ELSEVIER

Evaluation and Program Planning 27 (2004) 173–185

EVALUATION and PROGRAM PLANNING

www.elsevier.com/locate/evalprogplan

Book Review

Review of Experimental and Quasi-experimental Designs for Generalized Causal Inference
By W.R. Shadish, T.D. Cook, D.T. Campbell, 2002; Houghton-Mifflin, Boston

Will Shadish and Tom Cook, with the late Don Campbell, have written a book (the book and the set of the three authors are referred to hereafter by the authors' initials, SCC) that is in part a revision of two earlier books (Campbell & Stanley, 1966; Cook & Campbell, 1979), but is also a culmination of an entire tradition in social science methodology. Because of the role that this book will play in defining good research design, it will have an impact on how research and evaluation designs are taught, implemented, and judged. Writing a review of such a book is, of course, a challenge. On the one hand, no review is likely to do justice to the range of advances provided by the authors in this text. On the other hand, because this book will be received as a defining statement of our understanding of research design, it deserves a more involved critique than is typical for books reviewed for the evaluation community.

To convey some sense of the breadth of important issues addressed in SCC while still going into some of the areas in enough depth, I'll address selected points for three questions: the value of this book as a graduate-level text, its value in advancing the theory of causal inference, and its value in advancing the everyday practice of evaluation. Because this review is long and at times involved, I'll begin with an overview of the sections that follow so that readers can be selective with their attention.

In Section 1, we begin by acknowledging that the value of SCC as a text for a graduate methodology course will end up being a given—the book will be the foundation of many such courses. Of greater interest to those thinking about adopting the book for a course is how this book compares to the predecessor methods books by Campbell and colleagues. In brief, SCC is more self-consciously a stand-alone text for a research design course, and its writing is clearer for that purpose than its predecessors. However, the book is also filled with commentary on theoretical controversies that would likely require careful elaboration in the classroom to be of full value for most beginning graduate students.

Section 2, addressing the value of the book's contribution to the theory of causal inference, focuses on the reorganization of the typology of threats to validity. What's particularly interesting about the SCC contribution, part of which involves a tightening of what constitutes a threat to internal validity, is that it represents a further step in the increasing systematic nature of our understanding of the theory of causal inference. In Section 3, I argue that the main SCC contribution to research and evaluation practice may be its development and sanctioning of the pragmatic approach to research design. This pragmatic approach is clearest in the 'grounded theory of generalization' that the authors promote. Their grounded theory is not guided by the formal logic of statistical sampling that was presented as a necessary and sufficient foundation in the era of logical empiricism. Rather, the SCC approach begins with the heuristics of effective practice and makes this wisdom more systematic by distilling it into five principles to guide the generalization of causal inferences.

Building on these points, Section 4 of this review considers the strength of this text in supporting the future development of our understanding of experimental and quasi-experimental design. Our emphasis here will be on the continuing development of a more systematic theory of causal inference and a more pragmatic grounding for research and evaluation practice.

1. Value as Graduate Text: Changes from Previous Books

Given the broad adoption of the previous design books authored by Campbell and associates, the question of whether the book should be used as a text in graduate methodology courses is somewhat different than for most research design books. This book will be a definitive source on research design in the coming years, and so students of evaluation not exposed to it will soon be at a disadvantage. More interesting is the question of how this book is more, or less, appropriate than its predecessors as a graduate text.

1.1. Major Changes from Prior Campbell Books

In terms of title and focus, this book is most closely a revision of the Campbell and Stanley (1966) classic, *Experimental and Quasi-Experimental Designs for Research*. In this vein, the revision is dramatic, taking what was, after all, a chapter on research design for education and elaborating the themes into a coherent text

0149-7189/$ - see front matter
doi:10.1016/j.evalprogplan.2004.01.006

and reference book for all of social science research design. The change in title here reflects in part the current explicit focus on causal inference, a focus that helps explain and justify the Campbellian emphasis on internal validity. The title change reflects also the increased emphasis on a theory of generalization for causal conclusions.

But, of course, the current book is also a revisiting of topics developed in Cook and Campbell's (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. As indicated by differences in these titles, the current book covers a broader range of designs (not just quasi-experimental ones) and is less concerned with analysis issues. Accordingly, the structure of the book reflects the need to provide comprehensive coverage of design issues with fewer digressions from this central focus. This organization is both inevitable and desirable in what will, indeed, be the standard against which all other design books will be judged for the next decade. A major advantage of this focus is the detailed treatment of the major designs and of the 'design elements' that contribute to our confidence in the resulting conclusions. Chapters 4–8 of SCC provide the compilation of insights about the logic and use of different quasi-experimental and experimental designs that will make this text a required resource on design for years to come. Chapters 9 and 10 consider an array of practical issues, such as ethics, the management of treatment implementation, and attrition.

These changes in structure and focus contribute to the book's value as a research design text. Students who have been assigned chapters of SCC in my program evaluation course have offered positive assessments of the clarity of the writing in SCC. Furthermore, there has also been a moderated relationship involved in which the students with the better backgrounds in methodology were the more positive. Another aspect that many students seem to appreciate is the move away from a primary emphasis on the standard research designs and towards a focus on the 'design elements' noted above. This change in emphasis serves to avoid reliance on static designs and to strengthen appreciation of the elements of the designs, such as comparison groups or repeated measures, that support the general task of ruling out and ruling in alternative explanations for research findings. The result of this is to encourage students and others to be more flexible in developing research designs that are more useful for specific contexts. This emphasis on contextual analyses of threats and on design elements to address these threats may not be as comfortable for beginning graduate students, but it surely represents an advance in what we hope students will learn about research design.

Despite this overall virtue of the changes in SCC, one can be permitted some nostalgia for topics on analysis receiving less emphasis here than in Cook and Campbell (1979). For example, one of the more useful and practical chapters in Cook and Campbell, written by Chip Reichardt, dealt with statistical analysis for nonequivalent group designs. This chapter was practical in that it addressed a problem that haunts the vast majority of quantitative evaluations and useful because it offered reasonable solutions. Similarly, Mel Mark provided a valuable section in the Cook and Campbell chapter on passive observation that covered the analysis of patterns in time-series designs.

More generally, in becoming a book on research design, it is natural for the new book to place less emphasis on epistemology. As a result, there is no explicit mention of Campbell's 'evolutionary critical realism,' a position developed most fully in the first chapter of Cook and Campbell (1979) but also interwoven throughout that book. While this de-emphasis is understandable in that the first chapter of Cook and Campbell was likely off-putting to some, many others have gained much from that chapter. In particular, that chapter provided a frame of reference by summarizing differing perspectives on causality that, in turn, have differing implications for what should be the goal of good research or evaluation design. For example, the first chapter of Cook and Campbell differentiated theories of causality in terms of some being 'molar' (focusing on observed co-variation) and others being more 'molecular' (focusing on underlying mechanisms). This exposition was useful in that it set the stage for considering the value of different approaches to research design and evaluation, such as the controversy over the degree to which evaluators need to develop and make use of causal models of how a program yields its impacts.

1.2. Conclusions on Value as a Text

This book is the first in its Campbellian lineage to be structured as a course text on research methods. This is a good thing. It means that we have available a quantitative methodology text that is built upon the latest and most thorough conceptual foundations. The only caveat to those considering adoption is that the authors want this book also to be a definitive statement on the theory of causal inference, which means that there is considerable commentary about the various conceptual controversies in the field, including historical accounts of the debates between Campbell and Cronbach. These commentaries add much for advanced graduate students in evaluation who need to appreciate the evolution of our understanding of methodology related to causal inference. However, students who are new to methodology, or who are in fields where methodology is less emphasized, may view such accounts as digressions rather than as central to the task of mastering the fundamentals of research design. Similarly, the emphasis on design 'elements' rather than on traditional designs themselves will be welcomed by many but may frustrate those who might prefer to memorize lists of the strengths and weaknesses of standard designs in the abstract. But could we really want it otherwise from these authors?

2. Value of Contribution to Theory: Increasing Conceptual Organization of Theory of Valid Inference

The SCC book was not written only to provide an improved text on research design. As just noted, the authors were driven also by a desire to advance the theory of causal inference. This is particularly important to the evaluation community in that the theory of valid causal inference that predominates in our field will, over time, have an impact on how sponsors of evaluation, including government agencies, view the value of available methods.

How one judges the SCC contribution to this theory will likely be affected by how one feels about their effort to revisit and reorganize the various types of threats to valid inference. In brief, SCC follows the tradition of Cook and Campbell in dividing the threats into four categories (statistical conclusion validity, internal validity, external validity, and construct validity), but there are many changes, most small though some more significant, in how various threats are assigned to these four categories. For our purposes, the value of these specific adjustments is best viewed as part of a larger trend–the trend towards more systematic conceptualization of the threats to valid inference, from the simple lists of threats developed 50 years ago to a tight theory based on foundational concepts. As such, the short answer for the value of the SCC contribution to this trend is that the book offers a cleaner resolution to some of the past conceptual controversies but that this resolution should be viewed less as a final resolution (for theorists will not fail to criticize some of the finer points) and more as a summing up of current thinking to promote more constructive debate.

To elaborate a longer answer about this contribution to theory requires that we address an array of subtle issues that likely only theorists of methodology could love. To touch on these issues without making this review a longer monograph, we'll consider only a single example of the SCC reconceptualization of threats to valid inference, their revised distinction between internal validity and construct validity. Following this we'll address the general trend that this single example represents, the movement towards a more systematic account of the theory of valid casual inference.

2.1. Reframing Internal Validity: Contamination and Construct Validity

One of the notable changes found in the Cook and Campbell text was the addition of four threats to internal validity that, as a group, were referred to as contamination: diffusion or imitation of treatments, compensatory equalization of treatments, compensatory rivalry by respondents receiving less desirable treatments, and resentful demoralization. This addition was not the result of a major change in the definition of internal validity. To note the continuity, recall that for Campbell and Stanley (1966) internal validity concerned making false causal attributions to an intervention: "Did in fact the experimental treatment make a difference in this specific experimental instance?" (p. 5). The concern is that an ineffective treatment might mistakenly be concluded to be a cause of an observed effect. The Cook and Campbell definition had the same emphasis, defining internal validity as concerned with "[d]rawing false positive or false negative conclusions about causal hypotheses" (p. 80).

Tension in Definition of Internal Validity. Not based on changes in the basic theory, the additions introduced by Cook and Campbell were based on experience, on the recognition of additional ways that an ineffective treatment might appear effective or an effective treatment might appear ineffective. Consider, for example, the threat they labeled 'resentful demoralization.' Even if a treatment had no causal impact on desired outcomes (as defined in the sense of the ideal counterfactual, that in the absence of the treatment the subjects would have performed the same as they did with the treatment), it might appear effective if members of the comparison group became demoralized and performed worse than they would have in the absence of a treatment–control group study. As Cook and Campbell point out, in such cases "it would be quite wrong to attribute the difference to the planned treatment" (p. 55).

On the other hand, Cook and Campbell also developed the notion of construct validity as involving not just the relationship between operations and constructs (e.g. the degree to which a measure captures the construct intended or the degree to which an operation is justified as an exemplar of a construct) but also the disentangling of causal influences. As an example of the former, construct validity is the relevant concern when one questions whether an instructional approach is appropriately labeled as a 'child-centered' approach. In an example of the second aspect, if an innovative teaching approach labeled as 'child-centered' also has a nutrition component, there is a confounding of casual influences that might account for any improvements found among students involved in this innovative approach. "Construct validity is what experimental psychologists are concerned with when they worry about 'confounding'" (Cook & Campbell, 1979, p. 59).

Cook and Campbell offer a traditional example of confounding as when some medical patients are given pills and compared to those not receiving pills. The presumed active ingredient in the pills covaries, and so is confounded, with (1) the other ingredients in the pills, (2) the interaction with medical staff when receiving the pills, and (3) the perhaps self-fulfilling belief that one is receiving effective treatment. Standard medical research practice uses a double-blind methodology (everyone gets pills that are identical except with regard to the active drug being in the pills of the treatment group but not the control group; neither the patients nor the health professionals in contact with the patients know whether pills contain the presumed

active drug) to reduce this confounding. The point to emphasize is that in the Cook and Campbell (1979) conception, the double-blind procedure is used not to strengthen internal validity but rather to enhance construct validity (see also Campbell, 1986).

These two developments, the addition of internal validity threats such as resentful demoralization and the emphasis on construct validity as involving the teasing out of confounding relationships, created a certain tension by categorizing related (though not identical) problems in different ways. From one view, it is a matter of internal validity that a planned treatment was not responsible for an observed effect that was instead the result of resentful demoralization. From the other view, it is a matter of construct validity that a planned drug treatment was not responsible for an observed effect that was instead the result of beneficial social interaction or even a placebo effect.

Resolution in Terms of Total Package. SCC have modified their typology of validity threats to help resolve this tension. In defining internal validity, they note, "we use the term internal validity to refer to inferences about whether observed co-variation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured" (2002, p. 53). In referring to A causing B, the important point is that A refers to entire 'package' of all things that are the result of any manipulation of an intervention. Because the contamination threats that had previously been included under internal validity would not occur without the establishment of control or comparison groups, they are part of the package labeled A. Accordingly, the contamination threats have been reassigned to construct validity. Resentful demoralization, therefore, is now understood as something caused by the 'treatment' in the sense that it would not have occurred if the treatment manipulation had not occurred. As such, the SCC view achieves its internal consistency by viewing threats to internal validity not as sources of invalid causal conclusions about a planned treatment but as sources of invalid conclusions about the impact of the total intervention package, of which the planned treatment is only part. On the other hand, the consistency of construct validity is maintained by defining it as concerned with 'naming' (the meaning to be given to the variables measured). It therefore includes also the threats that arise in 'naming' which of the various components of a molar treatment package (such as the efficacious treatment or the demoralization of controls) are responsible for the observed impacts.

In sum, the potential of concluding that a treatment is successful based on unnaturally low control group scores due to resentment is no longer conceived as a threat to internal validity but as a threat to construct validity. To some, this change will require an adjustment in that it entails subtle changes in the relationship between internal validity and the counterfactual approach to estimating impacts of interventions. In the counterfactual view, the impact of a program is thought of as the difference between what happens to participants after an intervention and what *would have happened* if these same participants had not been exposed to the intervention. The biased impact estimate due to resentful demoralization would not occur if one had, instead, access to this ideal counterfactual of what would have happened to the experimental group without the intervention (which one never does). As such, from a counterfactual perspective, resentful demoralization leads to an invalid conclusion about causal impact in the same way that selection bias can, by creating a bad estimate of the ideal counterfactual.

2.2. Managing the Balance of Conceptual Organization and Empirical Openness

With this example of the SCC efforts to organize the threats to internal and construct validity in terms of a more systematic foundation, we can now consider the characteristics and value of a more systematic theory of causal inference. For this, it is useful to recall that our current theory of threats to valid inference evolved in fitful stages. Campbell (1986) provides some commentary on this evolution by noting that internal and external validity were originally used to differentiate the lists of threats that were addressed by random assignment (internal) and those that remained a concern even with random assignment (external). That this organization and labeling of threats is so different from our current conception is a testament to the contributions made by Campbell and colleagues.

What happened that changed the meaning of internal and external validity to what we understand today? A partial answer is that there has been a natural progression in our understanding of valid causal inference. There are three points to emphasize about this progression. First, this progression has been away from the initial, unorganized, lists of possible threats and towards a more systematic conceptual framework that organizes our understanding of these threats. Second, as SCC acknowledge, in the social sciences this progression is never to be complete, there is always a need to balance the virtues of a tight conceptual framework with the advantages of the openness to new understanding that is typical of empirically based lists. Third, because of the need for balancing the conceptual and the empirical, it is particularly important to revisit our guiding conceptual frameworks from time to time to ensure they remain open to experience-based revision.

Local Molar Causal Validity. Movement towards a more systematic framework is always based on some core organizing concept(s). Within the Campbellian paradigm, one core concept later became labeled 'local molar causal validity.' Campbell presented this label in 1986, but acknowledged that the phrase was cumbersome and unlikely to be accepted into general usage. Likewise, SCC describe this phrase on one page (p. 54) and promise not to rely on it in their text. Nonetheless, it is a central concept for

understanding the approach to organizing the threats to validity in the SCC system.

Unpacking the meaning of 'local molar causal validity' involves two dimensions of valid causal inference. For the first dimension, 'local' is contrasted with more global claims about causal relationships in other contexts. As elaborated below, Cronbach saw this move from local to global in two stages, first to targeted populations (of persons, settings, treatments, and outcomes) believed to be similar to those studied and then to populations believed to be dissimilar.

In the second dimension, 'molar' refers to claims about the treatment as a 'whole', or as a total package, with at least some components of the whole having a causal impact of interest on the aggregate group being studied. For example, using the standard drug treatment scenario, if people are randomly assigned to one of two groups to either receive or not receive a pill, and a lower incidence of heart problems is found with those taking the pill, then (assuming minimal attrition and using statistical inference to estimate and effectively rule out 'unhappy randomization') it is reasonable to conclude that *something* about the treatment package was indeed causal, that the total package of giving the pill caused a decrease in heart problems. What aspect of 'giving the pill' caused the positive health impact is unknown. It could have been one or more of the drugs in the pill, the regular contact with the nurse administering the pill, or even a placebo effect. Each of these explanations is viable because many components covaried together in the overall 'package' of the manipulated treatment. Molar causality is unconcerned with which of these mechanisms caused the observed effects, only that the molar package, as a whole, did indeed cause the observed aggregate differences between groups.

Defining internal validity in terms of local molar causal validity has both advantages and disadvantages. The main advantage is that it allows for something very important: in a randomized experiment, a superior finding for a treatment group (with negligible problems with subject attrition and such things as experimenter bias) allows for a strong knowledge claim that is as close to 'Truth' as we have (or have any right to desire) in our field. This advantage is much reduced in quasi-experimental designs. A major disadvantage of tying internal validity to local molar causal validity is that, as Cronbach (1982) pointed out, we almost always want to know more than what local molar causality conclusions, by themselves, have to offer. It is rarely sufficient to conclude merely that something about the intervention had a causal impact—in drug studies we want to know whether the drugs themselves, and not the interactions with medical staff, were causal. Similarly, in social science settings we would like to know that the planned treatment (e.g. training in more effective communication styles) 'caused' improvements in social skills, marital functioning, organizational efficiency, or whatever outcome is of interest. Claiming that our ambitions for knowledge are more modest represents the strategy that Reichardt (2000) refers to as 'relabeling' (discussed below).

Value of Evolving Conceptual Organization. Leaving aside the question of whether 'local molar causal validity' is a concept deserving of greater usage, the more important points for this review are the movement towards a more tightly organized theory of causal inference based on this core concept and the value of this movement. SCC rightfully note that all productive frameworks, theirs included, involve both a conceptual organization of what is known and an empirical list of useful things to think about: "The threats we present to each of the four validity types have been identified through a process that is partly conceptual and partly empirical" (p. 39). An important contribution of the SCC text is the clarity of the conceptual organization and the consideration of the implications. This organization into better defined conceptual categories represents a movement away from what was primarily a heuristic list of confounding factors and towards a conceptual framework.

To appreciate the trade-offs entailed by this movement towards systematic conceptions, it is useful to make reference to a classic distinction among theories. Pepper (1942) referred to theories based on empirically derived lists as 'dispersive' and theories based on conceptual organization as 'integrative.' Most taxonomies (whether of types of mammals or types of organizations) are empirically based, or dispersive in Pepper's terms, in the sense that particular categories are only loosely related to other catergories. The advantage of loose taxonomies is that it is a relatively simple matter to add new items to a list; the overarching theory does not need to be revised to accommodate such additions. As SCC point out, "empirically based threats can, should, and do change over time as experience indicates both the need for new threats and the obsolescence of former ones" (p. 39).

The advantage of an integrative conceptual framework is that it specifies similarities and relationships among the otherwise unordered threats to internal validity. As such, one disadvantage of empirical lists is that, without additional organization, they entail less—having a dispersive taxonomy that distinguishes three types of program participants does little to suggest the possibility of some fourth type. Integrative theories, such as those common in physics, have the opposite virtues. Identifying three types of subatomic particles might well suggest the existence and characteristics of a fourth type, but discovery of a fifth type might call the whole theory into question.

To illustrate these points, consider efforts to represent the theories of evaluation. One approach would be to review the many theories that have been proposed, cluster them into groups based on similarity on important characteristics, and present a list of types of evaluation theories. Such a list would have the advantage of being open to change as new types of theories are proposed and existing types are judged obsolete for some reason. However, by itself this list would

have few implications for strengthening theory or making it more relevant for practice.

A different approach is illustrated in a recent volume of *New Directions for Evaluation*, entitled "The practice–theory relationship in evaluation." In the lead chapter for that volume, Christie (2003) presents two dimensions of evaluation practice that were derived from multidimensional scaling of items that characterize practice. The practices of eight major evaluation theorists were then positioned on these two dimensions and compared to each other and to the positions of practicing evaluators. Locating these theorists in a two-dimensional space is an effort to offer a more systematic conceptual framework with which to represent the many theories. As such, it has the advantage of implications. For example, one finding was that all of the theorists included in the study provided higher ratings than the average for practitioners on the importance of including diverse stakeholders in multiple aspects of the evaluation. Perhaps practitioners have reasons to place less emphasis on stakeholder involvement than theorists do; perhaps there is a need for new theories that reflect the constraints confronted by practitioners; or perhaps it is more acceptable to practice management-oriented evaluation than it is to theorize about it. A potential disadvantage of this systematic effort, of course, is that the conceptual framework may be neglecting other important distinctions and so may inhibit progress towards more constructive frameworks. For example, House (2003) suggests that Christie's dimension of 'stakeholder involvement' requires an appreciation of different meanings of 'involvement.'

2.3. Conclusions on Value of Contribution to Theory

The SCC account of threats to valid causal inference is more systematic than previous accounts in this paradigm, with the domain of internal validity based on the core concept of local molar causal validity. There is value in moving forward with a more internally consistent framework, but it is useful also that the authors do recognize the importance of what Pepper (1942) identified as the need for balance between dispersive and integrative theories. As such, on the one hand, it is natural for our theory of causal inference to evolve into a more systematic conceptual framework. Through this process, the useful empirical lists will be absorbed into a tighter theory based on core concepts. On the other hand, we need to be on guard against having an eloquent theory diminish our openness to new understanding.

Because of this need for caution, it is important to revisit our integrative frameworks from time to time. This is a particular challenge for the development of the Campbellian paradigm. The deserved reverence accorded each of the major texts in the development of this view has led to a conservatism in which somewhat incidental features of the earlier works become formalized in later ones. For example, the inclusion of construct validity issues under external validity in Campbell and Stanley (1966) led to its linking with external validity in Cook and Campbell (1979).

The ongoing debates between Campbell and Cronbach did much to overcome some of this unintended conservatism and left a legacy to help others discern the core conceptual wisdom (evident in Shadish, Cook, & Leviton, 1991, as well as SCC). Given the current absence of constructive opposition from someone of Cronbach's stature and the potential for charges of heresy if lessor scholars had questioned some of the foundations of earlier efforts, it is particularly valuable that Shadish and Cook have continued the review and reorganization of the Campbellian theory of causal inference.

3. Value of Contribution to Evaluation Practice: Movement Towards a Pragmatic Stance

The contributions to theory described above are a bit abstract and of uncertain practical utility. After all, most consumers of evaluation want defensible judgments about program impacts and are unconcerned whether the factors limiting our confidence about the efficacy of a planned treatment are classified as part of internal or construct validity. In contrast, some of the contributions to practice offered by SCC, such as the theory of generalization, are of immediate practical value in that they address tasks required of every practicing evaluator.

For this section, we focus on the SCC contribution to the task of making judgments of the generalizability of study findings. The SCC approach builds on previous work by Cook (1991, 1993) that seeks a more practical, and yet more general, account of what is involved in supporting conclusions as being generalizable in some sense. Cook refers to this as a 'grounded' theory of generalization as it begins with an appreciation of the effective ways that we make judgments about generalization. To understand the nature of this contribution, we review the evolution of our understanding of external validity and consider what it means to promote a 'grounded' theory of generalization.

3.1. Development of External Validity

External validity was the counterpart concept to internal validity in the Campbell and Stanley framework. Whereas internal validity could be understood in terms of the strong knowledge claims of experiments with random assignment, external validity had to confront the problem of induction from a sample to a population from within a random selection perspective.

External Validity and Generalization. External validity is another domain in which the definitions have changed, as SCC point out: "In Cook and Campbell (1979), external validity referred only to inferences about how a causal

relationship would generalize to and across populations of persons and settings" (p. 38; though, see the Cook and Campbell definition as including generalizing to and across "persons, settings, and times," 1979, p. 71). Extending this to include also treatments and observations yields the current SCC definition: "External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes" (2002, p. 83).

SCC credit Cronbach (1982) with arguing for this broader definition, and so it is interesting to note that the SCC definition of external validity is a return to the definition provided by Campbell and Stanley: "To what populations [persons or other units], settings, treatment variables, and measurement variables can the effect be generalized?" (1966, p. 5). The Cook and Campbell justification for not including treatments and outcomes was that this kind of generalization is the domain of construct validity.

Despite this return to the Campbell and Stanley definition, there are some changes in how external validity is understood. In particular, the Campbell and Stanley emphasis on generalizing 'to' a population or whatever, suggests a search for the boundary conditions for which a conclusion is valid. In contrast, the SCC definition shares with Cook and Campbell (1979) the emphasis on generalizing 'across' populations. This emphasis goes beyond a concern with the boundary conditions for which an essentially similar effect can be found to a concern with the "extent to which a causal relationship holds over variations" (p. 83) in contextual and operational specifics. This associates external validity with a concern for moderated relationships across the dimensions of interest.

Two Problems for Generalization. This distinction between generalizing 'to' and generalizing 'across' populations parallels Cronbach's two problems for generalization: first, generalizing to target contexts and operations that are viewed as similar to the specific context that was studied, and second, generalizing to contexts and operations that are viewed as dissimilar to the specific context studied. The first problem relates to the interest in establishing boundary conditions, the second requires an openness to identifying important moderated relationships. SCC refer to Cronbach's two problems, but they define their two aspects of generalization somewhat differently. The first aspect is extrapolation, generalizing to other contexts and operations and understanding the moderated relationships involved (e.g. using a study of learning in middle school students to extrapolate to high school students). The second aspect is interpolation, predicting outcomes for those not studied but whose characteristics place them within the range of the sample that was studied (e.g. the study included students who were 11, 12, and 14 years old, and you want to generalize to 13 year old students).

3.2. Grounded Theory of Generalization

These changes in the definition of external validity prepared the way for SCC in developing a new theory of generalization. To understand why this development of a grounded theory is such a sweeping contribution, it is useful to emphasize what this theory is not: it is not based on the concept of drawing random samples from a population. This means that, by straying from a formal statistical model of generalizing from random sample to population, this theory is not based on a foundation that would have satisfied a logical positivist/empiricist approach to knowledge. Instead, this emphasis on grounded theory represents a further step towards a pragmatic stance in the theory of causal inference (Mark, Henry, & Julnes, 2000).

Characteristics of a Grounded Theory. The standard account and justification for generalization is based on sampling theory. It begins with the presumption that if a sample is representative of a population, we are justified in generalizing from the sample to the population. Random selection allows one to assume that a sample is fairly representative and to estimate the error that might result from the generalization. Cook (1993) recognized that the problem with this clean justification is that it is rarely applicable to the problems facing researchers and evaluators. Even in those cases that support random selection of persons or other subjects, it is generally more difficult by orders of magnitude to implement random selection of settings, treatments, or outcomes.

The general inapplicability of the sampling model of generalization highlights a common dilemma in social science methodology—having to choose between (a) an approach that has a strong logical or mathematical justification but is ill-suited for the task at hand and (b) an approach that is often applied to the desired task but cannot be justified based on formal logic or any other objective foundation. Cook's response was to strengthen this second option by developing a theory to guide actual practice by categorizing the features of effective efforts to generalize.

The grounded theory that Cook and now SCC have offered is based on an appreciation of what scientists and others do in real situations. To emphasize the effective practices in use over a particular logic is to take a more 'naturalistic' approach to methodology, where naturalistic refers to supporting our natural capacities rather than trying to replace them with formal methods. This position has been buttressed by both philosophy of science arguments (e.g. Harré, 1986; Putnam, 1995) and empirical research in cognitive science (e.g. Keil, 1996; Sperber, 1996).

Principles for Supporting Generalization Claims. While a grounded theory of generalization does not prescribe a set of strict rules for evaluating the generality of conclusions, Cook (1993) presented five "principles that scientists use in

making generalizations" (p. 353): Surface similarity, Ruling out irrelevancies, Making discriminations, Interpolation and Extrapolation, and Causal explanation. In order, these five refer to making generalization judgments based on how similar two situations seem to us (akin to face validity); generalizing based on observed irrelevancies of contextual factors (persons, settings, treatments, and outcomes) in determining outcomes; generalizing based on observations of how contextual factors are relevant in determining outcomes; using variations in the contexts studied to interpolate or extrapolate to contexts not studied; and generalizing based on an understanding of the underlying mechanisms involved.

As noted above, these principles are justified not in terms of some logical structure but rather as a listing of what scientists routinely do. In this sense, it is worth noting that the five principles are primarily a set of guidelines derived from experience. Accordingly, SCC's account of generalization can be viewed as the beginnings of a conceptual organization. In contrast with the more systematic treatments of internal validity that have been proposed, much of the SCC contribution to the practice of generalizing findings is at the level of an empirically derived list of activities to be supported.

3.3. Conclusions on Contributions to Practice

The grounded theory of generalization offered in SCC is a distinctive contribution to the practice of evaluation and research in that it is not an attempt to derive good practice from logical analysis (or from first principles) but rather is an effort to begin with good practice and derive organizing principles. Over time, while the SCC grounded theory will have some impact on current practice, it will be even more important in influencing how we justify our methods. What's interesting about this affirmation of the 'primacy of practice' (Putnam, 1995) is that it marks a return to the anti-formalism that characterized most of the American pragmatists of a century ago (Menand, 1997).

The pragmatic alternative refers not to the oversimplified theory of truth ascribed to pragmatism, that "a claim is true if it is useful to believe that claim" (SCC, p. 35; see Putnam, 1995, pp. 8–12, for a refutation of this straw-man account of a more complex position). Rather, the pragmatic alternative is based on the recognition of the practical decisions that research is to support. In developing a grounded theory of generalization, Cook, and now SCC, was acknowledging that a formal theory, such as the sampling theory of generalization, could illuminate important issues but that strict allegiance to the formal theory is rarely sufficient. Putnam (1995) describes the pragmatic anti-formalism in similar ways, where "the revolt against formalism is not a denial of the utility of formal models in certain contexts; but it manifests itself in a sustained critique of the idea that formal models…describe a condition to which rational thought either can or should aspire" (p. 63). This return to a pre-positivist paradigm cannot help but call attention to the hostility that the SCC grounded theory would have unleashed had it been presented 50 years ago.

4. Future Development

Having addressed the value of SCC as a graduate text, as a contribution to theory, and as a contribution to practice, it remains for us to consider the contributions of the SCC book as a vehicle for supporting future work in the areas of design and causal inference. This consideration requires an extrapolation that makes use of one other strength of the SCC offering. Continuing the tradition established by Cook and Campbell of confronting the assumptions and other foundations of their theory, SCC review the debates surrounding previous formulations of the Campbellian paradigm. Particularly useful is their account of the productive dialogue between Campbell and Cronbach. This examination is the focus of chapter 14 of SCC, and it serves both to highlight the intellectual honesty of the authors and to provide an indication of how future theories might evolve in the area of experimental and quasi-experimental design. While there are many themes to consider for the future development in this area, we'll follow the two issues addressed in the previous two sections of this review: the trend toward more systematic development of the theory of causal inference and the trend towards a pragmatic grounding for a theory of causal inference.

4.1. Continuing Trend Towards Systematic Conceptual Organization

Shadish (2000) notes and laments the lack of recent progress in developing our theories of research design, particularly with regard to the theory of quasi-experimentation. He observes that in the decades since Cook and Campbell there has been little "subsequent effort to systematize existing quasi-experimental theory" (p. 13). The SCC book is intended to be just such an effort, but this book will also contribute to theory by providing a touchstone to guide the work of others in the Campbellian paradigm. What might one expect, then, as the logic of their paradigm becomes even clearer and allows even greater conceptual organization of a theory of causal inference? Two things tend to happen as the conceptual organization of a paradigm advances. First, the structure of the domain is clarified with a better understanding of the basic elements in the theory and how they relate to each other. In this case, the elements in what was once a simple list become understood as members of more general categories. Second, the dynamics of the processes of

interest become understood in terms of underlying mechanisms. As examples of the first category, we consider efforts to organize the threats to internal validity (Mohr, 1995) and the strategies for addressing these threats (Reichardt, 2000) into superordinate categories. As an example of the second, we consider how the central Campbellian concept of 'local molar causal validity' is being elaborated into a hierarchical view of internal validity that is addressed at multiple levels.

Superordinate Categories for Internal Validity. The changing lists of threats to internal validity—from Campbell and Stanley to Cook and Campbell to the current text—have focused attention on how the various threats to internal validity relate to each other. This, in turn has led to suggestions that there are a few superordinate categories of internal validity threats. Superordinate in this sense means dividing up the many types of threats into a few, more general categories. Just what these superordinate categories might be can be debated, but several alternative schemas have been suggested.

For example, Larry Mohr (1995) has grouped threats to internal validity into four categories: history, selection, contamination, and spuriousness/time order. The first of these, history, refers to anything other than the treatment (and other than contamination, described below) that happened over the course of the study. As such, this superordinate validity category contains as subcategories several of the traditional Campbell and Stanley threats: history ('external events' for Mohr), testing, maturation, regression, and attrition. Selection as a threat refers to any differences that pre-exist between treatment and comparison groups that might be related to the outcomes of interest. Contamination as a threat overlaps with history in that something happened over time, in this case to affect, to make less pure, a planned comparison. Subcategories of contamination include the improper delivery of the treatment (such as compensatory equalization) and improper environmental controls (such as not insulating the treatment and comparison groups from each other, a failure of which can lead to resentful demoralization). The final basic category in Mohr's system is the time-order problem that is also a category in the SCC typology. The goal for Mohr in identifying these superordinate categories is not just the elegance the categories bring to the theory but also to clarify the strengths and weaknesses of alternative design models and their associated threats to internal validity.

Beginning with a similar goal but using different elements, Chip Reichardt (2000) also has offered a theory with superordinate categories relating to internal validity. Instead of focusing on the threats to internal validity, Reichardt has contributed a framework that organizes the many strategies for strengthening internal validity into a small set of three superordinate categories: elaboration, relabeling, and substitution. Elaboration is based on Sir Ronald Fisher's dictum, "Make your theories elaborate" (p. 94). The idea is that more elaborate predictions cannot be accounted for as easily by the standard threats to internal validity. One form of elaboration, 'competitive elaboration' (which he develops with a comprehensive analysis of the ways that we can make use of this strategy), is familiar to most researchers as an approach for adding comparisons or conducting additional analyses to force a competition of explanations. 'Noncompetitive elaboration,' the other form of elaboration, is less familiar and involves situations where the treatment hypothesis and the validity threat predict the same or similar outcome. Failure to find the predicted outcome argues against both the treatment hypothesis and the validity threat explanation. 'Relabeling' involves redefining your knowledge claim to incorporate a potential validity threat, as when concluding that "the combined effect of treatment and initial selection differences is X." Substitution makes use of alternative comparisons to a treatment group that reduce vulnerability to problematic selection threats.

Our purpose in considering the conceptual systems offered by Mohr and Reichardt is not to judge whether one or both of their systems is a proper grouping into superordinate categories but rather to illustrate the natural movement towards conceptual organization along these lines. The SCC text will encourage and support more such efforts, but it will also encourage more comprehensive efforts to think about what we want internal validity to mean within the larger enterprise of organizing the whole set of 'validities' that make up the Campbellian system.

Hierarchy of Levels of Valid Inference. As noted above, one of the notable conceptual changes in SCC was the reassignment of the contamination biases from internal to construct validity. Underlying this, and any, effort at revisiting the classification of threats to valid inference is the belief that proper classification matters, that there are advantages to realigning the categories. One justification is internal to the SCC system, to resolve the tension among concepts that was apparent in the Cook and Campbell text. A second justification is that a revised organization will highlight important similarities and distinctions. A primary distinction between internal validity and construct validity is that the former refers to causal factors that could occur without an intervention while the latter involves threats that only arise because of the intervention and the potential for confounding of the planned treatment with other factors that are part of the 'total package.' In this view, as Campbell (1986) put it, you add a placebo control to address construct validity, not internal validity.

And yet, as Campbell himself noted, this interpretation of internal validity has been controversial, to some extent because adding a placebo control seems so similar to other design elements that we include in quasi-experimental studies to rule out alternative explanations related to internal validity (e.g. controlling for exposure to pretests). Part of the difficulty in resolving this controversy comes from, as discussed above, the desire to remain faithful to the received

wisdom of the Campbellian paradigm. If we begin with the fourfold schema of internal validity, external validity, statistical conclusion validity, and construct validity and we want to tie internal validity to what Campbell called local molar causal validity, then it is to be expected that some additional problems in isolating causal influences will be assigned to construct validity, making construct validity a fairly full and diverse category.

One way to open up the space available for classifying threats is to give a more central role to the multiple levels of inference that we want to address in most studies. SCC acknowledge this hierarchical nature of 'molar causality' and lay the groundwork for considering the implications for internal validity:

"Of course, experiments can and should break down such molar packages into molecular parts that can be tested individually or against each other. But even those molecular parts are packages consisting of many components." (p. 54)

The prediction here is that as this insight is developed, the tension that led to restricting internal validity to the domain of local molar causal validity will be resolved and internal validity will be seen as the domain for all aspects of identifying and distinguishing causal factors. This is not inconsistent with the above-cited general definition offered by Cook and Campbell, "Drawing false positive or false negative conclusions about causal hypotheses is the essence of internal validity" (1979, p. 80). Specifically, internal validity will be understood at multiple, mutually supporting, levels, beginning with what Campbell labeled the molar level and with each subsequent level appearing molar relative to the components that comprise it (Julnes & Mark, 1998).

To illustrate these levels, consider an example in education reform where one might be concerned, at a molar level, with whether the adoption of a new curriculum in a school district caused an observed improvement in student achievement scores. The task for the researcher is to isolate some of the potential causal factors (e.g. divergent history) to support a conclusion about the causal hypothesis that the intervention itself has caused much of the observed improvement. Working at more molecular levels, researchers on the same study or on other studies would focus on isolating other, more molecular causal factors: (a) whether the improvement is due to the content of the new curriculum or to alternative, intervention-related causes, such as the decreased effort from comparison teachers who prefer to be using the new curriculum or a placebo effect among those using the curriculum; (b) which of the many components of the new curriculum (e.g. the focus on concepts or the regular individual attention provided by the teachers) appear most useful for increasing achievement scores; and (c) the cognitive factors responsible for the impact that the effective curriculum components have on the achievement tests.

Note the hierarchical nature of these levels of analysis, each appearing molar relative to the study of its components at a more molecular level. As a result, the basic tasks for evaluators and researchers are the same at each level. In all cases the researchers are isolating causal factors and are concerned with internal validity, and Reichardt's (2000) strategies remain equally appropriate. What changes as one moves from initial molar analyses to more molecular ones is not the strategies for isolating causal factors but only the level of mechanism being addressed. This movement from molar to molecular levels is an important aspect of research programs and represents what Mackie (1974, p. 73) referred to as the "progressive localization of a cause." Further, it is because internal validity is important at all levels that it is not as inherently trivial as Cronbach rightfully characterizes the most molar conclusions.

That individual studies can and often do address causality at multiple levels is an even stronger argument for grouping these activities under internal validity. Further, particularly with quasi-experiments, the work at more molecular levels is often conducted to support the more molar causal inferences. For example, pattern matching, promoted by Campbell as the underlying logic of research design, involves the active effort to collect data that are consistent with, or match, one causal explanation and inconsistent with the major competing explanations. Because more complex patterns are generally consistent with fewer plausible explanations, pattern matching follows from Sir Ronald Fisher's abovementioned dictum, "make your theories elaborate." The point for us is to recognize that this elaboration requires multiple levels of analysis. That is, pattern matching rarely, if ever, operates at only the molar level of the presumption that 'something about X is causing influences in Y' (note, SCC reject Cronbach's claim that this is the paradigmatic conclusion for the Campbellian approach). Rather, pattern matching involves presuming a molecular process and delineating the implications of such an underlying process. For example, if one wishes to strengthen conclusions about the impact of a new curriculum, the pattern matching approach would require making assumptions about the mechanisms responsible for the impact (such as the application of concepts to everyday experiences) and deriving implications about who is expected to benefit, what measures should reflect this benefit, and the conditions under which this benefit might be occur. One would then examine the fit between predictions of students's activities and outcomes and the data.

In sum, one does not add a placebo control simply to enhance construct validity but rather to address also internal validity on a more molecular level. Connecting the multiple levels of mechanisms that are prominent in most evaluation efforts is important in that many studies address more than one of the levels of interest. Furthermore, it is worth noting that this same hierarchical model applies to external validity as well. Similarity in aggregate, or molar, impact of an intervention across settings would be complemented by examinations of similarities in impacts at more molecular levels. As SCC recognize, this would involve determining

whether interventions impacted different subpopulations (e.g. ethnic or age groups) similarly across settings and even whether the mechanisms responsible for similar impacts were similar across settings.

4.2. Pragmatic Approach Extended

In addition to the support that the SCC book will provide for more systematic theories of research design, it will also encourage acceptance of a pragmatic alternative to the strict adherence to the 'logic of science' typically presented in introductory methodology texts. In particular, the precedent set for justifying method choices for generalization based on a pragmatic framework will influence justification of methods in other areas of research design and analysis. We consider the justification that follows from viewing our methods as tools and then consider the implications of this for informing policies on what constitutes a 'valid methodology' in social science research and evaluation. This implication for policy is a timely contribution by SCC given the current debates in government agencies and elsewhere on what constitutes a 'valid methodology.'

Methods and Methodologies as Tools. In lamenting the recent lack of development in the theory of evaluation, Shadish (2000) places the blame on our neglect of an empirical program to examine which designs are most effective under particular conditions. To argue that guidance on methodology should be driven at least as much by experience as by logic is a reflection of the 'primacy of practice' that informs most pragmatic approaches. It is also a reflection of the pragmatic view that debate about the most useful designs should be seen as debate about the best, or most useful, tools. This focus on the instrumental value of methods and methodologies is consistent also with the 'assisted performance' concept promoted in the neo-Vygotskian approach to developmental psychology (Tharp & Gallimore, 1988). This approach emphasizes that our conceptual frameworks assist us in seeing and responding to important patterns in our world. Further, the tool's usefulness is dependent on the task at hand, consistent with Shadish's interest in exploring the value of different designs under different circumstances.

To illustrate this view of valuing methods as tools (rather than as manifestations of formal logic), consider two opposing approaches to using statistical inverse inference in social science research. The Neyman-Pearson paradigm is often described as a 'forward-looking' approach in that you establish your decision rules before conducting the study (e.g. "I'll reject the null hypothesis if $p < 0.05$"). In contrast, Sir Ronald Fisher promoted a 'backward-looking' approach in which you gather the data first and then 'look back' at the data and try to make sense of them (i.e. you collect the data before settling on decision rules). While the Neyman-Pearson paradigm (with its hypothesis testing and confidence intervals) has been judged by social scientists as the more logically consistent and more objective paradigm, this judgment is not universal: for the past 50 years the Neyman-Pearson paradigm "merely dominated other views among statisticians; but it utterly overwhelmed other views among those whose interest in statistics was primarily practical" (Kyburg, 1974, p. 22; see also Seidenfeld, 1992, for another example of statisticians attempting to rehabilitate the Fisherian paradigm).

Gigerenzer (1993) contends that the Fisher paradigm (with its significance testing and fiducial estimation) is actually more consistent with the tasks that we confront in research. This claim is based in part on the view that the actual practice of most research deviates so much from the idealized version that the virtues of the Neyman-Pearson approach are illusory. But the claim also reflects the view that researchers should operate somewhat at odds with what the Neyman-Pearson paradigm prescribes. Using a psychodynamic metaphor, Gigerenzer refers to the Neyman-Pearson approach as a 'superego' approach in the sense that it is based on formal, unbending rules that offer defensible actions. The Fisher paradigm is an 'ego' approach in that it is not as easily defended based on formal, exacting rules but is more conducive to making sense of complexity. The implication is that acting consistent with the demands of the superego approach may make us feel proper and respectable, but it is the ego approach that allows us to negotiate the everyday demands of life.

Informing policies on 'valid methodologies.' The points made above, including methods understood as tools, suggest a move away from reaching general conclusions about the value of methods across contexts. This movement is particularly relevant now as there are debates within and between federal agencies over what constitutes a 'valid methodology.' Evaluators and researchers may argue that validity is more appropriate for describing conclusions than for describing methods, but there is no denying the desire for clear answers about the relative value of different designs. One manifestation of this desire for valid methods is the current emphasis on using random assignment experiments (as reflected by the instructions to grant review panels to assign bonus points to proposals incorporating this design). As might be expected given the hostilities shown in the recent quantitative–qualitative paradigm war, many evaluators are opposed to this emphasis on and valuing of experimental designs.

What insights might the pragmatic approach suggest for informing this debate? First, extreme positions can be ruled out. It will not do to claim that there is one best method, such as random assignment, that should be used in every situation. We have too much experience with the limitations of experiments and other methods to argue for only one methodology (Sechrest, 2003). Also, it will not

satisfy government audiences to say only that there are no general rules, that all traditional designs could be most valuable in certain contexts. Even without complete consensus, some designs are recognized as being better than others in addressing specified needs (Tharp & Gallimore, 1982).

Alternative to these extremes, we need an intermediate approach that is sensitive to a variety of contextual factors. One way in which a specific context might call for designs different from what a formal analysis might recommend is when people working in an area have reasons to be particularly concerned with specific threats to internal validity. Addressing those specific threats might call for certain designs (e.g. being primarily concerned with the threat of 'history' and selecting a multiple time-series design). Alternatively, specific threats might be most relevant because previous research has led us to discount other threats. In addition, there is a larger issue in recognizing that designs that are effective in some of the sciences may not be as effective in other sciences or fields (Datta, 2003; Julnes, in press). For example, many of the experiments in astronomy involve no manipulation of traditional independent variables, and yet they produce results that provide compelling support for one theory over another. If such a 'design' were used in social science, the results might appear hopelessly ambiguous due to multiple viable explanations. All of this suggests the need for Shadish, Cook, and others to develop pragmatic principles for other areas of design that parallel those offered by SCC for generalization.

4.3. Conclusions about Future Trends

As our understanding of valid causal inference advances, our conceptual frameworks will become both more systematic and more pragmatic. The movement towards greater conceptual organization will include refined organization of threats to validity and our strategies for addressing those threats, and also a hierarchical view in which internal validity, as well as other types of threats, are understood as having meaning at multiple levels.

As for a pragmatic approach, if the grounded theory approach of SCC is accepted as the mainstream position on methodology, what impact will SCC have in inspiring the next generation of pragmatic theories of methodology? This question is a speculative one but worth considering. A touchstone for theorists working in this vein will be Cook's statement of how he developed the grounded theory of generalization that was elaborated in SCC: "I have had to explicate what random selection achieves in order to explore alternative (and messier) ways of bringing about the same ends" (1993, p. 77). That is, it is by explicating the general functions that our methods are meant to serve that we are able to transcend the specific prescriptions and see the more general requirements of effective methods. The SCC text contributes to this explication and will support others engaged in this task.

Acknowledgements

Professors Tom Cook, Leslie Cooksy, Chip Reichardt, and will shadish provided helpful comments on an earlier draft.

References

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*. San Francisco: Jossey-Bass (New Directions for Program Evaluation, No. 31, pp. 67–77).

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

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice–theory relationship in evaluation*. San Francisco: Jossey-Bass (New Directions for Evaluation, No. 97, pp. 7–35).

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. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest, & A. G. Scott (Eds.), *Understanding causes and generalizing about them*. San Francisco: Jossey-Bass (New Directions for Program Evaluation, No. 57, pp. 39–82).

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

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Datta, L-e (2003, Nov.). The politics of methodology: Speculations on why a hard science agency may use softer, gentler approaches while an wgency offering human services embraces maximum strength performance measures and experimental designs. Paper presented at the annual conference of the American Evaluation Association, Reno, NV.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

Harré, R. (1986). *Varieties of realism*. Oxford, UK: Blackwell.

House, E. R. (2003). In C. A. Christie (Ed.), *The practice–theory relationship in evaluation*. San Francisco: Jossey-Bass (New Directions for Evaluation, No. 97, pp. 53–56).

Julnes, G. Experiment, overview. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. San Diego, CA: Academic Press, in press.

Julnes, G., & Mark, M. (1998). Evaluation as sensemaking: knowledge construction in a realist world. In G. Henry, G. Julnes, & M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice*. San Francisco: Jossey-Bass (New Directions for Evaluation, no. 78, pp. 33–52).

Keil, F. C. (1996). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 234–263).

Kyburg, H. E., Jr. (1974). *The logical foundations of statistical inference*. Dordrecht: Reidel.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Clarendon.

Mark, M., Henry, G., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey-Bass.

Menand, L. (1997). *Pragmatism: A reader*. New York: Random House.

Mohr, L. B. (1995). *Impact analysis for program evaluation* (2nd ed.). Thousand Oaks, CA: Sage.

Pepper, S. C. (1942). *World hypotheses: A study in evidence*. Berkeley: University of California.

Putnam, H. (1995). *Pragmatism*. Oxford, UK: Blackwell.

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

Sechrest, L (2003, Nov.). Complementary and alternative methodology. Paper presented at the annual conference of the American Evaluation Association, Reno, NV.

Seidenfeld, T. (1992). R.A. Fisher's fiducial argument and Bayes' theorem. *Statistical Science*, *7*, 358–368.

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Research design* (pp. 13–35). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Sperber, D. (1996). Introduction. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. xv–xx).

Tharp, R. G., & Gallimore, R. (1982). Inquiry process in program development. *Journal of Community Psychology*, *10*(2), 103–118.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching and learning in social context*. New York: Cambridge University.

G. Julnes*
Department of Psychology, Utah State University,
2810 OldMain Hill EDUC 484,
Logan, UT 84322-2810, USA
E-mail address: gjulnes@cc.usu.edu

* Tel.: +1-435-797-1633; fax: +1-435-797-1448.