhand smoke: Nearly as large as smoking. *Circulation, 111,* 2684–2698.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation.* Thousand Oaks, CA: Sage.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–77). San Francisco: Jossey-Bass.

Campbell, D. T. (1988). Can we be scientific in applied social science? In E. S. Overman (Ed.), *Methodology and epistemiology for social science: Selected papers of Donald T. Campbell* (pp. 315–334). Chicago: University of Chicago Press.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate. Vol. 3: Disadvantaged child* (pp. 185–225). New York: Brunner/Mazel.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cochran, W. G. (1965). The planning of observational studies in human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128,* 234–255.

Cochran, W. G. (1983). *Planning and analysis of observational studies.* New York: Wiley.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of

causal relationships. In L. B. Sechrest & A. G. Scott (Eds.), *New directions for program evaluation* (No. 57, pp. 39–81). San Francisco: Jossey-Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cook, T. D., Dintzer, L., & Mark, M. M. (1980). The causal analysis of concomitant time series. In L. Bickman (Ed.), *Applied social psychology annual* (Vol. 1, pp. 93–135). Beverly Hills, CA: Sage.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics, 69,* 1321–1326.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61,* 50–55.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86,* 366–379.

Frangakis, C., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21–29.

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika, 85,* 907–920.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12,* 247–267.

Higginbotham, H. N., West, S. G., & Forsyth, D. R. (1988). *Psychotherapy and behavior change: Social, cultural, and methodological perspectives.* New York: Pergamon.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, A. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics, 1,* 69–88.

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

Hume, D. (2007). *An enquiry concerning human understanding* (P. Millican, Ed.). New York: Oxford. (Original work published 1748)

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review, 30,* 420–437.

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7,* 178–193.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5,* 411–414 [Correction: *Psychological Methods, 6,* 17].

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

Kish, L. (1987). *Statistical designs for research.* New York: Wiley.

Lehman, D., Wortman, C., & Williams, A. (1987). Long-term effects of losing a spouse or a child in a motor vehicle crash. *Journal of Personality and Social Psychology, 52,* 218–231.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70,* 556–567.

Little, R. J., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5,* 459–476.

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Conceptual and analytical approaches. *Annual Review of Public Health, 21,* 121–145.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Little, R. J., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 247–259.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluations of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296.

Manski, C. F. (1999). Choice as an alternative to observational studies: Comment. *Statistical Science, 14,* 279–281.

Manski, C. F., & Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica, 68,* 997–1010.

Marcus, S. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics, 22,* 193–202.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9,* 403–425.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56,* 118–124.

Neyman, J. (1990). On the application of probability theory to agriculture experiments: Essay on principles (Section 9). *Statistical Science, 5,* 465–480. (Original work published 1923)

Overman, E. S. (1988). *Methodology and epistemology for social science: Selected papers of Donald T. Campbell.* Chicago: University of Chicago Press.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 89–115). Thousand Oaks, CA: Sage.

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods, 11,* 1–18.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science, 14,* 259–278.

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

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association, 102,* 191–200.

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

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

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrics, 29,* 581–592.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

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

Rubin, D. B. (1986). Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association, 81,* 961–962.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics, 17,* 363–374.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rubin, D. B. (2006a). *Matched sampling for causal effects.* New York: Cambridge.

Rubin, D. B. (2006b). Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 769–800). Amsterdam: Elsevier.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573–585.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schwarz, N. B., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59,* 93–97.

Shadish, W. R., & Cook, T. D. (1999). Design rules: More steps towards a complete theory of quasi-experimentation. *Statistical Science, 14,* 294–300.

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

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R. J., & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3,* 3–22.

Shadish, W. R., Matt. G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126,* 512–529.

Sheiner, L. B., & Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapy, 57,* 6–10.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association, 101,* 1398–1407.

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 39–54.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (2009). *Causal effects in between-group experiments and quasi-experiments: Theory.* Book in preparation.

Stuart, E., & Rubin, D. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155–176). Thousand Oaks, CA: Sage.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

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). New York: Cambridge.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D., Holtgrave, D., et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health, 98,* 1359–1366.

West, S. G., & Sagarin, B. J. (2000). Subject selection and loss in randomized experiments. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (Vol. 2, pp. 117–154). Thousand Oaks, CA: Sage.

Wolchik, S. A., West, S. G., Sandler, I. N., Tein, J.-Y., Coatsworth, D., Lengua, L., et al. (2000). An experimental evaluation of theory-based mother and mother–child programs for children of divorce. *Journal of Consulting and Clinical Psychology, 68,* 843–856.

Wu, W., West, S. G., & Hughes, J. N. (2008a). Effect of retention in first grade on children's achievement trajectories over four years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100,* 727–740.

Wu, W., West, S. G., & Hughes, J. N. (2008b). Short-term effects of grade retention on growth rate of Woodcock–Johnson III broad math and reading scores. *Journal of School Psychology, 46,* 85–105.

Received January 31, 2008
Revision received January 12, 2009
Accepted January 15, 2009 ■