The Truth About Validity

William R. Shadish

Abstract

The chapters in Advancing Validity in Outcome Evaluation: Theory and Practice, like the literature on validity in general, are extremely diverse. They range from a narrow focus on particular criticisms of validity in one recent work in the Campbell tradition to very broad overviews of methods for improving the generalizability of evaluation results. The author reviews and comments on each chapter, and discusses some general themes prompted by this issue of New Directions for Evaluation. Those themes include the ambivalent treatment of randomized experiments within the American Evaluation Association (AEA), the need for fresh ideas about outcome evaluations from outside the AEA community, and the desirability of an empirical program of evaluation theory in which data play a central role in our validity recommendations. © Wiley Periodicals, Inc., and the American Evaluation Association.

I hope that the reader will forgive the cheekiness of the title of this commentary. It is meant to be at once provocative and challenging, self-deprecating and tongue-in-cheek (do I really think I have access to “the truth”?)), yet also referential to the epistemological and ontological stances I find most congenial in their weak form. What I am really doing, of course, is speaking from my own little corner of the validity universe.

There is much to like in the chapters in Advancing Validity in Outcome Evaluation: Theory and Practice. They tackle ambitious and diverse intellectual agendas on an important contemporary question to the field of evaluation,
validity in outcome evaluation. Yet that very diversity makes it difficult to write a coherent commentary. Indeed, a thoughtful review of just one of these chapters could be a chapter in itself. Given that limitation, I will first begin with observations about each of the chapters seriatim, and then try to draw together some resulting general thoughts.

Chen, Donaldson, and Mark
These authors set the stage for the issue by defining the theme and briefly reviewing the chapters (Chen, Donaldson & Mark, this issue). They note how this issue evolved from an initial focus on Donald Campbell’s validity typology to its final and far more general focus on validity in outcome evaluation theory and practice. Some chapters still focus entirely on Campbell’s validity typology, others make only passing reference to Campbell but do talk about validity, and still others talk about other features of good outcome evaluation but try to fit them into validity. The latter, in particular, raises a question. Validity may be essential to good outcome evaluation, but it is not clear that everything essential to good outcome evaluation needs to be placed into a validity typology. Imagine, for example, that this issue had evolved so that the theme was not validity but good outcome evaluation. Who could argue against, for example, the proposition that good outcome evaluation ought to consider not only accuracy but also utility, propriety, and feasibility? Or against the Chen and Garbe system integration issues? But are all these issues the province of validity? I think not, for then validity just becomes a catchall phrase for good, and we evaluators should know better than anyone that the good and the true are two different things.

Chen et al. also introduce an issue that is itself minor in some respects but that subsequent authors (e.g., Julnes, this issue; Mark, this issue; Reichardt, this issue) repeat and discuss, that the distinction between construct and external validity has been lost in recent works in the Campbell tradition. The issue pertains to an alleged change in the definition of construct and external validity from Cook and Campbell (1979; henceforth CC) to Shadish, Cook, and Campbell (2002; henceforth, SCC). It is not, from my perspective, true that the definition has changed. The prompt for this allegation seems to lie in the fact that CC mostly limited their discussion of construct validity to treatments and outcomes, whereas SCC provide more extended discussion of construct validity of persons and settings as well, and CC mostly limited their discussion of external validity to persons and settings, whereas SCC provide more extended discussion of external validity of treatments and outcomes, as well.

This expansion of application does not mean the definition has changed, in two respects. First, both CC and SCC use nearly identical language to define the fundamental idea of construct validity. SCC defines it as the “validity of inferences about the higher order constructs that represent sampling particulars” (p. 38); and CC defines it as the “validity with which we can make generalizations about higher-order constructs from research
operations” (p. 38). Both references to higher-order constructs representing sampling particulars capture the essence of the original meaning of construct validity in Cronbach and Meehl (1955). That original meaning did not limit construct validity to treatments and outcomes, no surprise given that Cronbach and Meehl were writing in the context of test theory.

Second, given that CC focused more on the construct validity of treatments and outcomes, readers often overlook that they also said “construct validity concerns are not limited to cause and effect constructs. All aspects of the research require naming samples in generalizable terms, including samples of people and settings” (p. 59). The latter suggests that CC's choice to give only passing note to the construct validity of people and settings is not a denial of the intellectual merits of its inclusion. The more complete and systematic inclusion of people and settings in SCC was motivated by formalizing and systematizing what was already present in CC. It is not new, and it is not a changed definition.

Gargani and Donaldson

The authors Gargani and Donaldson rightly point to the need for more attention to external validity broadly construed. Their advice about factors that increase the use of evaluations is sensible, especially their specific recommendations for involving stakeholders in use—though it would have been useful to contrast how those recommendations differ in important ways from the kinds of recommendations that have been salient in the evaluation utilization literature for decades.

Gargani and Donaldson (this issue) suggest “... the discussion of validity as it relates to outcome evaluation seems to be focused largely on questions of internal validity (Did it work?) with less emphasis on external validity (will it work?)” (p. 17). They provide no empirical evidence that this is the case. My impression is just the opposite. Outside the evaluation literature, for example, consider the medical literature on evidence-based practice. We know so much about how to assess what works that matters related to internal validity receive very little attention. By contrast, discussion of how such research translates into clinical practice is rampant. For instance, I subscribe to a list serve that sends me an average of about 10 journal articles per week about evidence-based practice. The October 11–15, 2010 article alert referenced 29 such articles, virtually all of which were reasonably construed as external validity (see http://www.citeulike.org/user/ SRCMethodsLibrary). Even within the confines of the American Evaluation Association (AEA), I would venture the guess that a review of American Journal of Evaluation, New Directions for Evaluation, or the program for AEAs annual conference would find more discussion of external validity than internal validity.

It is, however, interesting to note that the work of authors like Campbell (e.g., Campbell & Stanley, 1963), Cook (e.g., Cook and Campbell, 1979), and Rubin (e.g., Rubin, 1974) on descriptive causal inference is
wildly popular and highly cited, whereas those same authors’ work on things related to external validity (e.g., Cook’s five principles of causal generalization, and Rubin’s advocacy of response surface modeling in meta-analysis) remains nearly entirely unused and unreferenced by practicing evaluators. Why is that? I think it is because evaluators who are designing an evaluation are intuitively aware of the contextual embeddedness of their work, and of the virtual impossibility of anticipating future uses beyond those specified by the stakeholders to the evaluand. In that sense, Gargani and Donaldson are right to place the responsibility of use on stakeholders.

Mark

Mark’s (this issue) chapter is also concerned with generalizability, and he provides a thoughtful set of recommendations for future practice and development on that topic. Mark focuses more than Gargani and Donaldson on things the evaluator can do to foster generalizability. They are excellent recommendations in principle, although they do vary in cost and feasibility. More important, they are premised on the claim that funders would not “allocate resources to evaluation if they assumed that it was a purely historical exercise with no generalizability to the times, clients, and setting(s) in which they are interested” (p. 31). This assumption is worth exploring for its validity. It seems just as plausible to say the opposite, that many funders of evaluation (as opposed to funders of research) are interested exactly in getting accountability data for a particular intervention. This might especially be the case for evaluators in AEA, more of whom probably do small-scale local evaluations than the kinds of large-scale randomized trials of potential policies more characteristic of the federal level. This suggests the need for more work on when it is and is not the evaluator’s responsibility to attend more to external validity.

Mark alludes to this in his suggestion about doing research on stakeholder beliefs about generalizability. Calls for such research are frequent, but rarely done. Too much debate in evaluation theory, including most of the present volume, is uninformed by anything much resembling systematic evidence. More than a decade ago, I made the same observation about methodological theory, with a special focus on the failure of the Campbellian tradition to investigate empirically its own theory of quasiexperimentation (Shadish, 2000). Today that situation is greatly improved, and we can make many more data-based recommendations about our design and analysis recommendations (Shadish & Cook, 2009). If evaluation theory is really going to advance debate, it must have a substantial evidentiary component that it currently lacks.

Reichardt

I encourage readers to use Reichardt’s (this issue) proposed typology if they find it useful. However, as the reader might suspect, I do not agree with his
analysis of SCC, and find that three of the four criticisms of it are incorrect, and the fourth is trivial. First is the claim that SCC made external and construct validity equivalent. The argument is wrong in many ways. It goes wrong at the start in its incorrect assertion that all construct validity claims in SCC must refer to a causal inference: “When a higher-order construct is to be used in a causal inference, there is simply no other way to determine which challenges to construct validity must be accepted and which may be denied.” Of course there is another way: “We use a pattern-matching logic to decide whether a given instance sufficiently matches the prototypical features to warrant using the category label” (SCC, p. 67). This is the method used since construct validity was first discussed (e.g., Cronbach & Meehl, 1955). SCC did not invent it. Of course, there is nothing wrong with reference to a causal inference in a construct validity argument. But it is unlikely that every term in every sentence used by evaluators refers to a causal inference, even when the evaluation has such an inference as a key focus. Under what rationale would evaluators have license to err about constructs that do not involve causal inference? Are we exempt from debates about, say, what constitutes “being black in the U.S. today”? (CC, p. 62). Of course not.

Reichardt’s second criticism of SCC is that some people use the term external validity differently from SCC, and that they may also use external validity to mean something different from generalization. The point is correct, but also inherent in the ordinary use of language. The logical positivists tried to eliminate such problems, and failed. I doubt Reichardt will succeed, either. He then concludes that SCC is flawed because “A complete typology of criticisms of inferences should include the criticism that an inference is too narrow and not just that it is invalid” (p. 48). No logical grounds for such a claim exist at all, at least no more so than for the claim that a complete typology of criticisms of inferences ought to include any of a myriad of other new validity types proposed in this issue and elsewhere. A moment’s reflection after reading this issue should reveal to any reader that a complete typology of criticisms of inferences is probably impossible.

Third, he claims SCC conflates validity and precision, concluding that “it is impossible for SCC to address properly the issue of precision under the rubric of validity” (p. 49). But he does not actually point to any errors in discussions of matters like power or heterogeneity or other things that affect precision in SCC. It seems this is just a disagreement about which word should be used. Lacking any evidence of an actual mistake, it is difficult to give this criticism much credence.

Fourth, he claims SCC omits time. We did omit it from the list of study facets: persons, settings, treatments, and measurements. We did so after considerable discussion, influenced by Cronbach (1982), who also omitted time from his UTOS model. That model said that all studies consist of units, treatments, observations, and settings (the first letters of each of these four words combining to form UTOS). Cronbach did not include time in his model. He argued that time has two parts. One is historical time (e.g., the Middle Ages,
the Great Depression, the 1960s), where he thought replication of time is impossible. The second is temporal variables associated with treatment (length of treatment in time), measurement (length of follow-up), persons (birthdate), and settings (age of infrastructure). He argued that historical time should be omitted from study facets because it is not accessible to the researcher in any feasible way, and that the other aspects of time could be incorporated well enough into the other four facets. We followed his lead, but the decision is clearly debatable (if trivial). However, the extensive examples using time in SCC belie Reichardt’s criticism that we omit it or otherwise give it short shrift.

**Julnes**

Julnes (this issue) has constructed another alternative validity typology. As I said about Reichardt’s, I encourage readers to use it if they find it useful. Julnes notes that the inferences we make tend to overlap and so are best represented by overlapping dimensions. No doubt exists that the inferences overlap; SCC (pp. 93–102) discussed several such relationships and priorities. In addition, threats to validity also overlap in ways not much discussed in SCC. For example, attrition is listed as a threat to internal validity. But because sample size drops, it can threaten power (statistical conclusion validity), may require changing how we describe who is and is not in the study (construct validity), and may raise questions about whether the intervention would have the same effect in those who dropped out (external validity). I doubt that any system can eliminate all these overlaps, so there is much to be said for Julnes’ embrace of them. Yet some simplicity is clearly lost. His figures show the potential proliferation, with 32 different types of possible inferences compared to SCC’s 4. How well such complexity can serve as a teaching device or guide to practice is unclear. But Julnes’ system is certainly an interesting intellectual exercise, especially as an illustration of the possible overlaps and innovations that might lie stated or unstated in the simpler SCC system.

Julnes’ system is much larger for a second reason: He wants to develop a system covering all “the types of valid inference important to the evaluation community,” (p. 57), not just causal inference with experiments. (It is not clear why Julnes describes SCC as “a general framework applicable to all research and evaluation” (p. 56) when SCC is clearly limited to cause-probing research and is not a book on evaluation at all.) In that goal he has, perhaps inevitably, failed, given that he does not include even all the validity types introduced in other chapters in this issue, much less those that have been posited in the past, such as Kirkhart’s multicultural validity. Perhaps he could do so with more time, though the added complexity might reduce its practical utility even more.

**House**

House’s characterizations of the Campbell tradition in all its variants are accurate, and reflect a careful read of and respect for original texts. His suggestions
for how a revised typology might discuss conflict of interest and deliberate bias are reasonable. However, House’s chapter might inadvertently leave the reader with the impression that researchers interested in internal validity are not sufficiently concerned with such matters. To the contrary, if we judge from the references House cites, it was the medical researchers themselves who surfaced the problems, and those researchers are certainly interested in internal validity.

House is trying to ensure that evaluators who are not sensitive to conflict-of-interest problems become so. This is a worthy goal. However, although it is desirable to call attention to this problem and seek remedies, we should recall that the sociology, psychology, and political economy of the contexts within which evaluations occur are constantly evolving. Any system we put into place to identify or prevent conflicts of interest and deliberate bias is likely to be gamed by the players to their own social, political, or economic advantage over time. Further, Campbell would emphasize that individual researchers are inherently limited in their ability to identify their own biases. He would place responsibility on the community of stakeholders to find such conflicts, stakeholders whose motivations may range from altruistic to revenge. I suspect House would agree.

Greene

Greene’s chapter describes her interpretive/constructivist perspective on outcome evaluation. It is unclear what parts of her chapter refer to Campbellian approaches to validity. Unlike House, Greene almost never relates her ideas to specific things that people in the Campbell tradition have actually said. Early on, for example, she makes the claim that “widely accepted conceptualizations of validity rest on the assumption of an objective, neutral evaluator.” She attributes the claim to House, who does not actually say this in his chapter, so the claim remains unattributed. If this refers to Campbell and colleagues, the claim is very badly wrong in very many ways, as House points out in some of his quotes from Campbell. Sometimes she does directly reference the Campbellian tradition, but incorrectly. For example, she dismisses Campbell’s self-identification as an epistemological relativist by claiming Campbell relied on a procedural approach to validation. If procedure is meant to refer to experimental design, then the claim is again just wrong. Campbell described an inherently social theory of knowledge construction whereby the relevant community was best placed to surface problems in inferences.

Real differences between Campbell and Greene no doubt exist. Some are not based in differences in theory or intellectual principles, but rather reflect how to distribute limited resources in practice. For example, the resources Campbell would devote to experiments would not be available to spend as much time in context and on relationships, even though Campbell would agree on the latter’s value to knowledge construction. Other differences may
reflect intellectual disagreements. Take Greene’s “interpretive evaluators likely rely relatively more on persuasive communications than empirical data” (p. 86). Because the bulk of most evaluation reports are probably persuasive communications, no matter who writes them, the question might be how little empirical data Greene thinks is enough—if we could find a way to measure the use of empirical data—and whether there is a cost in terms of knowledge construction due to having too little data. Finally, Campbell would have far less faith than Greene in the ability of any individual evaluator, interpretivist or not, to know “his or her own sociocultural history, beliefs about the social world and about what constitutes warranted knowledge of it, theoretical preferences, and moral and political values” (p. 82). He would worry that the interpretive evaluator who relied on his or her ability to know all this would be likely to increase bias, not decrease it.

Chen and Garbe
Chen and Garbe (this issue) propose an integrative validity model and a bottom-up approach to outcome evaluation. The chapter includes loose reference to Campbell as the root of a top-down approach in evaluation, though such a characterization never appears in Campbell’s work in any form resembling the one in this chapter. That being said, the sentiment behind Chen and Garbe is understandable. When a federal funding agency like the Institute of Education Sciences makes clear it wants evaluators to jump right into randomized experiments that can be read as overlooking much of the pre-experimental work without which randomized experiments are much the worse. The Chen and Garbe recommendations for such pre-experimental work are sensible, although like Gargani and Donaldson, it is not clear what is new here other than the term viable validity.

I do appreciate the authors’ use of empirical examples of interventions like Coordinated Approach to Child Health, and Reconnecting Youth, to bring evidence to bear on their contentions. It puts flesh on the bones of Mark’s call for empirical research on matters treated in this issue. But not enough flesh. It is far too easy to find single examples that support or refute a position. Sampling error alone will guarantee that two identically done randomized experiments powered at .80 will only agree on whether the effect is significant or nonsignificant 68% of the time. If we want compelling evidence about whether efficacious interventions are later supported in effectiveness studies, we need a lot more cases than just two. Similarly, implicit in the proposals of Chen and Garbe is that too many efficacy studies are undertaken with insufficient evidence of viability. We need data on whether this is really the case. My own experience with such studies is that they are frequently grounded in extensive prior research and practice and would meet at least some of what Chen and Garbe seem to want, though obviously this might vary across content areas or funders.
Discussion

Randomized Experiments in the American Evaluation Association

Most of the contributors to this volume (and I) have been active in AEA for its entire existence, often in leadership positions, and were also members of AEA’s predecessor organizations, Evaluation Network (ENet) and Evaluation Research Society (ERS). All of us know something about the organizational tensions that are perhaps illustrated in this volume, tensions born in the very different cultures of ENet and ERS. The former had members who were more likely to be local practitioners, to prefer qualitative methods, and to do small-scale or local evaluations; the latter had members more likely to be associated with federal and state agencies, to use quantitative methods, to look favorably on randomized experiments, and to be involved with large-scale evaluations of matters involving national policy. Tensions flared badly during the qualitative–quantitative debates of the late 1980s. They simmered during the 1990s, always obvious to the observant eye in the tendency for sessions at AEA’s annual conferences to be segregated by the same groupings present at its birth. They flared again in the last 10 years as a result of an emphasis on randomized experiments in federal evaluation policy, a flare igniting the pages of this issue.

Indeed, I imagine that this volume would never have occurred without that emphasis, and that Campbell is merely a lightning rod for discontent with those experiments. Julnes (this issue) says this most clearly: “... Recent controversies, such as the debates over federal policies promoting random-assignment experimental designs... have led some to criticize the SCC framework” (p. 56). This would go a long way toward explaining the uneven level of seriousness with which the chapters investigate what the Campbellian tradition actually says. In my experience, much of AEA has been unfriendly to randomized experiments, and the resulting culture has led to a brain drain from AEA among those interested in rigorous experimental methods, leaving AEA competent in many things but not those methods. AEA cannot afford to lose that expertise given its importance in modern evaluation policy.

Where Are the New Ideas for Outcome Evaluation?

Outside of AEA, I am not sure that anyone is debating validity typologies for outcome evaluation (though debates about validity for testing and assessment are active in education; e.g., Lissitz, 2009). Rather, researchers from such diverse disciplines as medicine, public health, statistics, and economics are investing their energy in solving the practical problems associated with experimentation, in finding methods by which research about effective treatments can be used to improve practice, and in creating ways to ameliorate the trade-offs that might be present regarding internal and external validity. I increasingly find those literatures more useful to my work
than the evaluation literature. AEA certainly produces new and useful ideas for evaluation in general. But when it comes to outcome evaluation specifically, I get the sense of having the same recycled debates I have heard for decades.

Less Talk, More Data

I enjoy a good theoretical debate as much as the next guy. But my comments on a number of the chapters called for empirical data that could be used to clarify the validity of assumptions (e.g., whether it is true that we attend more to internal than external validity) or that could provide tests of theoretical claims that are part of a theory. Some of the contributors to this volume also make that call. Such data are too often absent in the debate, where claims are treated as logically necessary or implicitly compelling instead of as testable hypotheses. The result is reminiscent of the state of Aristotelian theory prior to the scientific revolution where theory was the dogma that dictated what data were acceptable rather than data being used to correct errors in dogma. Of course, things are not that black and white, but we are far too much of the former side of the continuum with a near total absence of data to inform our evaluation theory. Gathering data on matters like those treated in this book is admittedly difficult, but probably easier than doing evaluations themselves. At least some pertinent data exist already (e.g., Christie, 2003). What I am suggesting requires a cognitive change in which we think of our own claims as testable, develop the interest in doing so, and constantly spot-check ourselves to see whether we are making a claim that could be tested.

Conclusion

Outcome evaluation is experiencing a very exciting time right now. We are beginning to understand the conditions under which various kinds of non-randomized experiments might approximate results from randomized experiments (Shadish, Clark, & Steiner, 2008; Shadish, Galindo, Steiner, Wong, & Cook, in press). New approaches to causation are appearing and being combined with older ones (e.g., Shadish & Sullivan, in press). I wish I knew how to bring that sense of excitement and optimism about outcome evaluation into AEA.

References


---

**WILLIAM R. SHADISH** is a professor of psychology at the University of California, Merced.