Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings

William R. Shadish University of California, Merced

This article compares Donald Campbell's and Donald Rubin's work on causal inference in field settings on issues of epistemology, theories of cause and effect, methodology, statistics, generalization, and terminology. The two approaches are quite different but compatible, differing mostly in matters of bandwidth versus fidelity. Campbell's work demonstrates broad narrative scope that covers a wide array of concepts related to causation, with a powerful appreciation for human fallibility in making causal judgments, with a more elaborate theory of cause and generalization, and with a preference for design over analysis. Rubin's approach is a more narrow and formal quantitative analysis of effect estimation, sharing a preference for design but best known for analysis, with compelling quantitative approaches to obtaining unbiased quantitative effect estimates from nonrandomized designs and with comparatively little to say about generalization. Much could be gained by joining the emphasis on design in Campbell with the emphasis on analysis in Rubin. However, the 2 approaches also speak modestly different languages that leave some questions about their total commensurability that only continued dialogue can fully clarify.

Keywords: Rubin's causal model, causal inference, validity, SUTVA, propensity score

This article compares the approaches to causation in field settings taken by Donald T. Campbell and Donald B. Rubin. Campbell received his bachelor's degree in psychology from the University of California at Berkeley in 1939 and received his doctorate in social psychology from the same university in 1947. For most of his career, he was a psychologist at Northwestern University. His major work on causal inference was generated from the 1950s through 1970s. He is the inventor of such terms as internal validity, external validity, quasi-experimentation, and threats to validity, as well as the inventor of many quasi-experimental designs, such as the regression discontinuity design. He is one of the most widely cited and influential psychologists of all time, a polymath with contributions not just to methodology, but also to social psychology, sociology of science, and epistemology. The methodology and the approach developed by Campbell and his colleagues (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) is the most influential approach to causal inference in field settings in psychology and

Rubin received a bachelor's degree in psychology from Princeton University in 1965 and received a doctorate in statistics from Harvard University in 1970. After stints at the University of Chicago and the Educational Testing Service, he returned to Harvard as a statistician. He is widely regarded as the most influential statistician ever to work on the topic of nonrandomized experiments and practical methods of addressing problems with randomized field experiments (e.g., noncompliance with treatment assignment, attrition) across such disciplines as statistics, public health,

Correspondence concerning this article should be addressed to William R. Shadish, University of California, Merced, P.O. Box 2039, Merced, CA 95344. E-mail: wshadish@ucmerced.edu

and medicine. Rubin is known for his contributions to the development of methods such as propensity score analysis and what has come to be known as Rubin's causal model. He is also the author of many other important statistical innovations, including many modern methods for missing data analysis, such as multiple imputation (e.g., Little & Rubin, 2002). He is among the 10 most cited statisticians in the world. Most psychologists are unfamiliar with his approach to causal inference in field research, because it has been published mostly in journals in statistics and other fields and rarely in psychology. The exception, ironically, was the very first presentation of his work on causal inference, which appeared in the Journal of Educational Psychology (Rubin, 1974). Rubin's work occurred later than most of Campbell's, and the two scholars rarely cited each other. Even today, work in these two traditions proceeds largely independently, and until recently, cross-citations were pro forma rather than crucial.

Despite these differences, the two models share much in common. Both are enormously influential in their respective spheres of influence, with key works being citation classics. Both bring experimental language to observational research. Despite Campbell's broad concern with both external and construct validity, both Campbell and Rubin prioritize on simple descriptive causal inference about the existence and magnitude of a causal relationship, rather than more complex explanatory causal inference about when, how, and why that effect occurs. Both focus primarily on discovering the unknown effects of known causes rather than the unknown causes of known effects, but both have extended their work to the latter question. Both emphasize the central role that manipulable causes play in all experimental work. Both prefer causal inferences from randomized experiments, yet both make special and extended efforts to address causal inference in nonrandomized experiments and observational studies. Both are practical, providing tools that researchers can use in their cause-

probing work. Taking different routes toward similar goals, they probably have more in common than not.

However, the literature contains few attempts to compare and contrast their work to see if concepts from one tradition can help the other, to outline areas of agreement and disagreement, or to clarify whether any of the disagreements are important. The present article attempts such a comparative effort. The perspective is self-consciously from a scholar who specializes in Campbell's work, not just in experimentation, but also in related fields, such as program evaluation (Shadish, Cook, & Leviton, 1991) and the social psychology of science (Shadish & Fuller, 1994). My knowledge of Rubin's work is less secure, and my goal is less to criticize than to learn how to use Rubin's work to improve the Campbell tradition in particular and causal inference in psychology more generally. Moreover, I do not purport to represent any more than my views of Campbell's work. Many superb scholars have been involved in the Campbellian tradition for far longer than I have, examples of whom include Robert Boruch, Mel Mark, Charles Reichardt, Lee Sechrest, and William Trochim. Such scholars may well disagree with parts of my interpretation of Campbell's work. Compared to them, I hold no privileged position for interpreting the body of Campbell's work. Of course, that disclaimer is even more true regarding my representation of Rubin's work.

Campbell's and Rubin's Causal Models

In one sense, it is a misnomer to suggest that the broad-ranging work of these two theorists can be reduced to a model, for the latter phrase implies a compactness, precision, and singular focus that belies the wide-ranging span covered by the corpus of their works. Nonetheless, both of their approaches to causal inference have a finite core of ideas and terms that is uniquely their contribution and is widely attributed to them and that drives much of the rest of their work on causal inference. In that sense, for the sake of convenience, I refer to Campbell's causal model (CCM) and Rubin's causal model (RCM). The latter is a common referent in the literature (e.g., Holland, 1986), but the former is merely a convenient counterpoint to facilitate the contrasts in this article. These abbreviations also allow inclusive reference to all those who worked on CCM and RCM. For instance, parts of CCM were developed by Cook (e.g., Cook, 1990, 1991; Cook & Campbell, 1979), and parts of RCM were developed by Rosenbaum (e.g., Rosenbaum, 2002). References to CCM and RCM in this article refer to the models rather than to Campbell or Rubin themselves.

Campbell's Causal Model

The core of CCM is made up of Campbell's validity typology, the associated threats to validity, and a critical approach to the design and criticism of cause-probing studies using the validity threats. These first appeared in a journal article (Campbell, 1957), then in a greatly expanded book chapter (Campbell & Stanley, 1963), and finally in a reprint of that chapter as a freestanding book (Campbell & Stanley, 1966), which was revisited and expanded in book form twice over the next four decades (Cook & Campbell, 1979; Shadish et al., 2002) and was elaborated in many additional works.

CCM began with Campbell's (1957) formulation of a key dichotomy: that scientists make two general kinds of inferences from experiments. These include inferences about "did, in fact, the experimental stimulus make some significant difference in this specific instance" (Campbell, 1957, p. 297) and inferences pertaining "to what populations, settings, and variables can this effect be generalized" (Campbell, 1957, p. 297). He labeled those two inferences as *internal validity* and *external validity*, respectively, though he often used *representativeness* or *generalizability* in place of external validity.

That dichotomy was later expanded into four validity types (Cook & Campbell, 1979; Shadish et al., 2002):

- Statistical conclusion validity: The validity of inferences about the correlation (covariation) between treatment and outcome.
- Internal validity: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B, as those variables were manipulated or measured.
- Construct validity: The validity with which inferences can be made from the operations in a study to the theoretical constructs those operations are intended to represent.
- External validity: The validity of inferences about whether the cause–effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

In various incarnations, the validity types are the central concepts that give CCM such broad sweep and that allow CCM to apply to causal inferences across such different designs as the case study, path models, and experiments. The boundaries between the validity types are artificial but are consistent with common categories of discourse among scholars concerned with statistics, causation, language use, and generalization.

The second part of CCM is the idea of threats to validity, also referred to in CCM as either plausible rival hypotheses or alternative explanations. These are ways in which one may be wrong about the four kinds of inferences about causation, statistics, language use, and generalization. Regarding internal validity, for example, one may infer that results from a nonrandomized experiment support the inference that a treatment worked. It is possible to be wrong in many ways: Some event other than treatment may have caused the outcome (the threat of history), the participants may have changed on their own without treatment (maturation or regression), or the practice provided by repeated testing may have caused the participants to improve their performance without treatment (testing). Originally, Campbell (1957) presented 8 threats to internal validity and 4 threats to external validity. The lists proliferated, although they do seem to be reaching an asymptote: Cook and Campbell (1979) had 33 threats, and Shadish et al. (2002) had 37. Presentation of all threats for all four validity types is beyond the scope of the present article and is unnecessary to its central focus.

A third key feature of CCM is the use of validity types and threats to analyze and prevent likely inferential problems in the design of cause-probing studies, both retrospectively and especially prospectively. From the start, Campbell prioritized on internal validity, saying first that "internal validity is the prior and indispensable condition" (Campbell, 1957, p. 310) and later the now oft-quoted "internal validity is the sine qua non" (Campbell & Stanley, 1963, p. 175). From the start, this set Campbell at odds with some well-known contemporary statisticians like Cronbach (1982). Internal validity still holds a special place in CCM, because it makes no sense to experiment without some interest in whether a treatment works and because confidence in the basic causal inference about the effects of treatment must be present across studies in a program of research or else there may be no effect to generalize. Reflecting this priority, the CCM tradition focuses more on ways to design experiments that improve internal validity than on other validity types.

In CCM, the first line of attack toward good causal inference is to design studies that reduce "the number of plausible rival hypotheses available to account for the data. The fewer such plausible rival hypotheses remaining, the greater the degree of 'confirmation' " (Campbell & Stanley, 1963, p. 206). Assessing remaining threats to validity after a study is completed is the second line of attack, which is harder to do convincingly but is often the only choice when better designs cannot be used or when criticizing completed studies. Accordingly, CCM has always been on the lookout for new design tools that might improve causal inference. This included inventing the regression discontinuity design (Thistlewaite & Campbell, 1960) but mostly includes extending existing work, such as Chapin's (1932, 1947) experimental work in sociology, McCall's (1923) book on designing educational experiments, Fisher's (1925, 1926) already classic work on experimentation in agriculture, and Lazarsfeld's (1948) writings on panel designs. This eventually resulted in CCM giving a priority among the nonrandomized designs to regression discontinuity, interrupted time series with a control series, and complex patternmatching designs, typically in that order. The latter are designs that make complex predictions in which a diverse pattern of results must occur, with a study that may include multiple nonrandomized designs each with different presumed biases: "The more numerous and independent the ways in which the experimental effect is demonstrated, the less numerous and less plausible any singular rival invalidating hypothesis becomes" (Campbell & Stanley, 1963, p. 206). This emphasis on the role of complex predictions was part of the tradition of Rubin's mentor (e.g., Cochran, 1965) as well, but the RCM tradition does not do as much with the idea compared to CCM.

All this accurately suggests that CCM emphasizes the primacy of design over analysis, which Light, Singer, and Willett (1990) pithily stated as "you can't fix by analysis what you bungled by design" (p. viii). Thus, CCM stresses careful selection and addition of design features that can reduce the plausibility of a contextually important threat to validity, that can increase the comparability of treatment and comparison groups, and that can replace assumptions with data. In general, it is better to have a design that rules out a threat to validity at the start than to rely on fallible human judgment and statistical analysis to decide whether the threat is plausible after the fact. In a nonrandomized experiment, for example, a carefully chosen control group (a focal local control: in the same locale as the treatment group, focused on the same kinds of persons) is crucial and is presumed to be better than, for example, a random sample from a national database, which econ-

omists have used to construct control groups. For CCM, design rules (Shadish & Cook, 1999).

Finally, Campbell's thinking about causal inference existed in the context of his broader interests. In one sense, it is a special case of his interests in bias in human cognition in general. For example, as a social psychologist, he studied biases that ranged from basic perceptual illusions to cultural biases, and as a meta-scientist, he examined social psychological biases in scientific work. In another sense, CCM fits into a larger context of Campbell's ideas about evolutionary epistemology in which experiments are an evaluative mechanism by which potentially effective ideas are evaluated for retention in the knowledge base. CCM cannot be fully understood without reference to these larger contexts, but they are beyond the scope of this article (see Shadish et al., 1991, for more details).

Rubin's Causal Model

Rubin presented RCM, a compact and precise conceptualization of causal inference (e.g., Holland, 1986), though he frequently credits the model to Neyman (1923/1990). A good summary is found in Rubin (2004). The three key elements of RCM are units, treatments, and potential outcomes. If Y is the outcome measure, define Y(1) as the potential outcome that would be observed if the unit were exposed to a treatment (W = 1), and define Y(0) as the potential outcome that would be observed if that same unit were not exposed to the treatment (W = 0). If so, then the (potential) effect is the difference between these two potential outcomes, Y(1) - Y(0). This effect is defined on each unit, and the average of individual causal effects is the average causal effect. However, these are potential outcomes only until treatment begins. After treatment begins, only Y(1) or Y(0) can be observed, with the other being missing. Therefore, the fundamental problem of causal inference in RCM is how to estimate the missing outcome. These missing outcomes are sometimes called counterfactuals because they are not, in fact, what was observed. Further, after treatment begins, individual causal effects cannot actually be estimated as defined [Y(1) - Y(0)], although the average causal effect over all units can be observed under conditions such as random assignment.

A key assumption of the model is stable-unit-treatment-value assumption (SUTVA), which asserts that the representation of potential outcomes and effects in the preceding paragraph reflects all values that could be observed in the study. For example, SUTVA asserts that there is no interference between units, that the outcome observed on one unit is not affected by the treatment received by another unit. A common violation of SUTVA is dependence among units caused by nesting (e.g., of children within classrooms) when that nesting is not taken into account in the analysis. However, SUTVA goes beyond this well-known dependency problem. Commonly used examples are that my receipt of a flu vaccine may affect the likelihood that you will be infected or that my taking aspirin for my headache may affect whether you get a headache from listening to me complain about my headache. What such violations of SUTVA imply is that a unit no longer has just one potential outcome for receipt of treatment but rather has a set of potential outcomes depending on which treatment condition other participants receive. This set rapidly increases in size with more treatments and more participants, eventually making the computations impossibly complex. For example, consider an ex-

periment with just 2 participants (P1 and P2). With SUTVA, P1 has only two potential outcomes, one if P1 receives treatment, Y(1), and the other if P1 receives the comparison condition, Y(0). However, without SUTVA, P1 now has four potential outcomes, Y(1) if P2 receives treatment, Y(1) if P2 receives the comparison condition, Y(0) if P2 receives treatment, and Y(0) if P2 receives the comparison condition. If the number of participants increases to 3, the number of potential outcomes assuming SUTVA is still two for each participant, but without SUTVA, it is eight. With the number of participants characteristic of real experiments, the number of potential outcomes for each participant without SUTVA is so large as to be intractable. Also, even if there is no interference between units, researchers may have to worry that the ith unit has more than one version of each treatment condition available, such as an ineffective or an effective aspirin tablet. SUTVA is a simplification to make causal inference possible under such real world complexities. Clearly, however, SUTVA is not always true in field settings; thus, the assumption that units really have only one potential outcome may indeed be a necessary simplification, but it is not clear that it is always plausible. Most readers find SUTVA, and the implications of violations of SUTVA, to be one of the more difficult concepts in RCM, a point to which this article returns later.

The assignment mechanism by which units do or do not receive treatment is crucial. Even though researchers cannot observe both potential outcomes on any one unit, random assignment of units to conditions allows the average causal effect to be an unbiased estimate of the population causal effect. By virtue of random assignment, one of the two potential outcomes for each unit is missing completely at random. Both intuition and formal statistical theory (Rubin, 2004) tell us that this random missingness of unobserved units should not affect the mean over the observed units, at least within sampling error in any one experiment, and on expectation in general. When assignment is not random, the situation is more complex. In some cases, assignment is not completely random but is based on an observed variable, as when the regression discontinuity design assigns units to conditions solely on the basis of a cutoff on an observed variable (Shadish et al., 2002) or when random assignment occurs in conjunction with a blocking variable. Such assignment is called ignorable because potential outcomes are unrelated to treatment assignment once those known variables are included in the model, so an unbiased estimate can still be obtained. In all other cases of nonrandom assignment, assignment is based on a combination of factors, including unobserved variables, in which case assignment is nonignorable, and estimating effects is more difficult or impossible.

The assignment mechanism affects both the probability of being assigned to a particular condition and how much the researcher knows about the variables affecting that probability. In an experiment in which units are assigned to one of two conditions by a coin toss, that probability is widely understood to be p=.50, and it is widely understood, as well, that no other covariates will be systematically related to that probability (e.g., whether the coin comes up heads or tail is not influenced by one's gender). Assignment probabilities are formalized in RCM as propensity scores. In practice, randomized experiments are subject to sampling error (or unhappy randomization) in which covariates from a vector of all possible covariates X (measured or not) may be imbalanced across conditions (e.g., a disproportionate number of male participants are

in the treatment group). In those cases, the observed propensity score is related to those covariates and varies randomly from its true value (Rubin & Thomas, 1992).

When random assignment is not used but the assignment mechanism is ignorable as defined earlier, the true propensity score is a function of the known assignment variables. Here, X takes on a somewhat different meaning than in a randomized experiment. In a regression discontinuity design, for example, where persons are assigned to condition solely on the basis of whether they fall above or below a cutoff on an assignment variable, X must contain that assignment variable in addition to any other covariates the researcher may wish to measure. Rubin (2004) referred to those designs where one or more covariates in X fully determines assignment to conditions as regular designs, which form the basis for further analysis of observational, nonrandomized studies in which assignment is not controlled. In those other nonrandomized experiments, X would ideally contain all the variables that determined whether units received treatment, but in practice those variables are, at best, only partly known with certainty. In these cases, the true propensity score is unknown and is estimated with methods such as logistic regression in which selection into treatment or comparison condition is predicted from covariates. RCM suggests rules for knowing when the propensity scores are good ones, though much work on that topic is ongoing. Good propensity scores can be used to create balance over conditions on all observed covariates used to create the propensity scores (e.g., Rubin, 2001), but this is not sufficient for bias reduction unless the strong ignorability assumption is also met. The latter is a critical assumption discussed in more detail shortly.

From such things as the use of matching or stratification to improve the analysis of regular designs, RCM considers the case of obtaining accurate effect estimates for studies in which assignment is not initially designed to be ignorable—for example, by matching or stratifying on propensity scores in nonrandomized designs. Units matched on propensity scores are matched on the set of covariates used to create the propensity scores, and units within a propensity score stratum are similar on those covariates as well. If it can plausibly be argued that the covariates include all the pertinent ones bearing on assignment, RCM argues that the assignment mechanism can now be treated as unconfounded—the strongly ignorable treatment assignment assumption. The latter assumption is crucial, and unfortunately there is no direct test of whether it has been met—though this should not be seen as a flaw in RCM because RCM merely formalizes the uncertainty implicit in lack of knowledge of assignment in any nonregular design. RCM treats the matching or stratification process as part of the design of a good observational study in the sense that creating propensity scores, assessing their balance, and doing the initial matching or stratification are all done without looking at the outcome variable, approximating as well as one can the design of a prospective experiment. Standard analyses can then estimate the effect of treatment. RCM emphasizes that this rests on the ignorability assumption and provides ways to assess how sensitive results might be to violations of ignorability (Rubin, 2001).

RCM then moves on to more advanced topics. One topic is dealing with treatment crossovers and incomplete treatment implementation, combining RCM with econometric instrumental variable analysis to deal successfully with this key problem if some strong but often plausible assumptions are met (Angrist,

Imbens, & Rubin, 1996). A second example is getting better estimates of the effects of mediational variables (coming between treatment and outcome, caused by treatment, and mediating the effect; Frangakis & Rubin, 2002). A third example is dealing with missing data in the covariates used to predict the propensity scores (D'Agostino & Rubin, 1999). A fourth is dealing with clustering issues in RCM (Frangakis, Rubin, & Zhou, 2002). These examples are only glimpses of the yield of RCM in the design and analysis of cause-probing studies.

As was the case with Campbell, a full understanding of RCM requires knowledge of the larger context of Rubin's other work. Two features are especially noteworthy. First, his mentor was William G. Cochran, a statistician with a persistent and detailed interest in estimation of effects from nonrandomized experiments; Rubin's dissertation concerned the use of matching and regression adjustments in nonrandomized experiments. This context helped create and shape the nature of his interests in practical field experimentation. Second, Rubin is a pioneer in methods for dealing with missing data (e.g., Little & Rubin, 2002; Rubin, 1987), the effects of which can be seen in his conceptualization of potential outcomes as missing completely at random in randomized experiments and of the role of multiple imputation in explaining a Bayesian understanding of computing the average causal effect (Rubin, 2004).

Having described the basic features of RCM and CCM, this article now moves on to compare and contrast the two models on a host of their characteristics, including philosophies of causal inference, definitions of effect, theories of cause, external validity, matching, quantification, design versus analysis, and commensurability of language, ending with a discussion of some underdeveloped topics in both CCM and RCM.

Philosophies of Causal Inference

Both RCM and CCM are, if nothing else, centrally about causal inference. Consistent with the bandwidth versus fidelity theme in this article, however, CCM is more wide ranging in its appeal to the philosophical literature on causation, whereas RCM has a different focus on a more narrow but powerful formal statistical model.

CCM journeys widely through the epistemological literature, both descriptive and normative (Campbell, 1988; Shadish et al., 1991). This includes philosophy of causal inference (Cook & Campbell, 1986), but Campbell's epistemological credentials extend quite diversely into sociology of science, psychology of science, and general philosophy of science. In sociology of science, for example, Campbell weighed in on the merits of weak relativism as an approach to knowledge construction. In psychology of science, Campbell discussed the social psychology of tribal leadership in science. In philosophy of science, Campbell not only coined the term evolutionary epistemology (Campbell, 1974) but is widely seen as a significant contributor to that philosophical literature. Campbell thus brought an unusual wealth and depth of philosophical background to his discussions of the use of experiments as an approach to knowledge construction in both science and everyday life.

CCM ties its approach to causation more explicitly to the philosophical literature than is typically the case for RCM. Cook and Campbell (1979), for example, described their work as "de-

rived from Mill's inductivist canons, a modified version of Popper's falsificationism, and a functionalist analysis of why cause is important in human affairs" (Cook & Campbell, 1979, p. 1). From philosopher John Stuart Mill, CCM took the idea that identifying a causal relationship requires showing that cause preceded effect, that cause covaries with effect, and that alternative explanations for the relationship between cause and effect are implausible. Accordingly, CCM's threats to internal validity include ambiguous temporal precedence reflecting Mill's first requirement, and the remaining threats (history, maturation, selection, attrition, testing, instrumentation, regression to the mean) are examples of those alternative explanations to be eliminated. CCM also acknowledges the influence of Mill's methods of experimental inquiry (White, 2000). Such influences include that certain design features can be thought of as inspired by Mill's methods, that experimental methods are not so much about identifying causes as eliminating noncauses, that causal inference from observation is distinctly different in character from causal inference by intervention, and that methods of experimental inquiry must be tailored to what is already known and to what extant conditions allow the researcher to do.

From philosopher Karl Popper (Popper, 1959), CCM endorsed the crucial role of falsifying hypotheses, specifically advising the experimenter to gather data that force causal claims to compete with alternative explanations epitomized by the threats to validity. CCM tells researchers to design a study to avoid a problem in the first place but, if this is not possible, to gather data about third variables that might have caused the effect. Though the application of this falsificationist logic is uneven in practice, good examples exist (Duckart, 1998; Reynolds & West, 1987). However, CCM is not naïve about the possibilities for falsification, recognizing that researchers can easily create a cloud of doubt about a claim that a threat to validity exists. Hence, CCM is skeptical of the results of any single study, encouraging programs of research in which individual studies are imbued with different theoretical biases and, more important, inviting criticisms of studies by outside opponents who are often best situated to find the most compelling alternative explanations.

Finally, CCM situates itself in the human activity of making causal inferences as a general context for scientific causal inference. It ties the development of human understanding of causal inference to evolutionary pressures that rewarded those who could perceive macrolevel causes in their environments (e.g., large predators) and who recognized the value to human survival of manipulations such as starting fires or making weapons (Cook & Campbell, 1979). CCM stresses the influence of social and psychological factors on construction of scientific knowledge (Campbell, 1984, 1988). It is a deeply psychological theory of how scientists make causal inferences, especially in nonrandomized designs, with as much in common with the social psychology of Heider (1958) as with the philosophies of Mill and Popper (Cordray, 1986).

RCM less often connects with the formal philosophical literature. An exception, however, is the common reference to RCM as a counterfactual theory of causation (e.g., Dawid, 2000; Holland, 1986; Morgan & Winship, 2007; Winship & Morgan, 1999). A counterfactual is a condition that would occur if some part of the world were different than it really is. Under a counterfactual theory, causal statements are counterfactual statements. That is, the effect of a cause is the difference between what did happen to the

person who received the cause (the fact) and what would have happened to that person had they not received the cause (the counterfactual). Lewis (1973) credited the 18th-century Scottish philosopher David Hume with the first clear statement of a counterfactual theory of causation:

We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed. (Hume, 1748/1963, Section VII)

The last sentence is a counterfactual claim. Hume did not develop counterfactual causation, however, developing a more positivist analysis instead (Cook & Campbell, 1979). Other philosophers have developed counterfactual theories since then (Collins, Hall, & Paul, 2004).

However, Rubin (e.g., 2005) prefers not to characterize RCM as a counterfactual theory, preferring the potential outcomes conceptualization. He notes that all potential outcomes could, in principle, be observed until treatment is assigned, that some can be observed after that, but that counterfactuals can never be observed by definition. RCM depends on the possibility of observing any of the potential outcomes. However, researchers need not choose between potential outcomes and counterfactuals. The RCM potential outcomes analysis is not a philosophy of causation; rather, it is a formal statistical model. In this sense, RCM is a statistical model of effects that is consistent with a counterfactual philosophy of causation given that when potential outcomes are not observed in an experiment, they become counterfactuals.

RCM also has falsificationist features, but very weakly compared to CCM. Hidden bias analysis (Rosenbaum, 1991a) indicates how much bias due to unmeasured covariates would have to be present to change the point and confidence interval estimates of the effect, that is, to falsify a claimed effect—although the analysis does not tell whether hidden bias actually exists. Rosenbaum (1991a) described a study that seemed invulnerable to hidden biases causing assignment probabilities ranging from .09 to .91; the latter probabilities cover almost the full range of possible nonrandom assignments (random assignment into two conditions with equal sample sizes would use a true probability of .50, to put this into context). However, Rosenbaum (1991a) said that later research showed that an even larger bias probably existed. In addition, part of propensity score analysis is examination of balance over groups after propensity score adjustment, with the possibility that the researcher will conclude that causal inference is not possible in the data set without heroic assumptions, thus falsifying the claim that a causal inference can be tested well in this quasiexperiment. However, RCM's emphases on falsification are minor compared to those of CCM, which is centrally built around the concept.

However, for all the apparent differences between CCM and RCM in philosophy of causation, little real disagreement between CCM and RCM results. For example, they agree on matters such as the need for manipulable experimental causes and the fallibility of human judgment in causal inference. They might well also agree on matters where CCM speaks but RCM is silent. For instance, Cook and Campbell (1979) ended their discourse on causation with eight claims, such as "the effects in molar causal laws can be the results of multiple causes" (p. 33), and "dependable intermediate mediational units are involved in most strong molar laws" (p.

35). It seems unlikely that RCM would disagree with any of these claims, although it also seems unlikely that RCM would have generated those claims. Conversely, RCM writings on philosophy of causation are too sparse to generate much discord.

Perhaps the important difference is the greater emphasis on human and scientific fallibility in CCM. CCM is paradoxically skeptical about the possibility of doing the very things that CCM sometimes requires to generate good causal inferences. Humans are poor at making many kinds of causal judgments, prone to confirmation biases, blind to apparent falsifications, and lazy about both design and identifying alternative explanations. Yet, in the cases where a stronger design cannot be used or fails, CCM is remarkably dependent on those judgments for identifying threats, deciding whether they have been rendered implausible in any given case, and summing over them to reach a causal conclusion, especially in weaker nonrandomized experiments. CCM argues, therefore, that the responsibility for being critical lies more with the community of scholars than with any given researcher, especially with those whose interests would most lead them to find fault (Cook, 1985). Of course, CCM benefits from technical advances, like propensity score analysis, that reduce reliance on fallible judgments. Hence, CCM has turned attention to incorporating such developments (e.g., Luellen, Shadish, & Clark, 2005; Shadish, Luellen, & Clark, 2006).

RCM is self-critical but in a different way, with less of a sense of the fallibility that permeates scientists as human beings and more of a sense of the clarity of assumption and the diligent search for tests of those assumptions—the very technical advances that help remedy CCM's reliance on human judgment but limited to far fewer such judgments than CCM's wide-ranging scope. RCM, for example, emphasizes the importance of making and testing assumptions about whether a given data set can support a credible propensity score analysis. It is ironic that these senses of fallibility are, perhaps, the hardest feature of both RCM and CCM to transfer to general practice. Many are the researchers who have proudly proclaimed the use of a quasi-experimental design that Campbell would have found wanting ("my methodological recommendations have been over-cited and under-followed"; Campbell, 1994, p. 295), and many are the researchers who have used propensity score analysis with little attention to the plausibility of assumptions like strong ignorability.

Definition of Effect

A defining strength of RCM is its explicit definition of an effect. The fidelity of RCM is exactly aimed toward this point. By contrast, CCM never had an explicit conceptual definition of an effect until it adopted RCM's (Shadish et al., 2002)—a substantial deficiency given its focus on finding the effects of causes. Implicitly, CCM used a counterfactual definition of effect in exactly the sense that RCM eschews in favor of a potential outcomes definition. This can be seen most clearly in Campbell's (1975) article, "Degrees of Freedom and the Case Study," in which he treated the case of causal inference in the one-group pretest—posttest design. He allowed inference of an effect when confident prior knowledge exists about how the outcome variable behaves in the absence of the intervention. The latter is knowledge of the counterfactual.

Otherwise, CCM treated effects as differences between two facts, for example, what happened to the treatment group versus what happened to the control group or what happened before treatment compared to what happened after treatment. Yet, this is less a conceptual definition of what an effect should be than a computation of observed differences. The computation worked well in randomized experiments. For other designs, CCM said that the measured effect was valid to the extent the researcher was confident the quasi-experiment mimicked the ability of the randomized experiment to rule out plausible alternative explanations. In this sense, RCM and CCM both appeal to the randomized design as the standard. RCM did so by building a propensity score logic for observational studies grounded in what researchers know about regular designs (Rubin, 2004). CCM did it by noting that "backhandedly, threats to internal validity were, initially and implicitly, those for which random assignment did control" (Campbell, 1986, p. 68). CCM gets to the correct counterfactual if those threats are implausible but does not get there otherwise. In this sense, the threats to internal validity are possible counterfactuals, things that could have happened to the treatment units if they had not received treatment. However, they are not all possible counterfactuals, which neither CCM nor RCM has a way of knowing fully.

Incorporating RCM's definition of effect into CCM is probably not enough. Rubin would presumably advise incorporating the potential outcomes model into CCM's explication of the basic rationales for various experimental and quasi-experimental designs. For example, CCM offers several explanations for why random assignment works (Shadish et al., 2002, Chapter 8), all of which are partly true but all of which might flow better as consequences of a potential outcomes model. For example, CCM could present RCM's potential outcomes model and then show how the randomized experiment flows from the model as a way of estimating the average causal effect. Rubin (2005) has done the work for randomized experiments, and he and others have done the same for many nonrandomized experiments (e.g., Angrist & Lavy, 1999; Hahn, Todd, & Van der Klaauw, 2001; Rubin, 2004; Morgan & Winship, 2007; Winship & Morgan, 1999).

Theory of Cause

Just as CCM never formally defined an effect, RCM has paid relatively less attention to the concept of a cause, no doubt partly because its focus on estimating effects can succeed with only minimal attention to the nature of causes. Illustrative is "the definition of 'cause' is complex and challenging, but for empirical research, the idea of a causal effect of an agent or treatment seems more straightforward or practically useful" (Little & Rubin, 2000, p. 122). Both CCM and RCM want a clear operationalization of the intervention, have a preference for full implementation of the intervention, and when that cannot be done, want good measurement of the implementation of the intervention. Both CCM and RCM agree that causes in experiments must be manipulable and that the cause includes whatever was manipulated. If experiments really are about discovering the effects of manipulations and if one's theory of causal inference is limited to experimental demonstrations that measure the effect, perhaps no additional theory of cause is needed. However, even if it is not needed, a more

developed theory of cause can still be quite useful in understanding experimental results.

One reason that a theory of cause can be useful is the incompleteness of knowledge about the manipulated cause in an experiment. That knowledge is mostly about some of the actions someone took to manipulate the treatment. That turns out to be quite partial knowledge. CCM aspires to more, a fact reflected, in particular, in its interests in construct validity of the cause. For example, Campbell (1957) said participant reactivity to the experimental situation is a part of the treatment, and he emphasized that experimental treatments are always multidimensional packages consisting of many components: "The actual X in any one experiment is a specific combination of stimuli, all confounded for interpretive purposes, and only some relevant to the experimenter's intent and theory" (p. 309). Cook and Campbell (1979) focused extensively on the construct validity of the cause, elaborating both the theory of construct validity and the validity types that were previously more implicit. Subsequent work in CCM adopted Mackie's (1974) conception of cause as a constellation of features, with researchers often focusing on one feature, even though all may be necessary to produce an effect (Cook & Campbell, 1986; Shadish et al., 2002). CCM also emphasizes the need for programs of research to identify the nature of a putative cause. Any given experiment sheds only a little light on this nature, so its identification requires multiple studies on the same question in which different features of the causal package are changed until a more complete understanding of it emerges. Some features of a causal package will prove irrelevant to its effectiveness, and others will prove central.

To some degree, Campbell's interest in the nature of the cause occurred because he was a social psychologist. Experiments in social psychology place great stock in knowing the cause, because key arguments are often about whether an intervention reflects the causal constructs of interest in social psychological theory. Applied experiments are often less theoretically driven. Yet, even they need some theory of cause. For example, in the late 1990s, a team of researchers in Boston headed by the late Judah Folkman reported that a new drug called endostatin shrank tumors by limiting their blood supply (Folkman, 1996). Other respected researchers could not replicate the effect, even when using drugs shipped to them from Folkman's lab. Scientists eventually replicated the results after they traveled to Folkman's lab to learn how to properly manufacture, transport, store, and handle the drug and how to inject it in the right location at the right depth and angle. An observer called these contingencies the "in-our-hands" phenomenon, saying "even we don't know which details are important, so it might take you some time to work it out' (Rowe, 1999, p. 732). The effects of endostatin required it to be embedded in a set of conditions that were not even fully understood by the original investigators and may still not be understood today (Pollack, 2008).

A second reason a theory of cause can be useful concerns the status of causes that are not manipulable. Both RCM and CCM agree that nonmanipulable agents cannot be causes in an experiment. However, CCM affirms the potential causal status of nonmanipulable agents and so is comfortable entertaining hypotheses about, say, the cause role of genetics in phenylketonuria even if the pertinent genes cannot be manipulated (Shadish et al., 2002). RCM is less clear, leading to some debate about the implications of manipulability for RCM and for causal inference more generally (Berk, 2004; Holland, 1986; Reskin, 2003; Woodward, 2003). Morgan and Winship (2007) make two

points about this. First, RCM may simply not apply to causes that are not manipulable because the individual causal effect cannot be defined when the probability of a person being assigned to a condition is zero. Second, the counterfactual framework in RCM encourages thought experiments about nonmanipulable causes to clarify the nature of the causal question by specifying all the circumstances that must be considered in defining what the counterfactual might have been. Is the counterfactual for a person with phenylketonuria a person who is identical in all respects except the presence of the genetic defect, or does it include all the other ways that person might be different as a result of the genetic defect? For example, given that the main treatment for phenylketonuria is adherence to a particular diet, perhaps a person without the defect would have adopted a less healthy diet in general. This perspective is entirely consistent with how CCM would approach the matter, as illustrated in Campbell's (1975) rationale for how researchers can justify causal inferences in case studies. Still, overall the theory of cause in RCM is not extensively developed relative to CCM.

All this being said, it would be wrong to say the underdeveloped theory of cause in RCM is a flaw needing remedy. Rather, this difference is one more clue that CCM and RCM have different purposes. CCM aspires to be a theory of generalized causal inference, one that covers all aspects of the many kinds of inferences a researcher might make from all sorts of cause-probing studies. RCM has a more narrow purpose, to define an effect in a clear and precise way that allows that effect to be better measured in single experiments. It is the bandwidth versus fidelity issue. RCM and CCM overlap heavily on the issue of internal validity and less so elsewhere. Neither CCM nor RCM can function well without a theory of the effect, but in experiments, both could do without much theory of cause if internal validity (or effect estimation) were the only issue. However, it is not the only issue.

Causal Generalizations: External and Construct Validity

Another example of the relative breadth of CCM compared to RCM is the attention that CCM gives to causal generalization in the form of external and construct validity. Early works in CCM (e.g., Campbell, 1957; Campbell & Stanley, 1963) did little more than identify these generalizations as important ("the desideratum," Campbell & Stanley, 1966, p. 5), with the notable exception of the multitrait-multimethod matrix for studying construct validity (Campbell & Fiske, 1959). Cook and Campbell (1979) moved the theory of causal generalization forward considerably with their theory of construct validity of the treatment and outcome and by identifying alternatives to random sampling that might be used to facilitate generalization in experiments. Cook (1990, 1991) outlined still more theory and methods for studying such generalizations, forming the basis for what became three chapters on the topic in Shadish et al. (2002; Cook, 2004). Both theory and methods are wide ranging. Meta-analysis plays a key role in examining how an effect varies over the differences in persons, settings, treatments, and outcomes across studies. Also key is identification of causal explanations that provide the mediating mechanisms between cause and effect; such explanations provide knowledge that allows both better labeling and better transfer of the effect.

Rubin has contributed to meta-analysis and mediational modeling both conceptually (e.g., Rubin, 1990, 1992) and statistically (e.g., Frangakis & Rubin, 2002; Rosnow, Rosenthal, & Rubin, 2000) but rarely with an explicit conceptual tie to causal generalization. A notable exception is Rubin's (1990, 1992) work on response surface modeling in meta-analysis. It starts with the premise that the literature may contain few or no studies that exactly match the methodological and substantive question of interest. Rubin has suggested using the data in the literature to project results to an ideal study that may not be present in the literature but that would provide the empirical test one would like to have. This is one crucial form of external validity generalization. However, it has been little developed either statistically (Vanhonacker, 1996) or in application (Shadish, Matt, Navarro, & Phillips, 2000; Stanley & Jarrell, 1998).

In fact, neither RCM nor CCM has been particularly successful in generating applications of their respective ideas on causal generalization, perhaps because of the priority given to internal validity in much of scientific funding and thinking. An exception is that meta-analysis has seen increased use and funding over time. However, that increase has nothing in particular to do with RCM or CCM and seems to be focused mostly on getting a better estimate of the size of an effect than on exploration of how that effect varies over treatment, person, or study characteristics. Many meta-analyses include few or no tests of such moderators, reporting only an overall effect size for the intervention, as perusal of thousands of systematic reviews in the Cochrane Collaboration Library confirms. Knowledge about causal generalization seems to emerge organically as a literature on a topic develops and is occasionally reviewed. Researchers seem to feel little need for guidance in that process.

Quantification

Consistent with its more narrow statistical focus on effect estimation, RCM is more thoroughly and successfully quantitative than is CCM. Rubin is a statistician with a quantitative formulation of the potential outcomes model, who has paid careful attention to statistical assumptions, who has developed statistical tools, such as propensity scores and hidden bias analysis, to improve effect estimation, and whose model has generated a large number of quantitative developments, such as the use of instrumental variable analyses in the presence of partial treatment implementation (Angrist et al., 1996). Fruitfulness of quantitative developments is perhaps its most impressive feature.

At first glance, it may seem that CCM is not very quantitative, yet this is partially deceiving. Few of the many students coming out of the tradition were quantitative (Brewer & Collins, 1981; Campbell, 1988), but there were notable exceptions (e.g., Reichardt & Gollob, 1986; Trochim & Cappelleri, 1992). Campbell successfully collaborated with statisticians of his day to work on the analysis of regression discontinuity designs (Cook, 2007) and the multitrait—multimethod matrix (Boruch & Wolins, 1970), although the latter is not really much a part of CCM. Both lines of research continue to see activity today. CCM attracted the eye of other statisticians and economists (Meyer, 1995; Meyer, Viscusi, & Durbin, 1995), although that eye was often critical (e.g., Cronbach, 1982; Rogosa, 1980).

It may be more accurate to say that the Campbell tradition is quantitative on matters that are mostly peripheral to the core of

CCM-the validity typology, the use of threats to evaluate the accuracy of an effect estimate, and the notion that some design features can reduce plausibility of those threats. In principle, these omissions are at least partly remediable. For example, a great deal of work has already been done, both within and outside the CCM tradition, to quantify the effects of attrition on bias in randomized experiments (e.g., Delucchi, 1994; Shadish, Hu, Glaser, Kownacki, & Wong, 1998; Shih & Quan, 1997; Verbeke & Molenberghs, 2000; Yeaton, Wortman, & Langberg, 1983). Some work exists quantifying the effects of the testing threat to internal validity (Braver & Braver, 1988; Solomon, 1949). Reichardt (2000; Reichardt & Gollob, 1987) outlined ways that CCM could be reframed to be more quantitative. Winship and Morgan (1999) and Haviland, Nagin, and Rosenbaum (2007) discussed how to join an analysis to a design that includes multiple pretests to improve inferences in nonrandomized designs.

However, quantification of CCM would necessarily have its limits because many remaining issues pertaining to causal inference are more qualitative than quantitative. Regarding internal validity, for example, a good propensity score analysis has no way of detecting or adjusting for a history threat or a change in instrumentation that might have occurred during a time series (e.g., Neustrom & Norton, 1993), and it would have little to do with construct or external validity. Further, it could not detect cases in which the qualitative inference may differ from the quantitative inference, as when the quantitative analysis suggested an effective treatment but an ethnographer identified serious problems with it (Cook & Reichardt, 1979). Quantifications of these qualitative features into a grand quantitative theory of causal inference might be possible, but it is not on the horizon. In this sense, the fidelity of RCM necessarily means that it quantifies only a small part of causal inference, the part closely connected to the measurement of effects under different forms of selection. This is a keystone contribution, but it also points to the need for both fidelity and bandwidth in the larger endeavor.

Design and Analysis

Here, the differences between RCM and CCM are not accurately characterized as bandwidth versus fidelity. Both RCM and CCM prefer the use of strong designs, especially randomized experiments, whenever possible. However, with the notable exception of matching, RCM has focused more on analysis, whereas CCM has focused more on design. The regression discontinuity design best highlights this difference—the design was invented in the 1950s by Campbell (Thistlewaite & Campbell, 1960), but a statistical proof of its unbiased estimate was provided by Rubin (1977) in the 1970s (an earlier unpublished proof was provided by Goldberger, 1972). Similarly, Campbell's treatment of the interrupted time series quasi-experiment includes such variations as adding a control group, adding a nonequivalent dependent variable, and using removed and repeated treatments, negative treatments, and staggered implementation over sites, all aimed at improving confidence about any conclusions about effects that are drawn. By contrast, discussions of time series in the RCM tradition (e.g., Winship & Morgan, 1999) emphasize analysis almost exclusively.

In many respects, these complementary emphases on design and analysis need to be merged. Strong on design, CCM lacks the analytic sophistication and tools that move conclusions about confidence in effects from qualitative to quantitative for all the nonrandomized designs. Strong on analysis, RCM is often invoked and used without as much attention to design as CCM would like to see. Yet, good design makes for better analytic results, and better analyses can improve an otherwise low yield from a weaker design. This is why tools like propensity scores can help remedy the weakness in CCM of reliance on qualitative judgments by scientists who are frequently not very good at the kind of critical thinking required to identify and rule out alternative causal explanations. However, there is a parallel weakness in RCM. Once tools like propensity score adjustments are available, scientists are tempted to take the lazy way out of thorny design problems, hoping that they can fix by analysis what they bungled in design—which is ironic because RCM no doubt shares the same priority on good design emphasized in CCM.

Like most statistical approaches to causation, RCM seems to value design mostly when it contributes to a quantitative result, such as an effect size, a confidence interval, or a significance test. CCM shows how to make use of information even if it can be done only in a more qualitative fashion. For example, propensity score adjustments are designed to deal only with the selection threat to internal validity and perhaps also maturation and regression as special forms of selection bias. For those cases, it produces a quantitative adjustment that may improve inferences about treatment effects. However, propensity score analysis is not designed to deal with the internal validity threats of history, testing, or instrumentation, all of which can produce a spurious effect. CCM brings those concerns to the table in a way RCM does not. Yes, it would be better if they were quantified, but so far they are not in either CCM or RCM.

A notable exception to the general preference for analysis in RCM concerns matching. Both CCM and RCM attend carefully to the method of matching, reflecting its wide use as a seemingly plausible method for trying to create comparable groups when random assignment is not possible. Until recently, researchers in the CCM tradition were skeptical about matching in nonrandomized experiments, for example, saying "the two groups have an inevitable systematic difference on the factors determining the choice involved, a difference which no amount of matching can remove" (Campbell, 1957, p. 300). The best developed early analysis of this issue in CCM is from Campbell and Erlebacher (1970), who analyzed the pernicious effects of matching on the estimates of effects in the early Head Start program by Cicirelli and Associates (1969). The latter found that a matched group of children who did not receive Head Start performed better at outcome than the Head Start children. Campbell and Erlebacher showed how this could easily result from a combination of selection bias on the true underlying variables coupled with unreliability of measurement that caused the two groups to regress in different directions, creating an artifactual negative estimate of treatment effect. Reanalyses of these data by Magidson (1977) supported this interpretation. This led to early skepticism about matching within the CCM tradition. More recently, researchers using CCM have expressed support for matching on variables that are stable and reliable, for example, the use of achievement test scores that are aggregated to the school level and also aggregated over several years of pretest data (e.g., Millsap et al., 2000) or well-developed propensity scores.

RCM has brought new life to matching (and similar design strategies like stratification), especially propensity score matching (Rubin, 2006). Given that propensity scores are a composite of many covariates, such scores are likely to be more stable and reliable than individual covariates, converging in rationale with the CCM tradition. However, the kinds of stable matching variables that CCM eventually endorsed are not always common, whereas propensity scores are, in principle, much more widely available. Moreover, propensity score matching has clearer theoretical rationales for obtaining better effect estimates. In practice, results seem to be mixed (Dehejia & Wahba, 1999; Glazerman, Levy & Myers, 2003), though tests of this hypothesis by methods that are arguably better have been more optimistic (Luellen et al., 2005; Shadish, Clark, & Steiner, 2008; Shadish et al., 2006). Even the latter case, however, finds some sensitivity of propensity score adjustments to how missing data is handled in the covariates used to create propensity scores, which suggests that the relation of those covariates to treatment choice and outcome may be more important than the particular statistics used in propensity score creation and adjustment, and shows no particular advantage to propensity scores in reducing bias compared to ordinary regression. It seems crucial, therefore, to determine the conditions under which researchers can be reasonably sure that matching nearly always reduces rather than increases bias.

Commensurability of Language and Concepts

This article suggests the possibility of synthesizing RCM and CCM in ways that improve the causal inference enterprise, joining CCM's bandwidth with RCM's fidelity and giving heavy weight to design and analysis in both. However, some potential differences remain that need to be clarified as part of such a process. Kuhn (1962) suggested that it can be difficult to map the key ideas in one theory onto another, partly because they may use the same words differently, partly because they use different terminology, and partly because terminology is tied to other ideas within one theory that may or may not be present in alternatives. Kuhn called this incommensurability. Much of this article to this point can be thought of as working through the commensurability of ideas in the two models. To be clear, the present analysis does not identify any clear points of incommensurability. However, there are sufficient instances of possible problems to indicate that a dialogue would be useful to work through them.

For Kuhn, true incommensurability involves the use of either very different terminology, or identical or highly similar terms, in either case where the terms mean different things when embedded in two different theoretical systems. This is crucial because clear communication between the two systems may then never be fully possible. This is to be distinguished from a lack of clarity about a definition of a term that can be remedied more easily. Working through commensurability between RCM and CCM requires examining whether the very different terms used by RCM and CCM are truly incommensurable or are simply not understood definitionally. For example, one might think that RCM is primarily concerned with CCM's internal validity, in part because the RCM goal of getting an unbiased measure of effect is so similar to the priority given in CCM to internal validity and in part because RCM does not much mention ideas like construct validity or external validity. However, some researchers conversant in the RCM tradition explicitly invoke internal, construct, and external validity in discussions of the SUTVA. Speaking of experiments in which treatment is given to entire communities, Oakes (2004) said,

Causal effect estimates from trials may suffer from Hawthorne effects ... which are typically conceived as changes in behavior due simply to being the focus of attention. Such social placebo effects violate SUTVA and muddle the construct validity of causal effects. (p. 1943)

Assume for the sake of argument that this is a study in which communities are randomly assigned and that the study is properly analyzed. In CCM, if placebo effects occur, this would affect how researchers label the cause (the construct validity of the cause), but it would not preclude inferring that the treatment (properly relabeled) caused the observed mean difference between conditions (internal validity). Is it the case that placebo is a violation of SUTVA and that this violation is taken in RCM as precluding getting an estimate of the effect? If so, is this the same as saying the inference about a causal relationship no longer has internal validity?

Similarly, Dawid (2000) said that meeting SUTVA "allows one to generalize readily from the experiment to the target population, even in the face of differential treatment and selection criteria" (p. 414). The idea that a single experiment, randomized or not, could yield an estimate of effect that is both unbiased and generalizable is widely regarded in CCM as nearly impossible. In a randomized experiment, for example, the target population is almost always different from those willing to be randomized to conditions. Causal generalizations also concern samples of treatments, settings, and outcome measures about which researchers want to generalize. The fact that SUTVA is met seems unlikely to yield this plethora of generalization, but working through the interface between SUTVA and generalization in CCM's sense would help clarify this.

A third example is a claim by a reviewer of an article I published (Blitstein, Murray, Hannan, & Shadish, 2005). The case was a large group randomized trial of the effects of a healthy lifestyle intervention on adolescents within schools, with schools being randomly assigned to conditions. The data were properly analyzed with a random coefficients mixed model that is commonly considered to take into account the dependence of units within groups. Yet, the reviewer said that SUTVA was violated in this study because the nature of the intervention could be changed appreciably by the interaction among participants within schools, so that the effect of the intervention could not be estimated apart from that interaction. The reviewer asked that we prove that this violation of SUTVA did not bias the effect estimate. After several futile rounds of trying to respond, the article was accepted after we quoted an article by Rubin, Stuart, and Zanutto (2004) that recommended the group randomized design in a similar situation.

It is noteworthy that all these questions arise about SUTVA, which I previously noted is a difficult concept for many researchers to grasp. In the present case, we might ask what impact a violation of SUTVA has on CCM's validities. The answer is probably that it depends on the kind of violation. For example, a traditional nesting violation would be a question of statistical conclusion validity (violation of assumptions of statistical tests) but would not harm internal validity if nesting is properly taken into the account in the analysis. Experimenter expectancy effects

would for Campbell be a problem of construct validity (how the treatment is described) but not internal validity. It would help greatly to clarify the many variations of SUTVA violations and their implications.

Facilitating such dialogue is one purpose of a biannual conference on causality, which was first held in Seville, Spain, in 2004 and has been hosted subsequently in Jena, Germany. It began with a focus on comparing CCM and RCM, but in recent years it has expanded to include discussion of other approaches, such as Pearl (2000), Steyer (2005), Morgan and Winship (2007) and others and to include empirical studies on these issues (see http://www.metheval.uni-jena.de/konferenzen.php).

Gaps in RCM and CCM

These gaps are topics that, although they are in no sense incapable of treatment by both CCM and RCM, have not received much attention from either model. First, both RCM and CCM have mostly addressed causal inference in field rather than laboratory settings. Indeed, Campbell (1957) said CCM was developed to characterize and cope with the problems of causal inference that field researchers encountered, especially when randomized experiments could not be conducted. Similarly, to the best of my knowledge, RCM has never published an application of RCM to, say, basic psychological laboratory research. This is not because the models are irrelevant in that setting. Rather, it is because the models' signature contributions are to ameliorating the kinds of obstacles to causal inference about whether a treatment causes an effect that field researchers experience, such as effect size estimation in the face of attrition, partial treatment implementation, or the inability to control the assignment mechanism. However, to the extent that laboratory studies encounter such problems, both CCM and RCM are just as relevant.

First, a major exception to the preceding paragraph is construct validity. In field research, the focus is often on internal validity with both CCM and RCM. Laboratory researchers in psychology make extensive efforts to ensure that their treatments and measures reflect the important components of the theories that they are trying to represent. Methods textbooks aimed more at laboratory researchers in psychology (e.g., Rosenthal & Rosnow, 1991) routinely refer to CCM's threats to construct validity as a central component of laboratory research. Campbell and Fiske's (1959) work on the multitrait–multimethod matrix for assessing construct validity was, for a time at least, the most highly cited article ever published in *Psychological Bulletin*. Some recent work has attempted to tie a theory of causal effects to the multitrait–multimethod matrix (Pohl, Steyer, & Kraus, 2008).

Second, in part as a function of their focus on field settings, both RCM and CCM attend more to between-groups designs, such as randomized experiments, than they do to the kind of within-groups designs that tend to be used often in psychological laboratory work, such as reaction time experiments in which participants respond on a computer to stimuli that appear serially and vary factorially. Campbell and Stanley (1963) discussed within-groups designs (see their Designs 8 and 9), but such designs subsequently largely disappeared within most of the CCM tradition. The reason is probably that such designs are not widely used in field settings and tend to have different kinds of issues than field studies. From CCM, some threats to internal validity clearly apply, such as

testing effects; however, these are by now well known by laboratory researchers who have developed many means of coping with them. Similarly, Rubin has written about particular statistical issues in within-groups designs (Lo, Matthysse, Rubin, & Holzman, 2002), but that work does not flow from RCM in any explicit way (but see Pohl et al., 2008; Steyer, 2005).

Third, both RCM and CCM have focused primarily on simple descriptive causal inference about measuring the effects of treatment. The latter question is often of most interest in field settings like medicine, epidemiology, public health, and labor economics, where the answers are used to inform decisions about adopting policies or practices. They have both devoted significantly less attention to causal mediation about how effects occur. Examples of exceptions are Frangakis and Rubin's (2002) discussion of how RCM applies to mediational models and Cook's (1990, 1991) invocation of mediational modeling as a major key to causal generalization. These exceptions are comparatively minor foci in RCM and CCM, although others have pursued the issue in more detail (Holland, 1988; Sobel, 2008). A particularly interesting work related to RCM is the book by Pearl (2000), who analyzed the relationship between RCM and structural equation modeling. This article does not explore the relationship between Pearl's work on causal inference and either RCM or CCM. Pearl's work is extensive, not part of either the RCM or CCM traditions, and worth an extended review that is beyond the scope of this article.

Fourth, both RCM and CCM focus more on assessing the unknown effects of known causes than trying to find the unknown causes of known effects—case-control studies in epidemiology are an example of the latter. The reason may be that they base their treatment of nonrandomized designs on the desirable characteristics of the randomized design. Methods like case control depart more than most other nonrandomized designs from the logic of randomized experiments. However, both RCM and CCM have something to say about case-control designs. RCM has been far more successful at this than CCM, so much so that it may not be fair to call this a gap in RCM. (e.g., Holland & Rubin, 1988; Rosenbaum, 1991b). CCM treats case control designs more briefly (Potvin & Campbell, 1999; Shadish et al., 2002), mostly showing how threats to internal validity apply to causal inferences from case-control studies. Specific threats are highly contextualized, however, so Sackett (1979) developed a list of threats for casecontrol studies that differs entirely from the lists in CCM but better fits what such researchers know about the reasons why they can be wrong about the causal inferences they draw.

Discussion

If a theme emerges from this review of RCM and CCM, it is their difference in scope: CCM's bandwidth and RCM's fidelity. To describe CCM as wide ranging is to understate the matter. Encyclopedic seems more accurate. CCM seemingly has a capacity to include all conceptual, methodological, and statistical issues about field experiments within its scope, as witnessed by the ease with which CCM has or can incorporate RCM into its framework. Moreover, its scope includes essential features of generalized causal inference, such as the importance of external and construct validity, the many ways to improve the design of observational studies, and the human context of the scientific enterprise, that are not much covered by RCM but are crucial to the enterprise.

RCM is a concise statistical model about measuring causal effects. Within that narrow focus, RCM generates empirical implications with ease, unraveling a particular technical problem in a compelling manner that elicits assent. RCM is like antibiotics coming to dominate the treatment of infectious diseases in the 1940s, like television spreading throughout American households in the 1950s, and like the personal computer coming to be on desktops in offices and homes worldwide in recent decades. In each of these cases, the innovation became omnipresent within its host. In this sense, if CCM is the big tent, RCM is the current star of the show. It is the star because it has made a great deal of progress on some of the central problems that CCM identified but never really solved, such as how to generate accurate effect estimates from nonrandomized experiments or how to get a valid quantitative estimate of treatment on the treated in the presence of partial treatment implementation.

Indeed, when I first began this article, I wondered if CCM's time had passed. After all, the attention to RCM from mathematical statisticians (e.g., Rosenbaum, 2002), biostatisticians in medicine and public health (e.g., Kurth et al., 2005), economists (e.g., Angrist et al., 1996; Angrist & Lavy, 1999; Hahn et al., 2001), educators (e.g., Hong & Raudenbush, 2005), sociologists (e.g., Berk, 2004; Morgan & Winship, 2007), and even many within the CCM tradition (e.g., Reichardt, 2006; Shadish et al., 2008; West, Biesanz, & Pitts, 2000) is so very widespread and seemingly disconnected from CCM that questions arise about the latter's continued relevance. Perhaps it is time to teach RCM instead of CCM in undergraduate and graduate training programs.

On the one hand, such a critique has considerable merit. CCM is, at its core, a 50-year-old model, and that core has changed only a little over that time. Although that undoubtedly reflects a certain timeless character to the core of CCM, it may also reflect insufficient attention to developing CCM, especially empirically (Shadish, 2000). CCM in its various incarnations has sometimes been referred to as the bible of field experimentation in psychology and related fields. That honorific carries an unfortunate corollary, a tendency to treat biblical pronouncements as authoritative rather than as hypotheses and concepts that can be subject to empirical inquiry. There have been too few efforts, for example, to identify the key hypotheses of CCM and provide them with rigorous empirical tests or to translate the conceptual core of CCM into tools the field researcher can apply empirically rather than conceptually—though that has begun to change significantly in the past decade (e.g., Cook, Shadish, & Wong, 2008; Shadish et al., 2008). What comes out of such efforts is ultimately likely to be neither CCM nor RCM, but something new that improves on both of them while incorporating their best parts. Conversely, without such attention to critique, improve, and even replace key CCM concepts, the approach risks dying with its last living proponents.

On the other hand, that death is unlikely to be imminent for two reasons. First, the issues of generalization that CCM treats in detail are so many and so crucial that researchers cannot do without them. Second, in gathering references, I was struck by the continued widespread work on generalized causal inference that traces its inspiration at least partly to CCM (e.g., Boruch & Mosteller, 2002; Cook, 2007; McKnight, McKnight, Sidani, & Figueredo, 2007; Reichardt, 2006; Shadish & Rindskopf, 2007) and the continued high number of citations to the key works in CCM. Much of this work was and still is carried out by members of the core theory

group that Campbell attracted to Northwestern University (Shadish, Phillips, & Clark, 2003). However, CCM achieved paradigmatic status in part because it attracted so many researchers who were not themselves part of the Northwestern theory group (e.g., Bickman, 2000; Braver & Braver, 1988; Henry & Mark, 2003; Lipsey, 1990; Maxwell, 2003; McKillip, 1992; McKnight et al., 2007; Reynolds & West, 1987). That continues to be the case today.

It is, however, unfortunate for CCM that the discipline that CCM has long called home, psychology, is less dedicated to both widespread field experimentation and quantitative psychology than perhaps it used to be (Aiken, West, & Millsap, 2008). CCM thrived in part because the social psychology of the 1950s and 1960s made extensive use of field experiments. Today, the ability of CCM to develop the kinds of quantitative contributions represented by RCM is hampered by a general lack of attention to quantitative matters in faculty hires and program development in psychology. In this sense, even those working most directly in CCM find themselves focusing either on the few fields in psychology that welcome field experiments, such as health psychology, or on other fields entirely, such as education. So if CCM is in good epistemological shape, it is less clear that it is in good sociological condition. Only time will tell.

Finally, although this article no doubt sheds light on the relationship between RCM and CCM, it falls short of both a penetrating statistical critique and a vision about how causal inference might proceed beyond both CCM and RCM. This is due partly to my own limitations (I am trained as a psychologist rather than a statistician and my ability to critique the statistical aspects of RCM is limited), partly to my immersion in the work of CCM so that I may fail to see important critiques to be made, partly to the difficult hermeneutic task of interpreting the meaning of the corpus of work from such prolific authors, and partly to the fact that it is extremely difficult to improve on the ideas of scholars like Campbell and Rubin. They are, after all, the authors of the two greatest paradigms in field experimentation in the last half of the 20th century.

This is an exciting time to be involved in field experimentation. For a wide variety of reasons of which RCM is just one part, randomized and especially nonrandomized experiments seem poised for a renaissance both intellectually and practically (Shadish & Cook, 2009). Problems that chronically plagued randomized experiments, such as partial treatment implementation and low power cluster randomized designs, are increasingly productively addressed. Regression discontinuity designs are more widely used and better analyzed. Short interrupted time series have new analytic options when more than one series is available. Problems of selection bias in nonrandomized experiments show real signs of practical progress. An empirical program of research on all these matters is increasing researchers' understanding of the conditions under which the theoretical and methodological advice they dispense can work in practice. Demand for high-quality evidence about what works among policymakers and practitioners is at an all time high. All this and more bodes well for the future of field experimentation in general.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology: Replication and extension of

- Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, 63, 32–50.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American* Statistical Association, 91, 444–455.
- Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to identify the effects of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 533–575.
- Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.
- Bickman, L. (Ed.). (2000). Validity and social experimentation: Donald Campbell's legacy (Vol. 1 & 2). Thousand Oaks, CA: Sage.
- Blitstein, J. L., Murray, D. M., Hannan, P. J., & Shadish, W. R. (2005). Increasing the degrees of freedom in existing group randomized trials: The *df** approach. *Evaluation Review*, *29*, 241–267.
- Boruch, R. F., & Mosteller, F. (Eds.). (2002). Evidence matters: Randomized trials in education research. Washington, DC: Brookings Institution Press.
- Boruch, R. F., & Wolins, L. (1970). A procedure for estimation of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement*, 30, 547–574.
- Braver, M. C. W., & Braver, S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, 104, 150–154.
- Brewer, M. B., & Collins, B. E. (Eds.). (1981). Scientific inquiry and the social sciences: A volume in honor of Donald T. Campbell. San Francisco: Jossey-Bass.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.
- Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (Ed.), The philosophy of Karl Popper. La Salle, IL: Open Court.
- Campbell, D. T. (1975). "Degrees of freedom" and the case study. Comparative Political Studies, 8, 178–193.
- Campbell, D. T. (1984). Toward an epistemologically relevant sociology of science. Science, Technology, and Human Values, 10, 38–48.
- 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). Methodology and epistemology for social science: Selected papers. Chicago: University of Illinois Press.
- Campbell, D. T. (1994). Retrospective and prospective on program impact assessment. American Journal of Evaluation, 15, 291–298.
- Campbell, D. T., & Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The disadvantaged child: Vol. 3. Compensatory* education: A national debate (pp. 185–210). New York: Bruner/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., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.
- Chapin, F. S. (1932). The advantages of experimental sociology in the study of family group patterns. *Social Forces*, 11, 200-207.
- Chapin, F. S. (1947). Experimental designs in sociological research. New York: Harper.
- Cicirelli, V. G., & Associates. (1969). The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (Vol. 1 & 2). A report to the Office of Economic Opportunity. Athens, OH: Ohio University and Westinghouse Learning Corporation.

- Cochran, W. G. (1965). The planning of observational studies in human populations. *Journal of the Royal Statistical Society: Series A, 128*, 134–155.
- Collins, J., Hall, E., & Paul, L. (Eds.). (2004). Causation and counterfactuals. Cambridge, MA: MIT Press.
- Cook, T. D. (1985). Postpositivist critical multiplism. In L. Shotland & M. M. Mark (Eds.), Social science and social policy (pp. 21–62). Newbury Park, CA: Sage.
- Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (pp. 9–31; DHHS Publication No. [PHS] 90– 3454). Rockville MD: Department of Health and Human Services.
- Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter-century* (pp. 115–144). Chicago: National Society for the Study of Education.
- Cook, T. D. (2004). Causal generalization: How Campbell and Cronbach influenced my theoretical thinking on this topic, including in Shadish, Cook, and Campbell. In M. Alkin (Ed.), Evaluation roots: Tracing theorists' views and influences (pp. 88–113). Thousand Oaks, CA: Sage.
- Cook, T. D. (2007). "Waiting for life to arrive": A history of the regression—discontinuity design in psychology, statistics and economics. Evanston, IL: Institute for Policy Research, Northwestern University. Retrieved April 17, 2009, from http://www.northwestern.edu/ipr/publications/papers/2007/wp0703.pdf
- Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand-McNally.
- Cook, T. D., & Campbell, D. T. (1986). The causal assumptions of quasi-experimental practice. Synthese, 68, 141–180.
- Cook, T. D., & Reichardt, C. S. (Eds.). (1979). Qualitative and quantitative methods in evaluation research. Newbury Park, CA: Sage.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal* of Policy Analysis and Management. 27, 724–750.
- Cordray, D. W. (1986). Quasi-experimental analysis: A mixture of methods and judgment. In W. M. K. Trochim (Ed.), Advances in quasi-experimental design and analysis (pp. 9–27). San Francisco: Jossey-Bass.
- Cronbach, L. J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass.
- D'Agostino, R., Jr., & Rubin, D. B. (1999). Estimation and use of propensity scores with incomplete data. *Journal of the American Statistical Association*, 95, 749–759.
- Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95, 407–448.
- Dehejia, R., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.
- Delucchi, K. L. (1994). Methods for the analysis of binary outcome results in the presence of missing data. *Journal of Consulting and Clinical Psychology*, 62, 569–575.
- Duckart, J. P. (1998). An evaluation of the Baltimore Community Lead Education and Reduction Corp (CLEARCorps) Program. *Evaluation Review*, 22, 373–402.
- Fisher, R. A. (1925). Statistical methods for research workers (1st ed.). London: Oliver & Boyd.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503–513.
- Folkman, J. (1996). Fighting cancer by attacking its blood supply. *Scientific American*, 275, 150–154.
- Frangakis, C., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58, 21–29.

- Frangakis, C., Rubin, D. B., & Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and applications to advanced directive forms. *Biostatistics*, 3, 147–177.
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. Annals of the American Academy of Political and Social Science, 589, 63–93.
- Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations (Discussion Paper 123–72). Madison: University of Wisconsin, Institute for Research on Poverty.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression—discontinuity design. *Econometrica*, 69, 201–209.
- 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.
- Heider, F. (1958). The psychology of interpersonal relations. New York: Wilev.
- Henry, G., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. American Journal of Evaluation, 24, 293–314.
- Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945–970.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. Sociological Methodology, 18, 449–484.
- Holland, P. W., & Rubin, D. B. (1988). Causal inference in retrospective studies. Evaluation Review, 12, 203–231.
- Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27, 205–224.
- Hume, D. (1963). An enquiry concerning human understanding. LaSalle, IL: Open Court Press. (Original work published 1748)
- Kuhn, T. S. (1962). The structure of scientific revolutions. Chicago: University of Chicago Press.
- Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. (2005). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensitybased weighting under conditions of nonuniform effect. *American Jour*nal of Epidemiology, 163, 262–270.
- Lazarsfeld, P. F. (1948). The use of panels in social research. Proceedings of the American Philosophical Society, 92, 405–410.
- Lewis, D. (1973). Counterfactuals. Cambridge, MA: Blackwell.
- Light, R. J., Singer, J. D., & Willett, J. B. (1990). By design: Planning research in higher education. Cambridge, MA: Harvard University Press.
- Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research. Thousand Oaks, CA: Sage.
- Little, R. J. A., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytic approaches. *Annual Review of Public Health*, 21, 121–145.
- Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). New York: Wiley.
- Lo, Y., Matthysse, S., Rubin, D. B., & Holzman, P. S. (2002). Permutation tests for detecting and estimating mixtures in task performance within groups. Statistics in Medicine, 21, 1937–1953.
- Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. Evaluation Review, 29, 530–558.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, England: Oxford University Press.
- Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation. *Evaluation Quarterly*, 1, 399–420.
- Maxwell, S. E., & Delaney, H. D. (2003). Designing experiments and

- analyzing data: A model comparison approach (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- McCall, W. A. (1923). How to experiment in education. New York: Macmillan.
- McKillip, J. (1992). Research without control groups: A control construct design. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L., Heath, & E. Henderson (Eds.), *Methodological issues in applied psychology* (pp. 159–175). New York: Plenum.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* New York: Guilford.
- Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*. 13, 151–161.
- Meyer, B. D., Viscusi, W. K., & Durbin, D. L. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. American Economic Review, 85, 322–340.
- Millsap, M. A., Chase, A., Obeidallah, D., Perez-Smith, A. P., Brigham, N., & Johnston, K. (2000). Evaluation of Detroit's Comer Schools and Families Initiative: Final report. Cambridge, MA: Abt Associates.
- Morgan, S. L., & Winship, C. (2007). Counterfactuals and causal inference. Cambridge, England: Cambridge University Press.
- Neustrom, M. W., & Norton, W. M. (1993). The impact of drunk driving legislation in Louisiana. *Journal of Safety Research*, 24, 107–121.
- Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. *Statistical Science*, 5, 465–480. (Original work published 1923)
- Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. Social Science and Medicine, 58, 1929–1952.
- Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge, England: Cambridge University Press.
- Pohl, S., Steyer, R., & Kraus, K. (2008). Modeling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series* A, 171, 41–63.
- Pollack, A. (2008, January 16). Judah Folkman, Researcher, Dies at 74. New York Times. Retrieved April 19, 2009, from http://www.nytimes.com/2008/01/16/us/16folkman.html?_r=1&scp=1&sq= Judah%20Folkman%20&st=cse
- Popper, K. R. (1959). The logic of scientific discovery. New York: Basic Books.
- Potvin, L., & Campbell, D. T. (1999). The case-control method as a quasi-experimental design. In D. T. Campbell & M. J. Russo (Eds.), *Social experimentation* (pp. 363–374). Thousand Oaks, CA: Sage.
- Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), Research design: Donald Campbell's legacy (Vol. 2, 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.
- Reichardt, C. S., & Gollob, H. F. (1986). Satisfying the constraints of causal modeling. In W. M. K. Trochim (Ed.), Advances in quasiexperimental design and analysis (pp. 91–107). San Francisco: Jossey-Bass
- Reichardt, C. S., & Gollob, H. F. (1987). Taking uncertainty into account when estimating effects. In M. M. Mark & R. L. Shotland (Eds.), Multiple methods for program evaluation (pp. 7–22). San Francisco: Jossey-Bass.
- Reskin, B. F. (2003). Including mechanisms in our models of ascriptive inequality. *American Sociological Review*, 68, 1–21.
- Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11, 691–714.
- Rogosa, D. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88, 245–258.
- Rosenbaum, P. R. (1991a). Discussing hidden bias in observational studies. Annals of Internal Medicine, 115, 901–905.

- Rosenbaum, P. R. (1991b). Sensitivity analysis for matched case-control studies. *Biometrics*, 47, 87–100.
- Rosenbaum, P. R. (2002). Observational studies (2nd ed.). New York: Springer-Verlag.
- Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.
- Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect size estimation. *Psychological Science*, 11, 446– 453.
- Rowe, P. M. (1999). What is all the hullabaloo about endostatin? *Lancet*, 353, 732.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.
- Rubin, D. B. (1990). A new perspective. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 155–165). New York: Sage Foundation.
- Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, 17, 363–374.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Out*comes Research Methodology, 2, 169–188.
- Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29, 343–367.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100, 322–331.
- Rubin, D. B. (2006). Matched sampling for causal effects. New York: Cambridge University Press.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational* and Behavioral Statistics, 24, 103–116.
- Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal covariates. *Bi-ometrika*, 79, 797–809.
- Sackett, D. L. (1979). Bias in analytic research. Journal of Chronic Diseases, 32, 51–63.
- Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy* (pp. 13–35). Thousand Oaks, CA: Sage.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statis*tical Association, 103, 1334–1343.
- 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. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, 60, 607–629.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton Mifflin.
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). Foundations of program evaluation: Theories of practice. Newbury Park, CA: Sage.
- Shadish, W. R., & Fuller, S. (Eds.). (1994). The social psychology of science. New York: Guilford.

- 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., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experimentation. In R. R. Bootzin & P. McKnight (Eds.), Strengthening research methodology: Psychological measurement and evaluation (pp. 143–157). Washington, DC: American Psychological Association.
- 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.
- Shadish, W. R., Phillips, G., & Clark, M. H. (2003). Content and context: The impact of Campbell and Stanley. In R. J. Sternberg (Ed.), The anatomy of impact: What makes the great works of psychology great? (pp. 161–176). Washington, DC: American Psychological Association.
- Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. In G. Julnes & D. J. Rog (Eds.), Informing federal policies on evaluation method: Building the evidence base for method choice in government sponsored evaluation (pp. 95–109). San Francisco: Jossey-Bass.
- Shih, W. J., & Quan, H. (1997). Testing for treatment differences with dropouts present in clinical trials—A composite approach. Statistics in Medicine. 16, 1225–1239.
- Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behav*ioral Statistics, 33, 230–251.
- Solomon, R. L. (1949). An extension of control group design. Psychological Bulletin, 46, 137–150.
- Stanley, T. D., & Jarrell, S. B. (1998). Gender wage discrimination bias: A meta-regression analysis. *Journal of Human Resources*, 33, 947–973.
- Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. Methodology: European Journal of Research Methods in the Behavioral and Social Sciences, 1, 39–54.
- Thistlewaite, D. L., & Campbell, D. T. (1960). Regression—discontinuity analysis: Alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.
- Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials*, 13, 190–212.
- Vanhonacker, W. R. (1996). Meta-analysis and response surface extrapolation: A least squares approach. American Statistician, 50, 294–299.
- Verbeke, G., & Molenberghs, G. (2000). Linear mixed models for longitudinal data. New York: Springer.
- West, S. G., Biesanz, J., & 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 psychology* (pp. 40–84). New York: Cambridge University Press.
- White, P. A. (2000). Causal attribution and Mill's methods of experimental inquiry: Past, present and prospect. *British Journal of Social Psychol*ogy, 39, 429–447.
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. Annual Review of Psychology, 25, 659–706.
- Woodward, J. (2003). Making things happen: A theory of causal explanation. New York: Oxford University Press.
- Yeaton, W. H., Wortman, P. M., & Langberg, N. (1983). Differential attrition: Estimating the effect of crossovers on the evaluation of a medical technology. *Evaluation Review*, 7, 831–840.

Received January 31, 2008
Revision received September 25, 2008
Accepted October 7, 2008