What Counts as Credible Evidence in Applied Research and Evaluation Practice? Edited by Stewart I. Donaldson, Christina A. Christie, & Melvin M. Mark, 2008. Thousand Oaks, CA: Sage. $36.94

Reviewed by Jason T. Burkhardt and Jan K. Fields
Western Michigan University

The relatively young field of evaluation, which by one account began to crystallize and emerge as a distinct profession around 1973 (Stuffelbeam & Shinkfield, 2007), has already seen more than its share of growing pains. Most notable are the "Paradigm Wars," a protracted and often contentious debate between quantitative and qualitative methodologists, which lasted from the late 1970s through the 1980s.

The relative calm that followed in the 1990s did not last long. Old wounds opened again in 2003 when the U.S. Department of Education's Institute of Educational Sciences (IES) issued a notice of proposed priority that privileged experimental and some types of quasi-experimental designs over other evaluation methods in education evaluation funding competitions.

In the opening chapter in Part I of *What Counts as Credible Evidence in Applied Research and Evaluation Practice?*, Stewart Donaldson recounts the response of the evaluation community. The leadership of the American Evaluation Association (AEA) in 2004 published a statement opposing the new policy. In turn, a group of senior evaluators distanced themselves from the official AEA statement by posting their own statement in support of privileging experimental designs as applicable, thus threatening to reignite the old methodological debate.

In August 2006, a group of recognized thought leaders in the field of evaluation assembled to share their perspectives on the question of "What Counts as Credible Evidence?" The goal of this symposium, as Donaldson states, was to attempt to build bridges so that leaders on both sides of the debate would "stay together in a united front against the social and human ills of the 21st century" (p. 12), under a "shared blueprint for an evidence-based global society" (p. 12) The discussions from that symposium are compiled in this book.

In addition to Donaldson's historical background of the debate on credible evidence, Part I of the book also includes a paradigm background by Christina Christie and Dreolin Fleischer. These authors define a paradigm as a theory of knowledge which consists of an ontology (nature of reality), an epistemology (what is knowable and who can know it), and a methodology (how one can obtain knowledge). They go on to assert that most evaluators align themselves to one of two paradigm camps: post-positivism and constructivism.

Ontologically, post-positivists believe in a single, objectively observable reality, although recognizing that it may not be understood it in its entirety. The constructivists believe that there are multiple, subjective realities that may change according to the knower. Epistemologically, the post-positivists believe that the knower is independent of that which he or she is trying to know, while the constructivists believe that the knower and the known are interrelated. Finally, methodologically, the post-positivists have a strong preference for quantitative methods and deductive reasoning, while the constructivists more frequently use qualitative methods and inductive reasoning.

This paradigm background sets the stage for the rest of the book, which consists of several post-positivist contributors writing in support of discovering credible evidence via experimental methods, followed by several constructivist contributors writing in support of determining credible evidence via non-experimental methods. Christie and Fleischer briefly mention that the mixed-methods approach evolved to circumvent the methodology debate; although frequently in practice one of the two basic methodologies counts as credible evidence while the other plays a supporting role only, thus providing little resolution to the polarizing nature of the ongoing debate.

Part II consists of four post-positivist essays making the case for experimental approaches as the preferred route to credible evidence. Part III consists of five constructivist (or at least constructivist-friendly) essays supporting the use of non-experimental approaches to obtain credible evidence. We will integrate reviews of the essays from both perspectives here to illustrate the contrasting nature of the perspectives.

Gary Henry roars out of the gate on behalf of the post-positivist case for credible evidence by stating that high-quality experimental evaluations are the only way to eliminate selection bias when assessing policy and program impact, period. Bias is a systematic difference between a parameter value and an obtained statistic that, as a first line of defense against evaluations of poor quality, the evaluator should confront, quantify, and when possible, reduce to practical insignificance. Modern democracies need credible assessments of the causal impacts of public policies and programs and RCTs are the most conclusive means to filter out bad ideas and are available for almost all situations.

Michael Scriven, though not exactly a constructivist, disagrees with Henry in his essay. His primary argument involves positing that methods other than the RCTs can provide evidence capable of meeting the "gold standard" of credible evidence. This is an argument that Scriven has made before.[1] Scriven evinces a credible argument that is easily accessible

[1] Scriven has made this point in *Hard Won Lessons in Program Evaluation* (1993) as well as in Cook, Scriven, Coryn, and Evergreen (2010).

to those who are looking for ways in which to argue solid cases for their evaluations, but may be on a limited budget or lack the infrastructure needed to perform an RCT.

Leonard Bickman and Stephanie Reich tone down the case for RCTs a bit but they still toe the post-positivist line. They argue for an acute awareness of the numerous threats to validity in evaluation and of the high costs of making the wrong decision. They explain that while RCTs may not be perfect, they seem to address threats to validity more adequately than other methods (although Scriven makes the case that the General Elimination Method [GEM] is the actual mechanism by which we do this, not the RCT). Still, the authors caution that judging credibility necessitates more than simply considering design. It also necessitates information about the questions asked: who asked the question and gathered the evidence, what evidence was gathered, how was the evidence gathered and analyzed, and under what conditions was the evaluation undertaken.

Jennifer Greene agrees that it is not just about design and introduces the political influences surrounding evidence basis for evaluation. She begins by discussing the background of the fight for evidence-based social science, and subsequently provides a rebuke of the main arguments for current evidence-based thinking by pointing to the role of politics in dictation of methodology. She goes on to explore human complexity as a major argument for changing the conversation about evidence. Carol Weiss sang a similar tune, and Lee Cronbach appears to agree through his analyses of complex interactions in determining program effects. (Shadish, Cook, & Leviton, 1991)

Russell Gersten and John Hitchcock would beg to differ on the role of politics in evaluation. In their essay, they champion a database created by the IES called the What Works Clearinghouse (WWC). Like the National Institutes of Health, the IES is an independent institution that is protected from direct political influence. They go on to explain the thinking behind the WWC, especially focusing on the value the project places on randomized designs. The WWC has two goals: to provide guidance for judging quality and effectiveness of an intervention and to provide guidance for designing intervention research. While the authors of this essay are not saying that program evaluation should use only RCTs or high-caliber quasi-experimental designs, they are saying there are not enough of these designs in current educational evaluation research and that the WWC can remedy that situation if properly used.

Sharon Rallis will not have any part of that argument. Her essay begins with a discussion of how an evaluation of a program using one specific outcome produced negative results, thus resulting in a loss of funding. She posits that this situation is representative of the shortcomings of the RCT in answering human questions. She also asserts that evaluation informs policy, not science, and therefore methods that only focus on narrowly defined indicators are not effective. Rallis' main point becomes that credible evidence is that which speaks to wholesale truth and moral soundness (which she spends a great deal of time defining). This view is supportive to those who utilize participant-focused evaluation methods such as democratic deliberation and transformative research methods (Shadish, Cook, & Leviton, 1991).

George Julnes and Debra Rog take the first steps towards reconciliation between the methodology antagonists in their

essay. While not giving ground on the mostly post-positivist notion that there IS such a thing as a "best" methodology, these authors concede that actionable evidence is as important as credible evidence and provide a pragmatic framework for method choice to produce findings that in turn guide action. This framework posits four questions: (1) how are the methods to be related to the questions being asked, (2) what contextual factors condition the evidence needed to support causal conclusions, (3) how are we to judge the adequacy of methods for providing the needed evidence, and (4) when is it appropriate to use particular methods of causal analysis. The main point here is that method choice should be contextual, contingent, and political to produce both credible and actionable evidence.

Sandra Mathison is not taking the olive branch. She takes the contextual argument to breath-taking heights by describing the use of found, participant, and researcher derived images as credible evidence. Her reasoning involves the use of these images to elicit processes and gain information that is not readily available through other means of research. She bases her arguments on the statement that the credibility of evidence is dependent on the context of the research and the research design itself. This extends to support her broader view of what counts as credible evidence as well. However, some readers may not be satisfied with this argument if they are looking for more standardized approaches to gathering credible evidence.

Thomas Schwandt masks his ontology (boxers? briefs? we don't know) and plays the consummate peacemaker by discussing the need to abandon old debates that have prevented the move to a more enlightened view of evidence in social science. His analysis is more philosophical than those of the other contributors, but is salient to the debate nonetheless in his assertion that "what counts as credible evidence is not the same as what counts for credible evaluation" (p. 209). This chapter provides the most comprehensive view of what counts as credible evidence. It is also the most egalitarian in its treatment of the various methods as being equal in their ability to answer research questions.

Jason's wrap-up: The conclusion of this book seeks to resolve the debate over what counts as credible evidence by asserting that those who ask the questions and those who answer them really determine what counts as credible evidence. I believe this is a salient point that illuminates another point made elsewhere—reality is measurable but the interpretation of reality is constructed. This book finds the ultimate middle ground, and makes valuable prescriptions for the future practice of evaluation.

Jan's wrap-up: This symposium/book had two primary goals: (1) to create bridges between the warring parties of methodology and (2) to create a blueprint for an evidence-based global society. I'm not so sure it met its first goal. It certainly does not create a consensus or a universal answer to what counts as credible evidence. However, it does keep the lines of communication open. So if by "bridge" we mean openness to discussion, then yes, that goal is met.

I feel more confident that the second goal is met because the dialogue captured by this book stays away from fundamentalist extremes and strives to reach a pragmatic middle ground where we can, even with our ontological persuasions, work together towards an evidence-based approach to making this world a better place in which to live.

References

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*(1).

Scriven, M. (1993). Hard won lessons in program evaluation. *New Directions in Program Evaluation, No. 58.* San Francisco, CA: Jossey-Bass.

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

Stuffelbeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, & applications.* San Francisco, CA: Jossey-Bass.