What Have We Learned About RCTs, Gold Standards, and Credible Evidence: Moving Beyond the Debates to Improve Evaluation Practice

Stewart I. Donaldson

Claremont Graduate University, USA

Abstract

This presentation will summarize key findings from a new book from SAGE on "What Counts as Credible Evidence in Applied Research and Evaluation Practice" (Donaldson, Christie, & Mark, 2009). Many thorny debates about what counts as credible evidence have occurred in recent years, but few have sorted out the issues in a way that directly informs evaluation practice. In this volume, internationally renowned evaluators explore the challenges of designing and executing high quality evaluations in contemporary evaluation practice. A summary of what can be learned from the chapter authors about the strengths and weaknesses of both experimental and non-experimental approaches for gathering credible and actionable evidence will be presented. A proposal to revise the notion of an "Experimenting Society" to an "Evidence-based Global Society", which includes replacing the "RCT Gold Standard" with the gold standard of "Methodological Appropriateness" will be offered as a avenue toward improving evaluation policy and practice.

Key words: Credible Evidence, Randomized Controlled Trial, Gold Standard

In 2006, debates about whether randomized controlled trials (RCTs) should be considered the gold standard for producing credible evidence in applied research and evaluation remained front and center. At the same time, the zeitgeist of accountability and evidence-based practice was now widespread across the globe. Organizations of all types and sizes were being asked to evaluate their practices, programs, and policies at an increasing rate. While there seemed to be much support for the notion of using evidence to continually improve efficiency and effectiveness, there appeared to be growing disagreement and confusion about what constitutes sound evidence for decision making. These heated disagreements among leading lights in the field had potentially far-reaching implications for evaluation and applied research practice, for the future of the profession (e.g., there was visible disengagement, public criticisms, and resignations from the main professional associations), for funding competitions, as well as for how best to conduct and use evaluation and applied research to promote human betterment.

In light of this state of affairs, an illustrious group of experts working in various areas of evaluation and applied research were invited to Claremont Graduate University to share their diverse perspectives on the question of "What Counts as Credible Evidence?" The ultimate goal of this symposium was to shed more light on these issues, and to attempt to build bridges so that prominent leaders on both sides of the debate would stay together in a united front against the social and human ills of the 21st century. In other words, a full vetting of best ways to produce credible evidence from both experimental and non-experimental perspectives was facilitated in the hope that the results would move us closer to a shared blueprint for an evidence-based global society. This illuminating and action-packed day in Claremont, California, included over 200 attendees from a variety of backgrounds—academics, researchers, private consultants, students, and professionals from many fields—who enjoyed a day of stimulating presentations, intense discussion, and a display of diverse perspectives on this central issue facing the field (see webcast at www.cgu.sbos). Each presenter was asked to follow up his or her presentation with a more detailed chapter for this book. In addition, George Julnes and Debra Rog were invited to contribute a chapter based on their findings from a recent project focused on informing federal policies on evaluation methodology (Julnes & Rog, 2007).

Our search for a deeper and more complete understanding of what counts as credible evidence begins with an analysis of the passion, paradigms, and assumptions that underlie many of the arguments and perspectives expressed throughout this book. In Chapter 2, Christina Christie and Dreolin Fleischer provide a rich context for understanding the nature and importance of this debate. Ontological, epistemological, and methodological assumptions that anchor views about the nature of credible evidence are explored. This context is used to preview the positions expressed about credible evidence in the subsequent sections of the book.

Experimental Routes to Credible Evidence

Part II contains four chapters that discuss the importance of experimental and quasi-experimental approaches for producing credible and actionable evidence in applied research and evaluation. In Chapter 3, Gary Henry sketches out an underlying justification for the U.S. Department of Education's priority for randomized experiments and high quality quasiexperiments over nonexperimental designs "when getting it right matters." His argument has deep roots in democratic theory, and stresses the importance of scientifically based evaluations for influencing the adoption of government policies and programs. He argues that high-quality, experimental evaluations are the only way to eliminate selection bias when assessing policy and program impact, and that malfeasance may occur when random assignment evaluations are not conducted. Henry urges his readers to consider his arguments in favor of the proposed priority in an open-minded, reflective, and deliberative way to do the greatest good in society.

In Chapter 4, Leonard Bickman and Stephanie Reich explore in great detail why RCTs are commonly considered the "gold standard" for producing credible evidence in applied research and evaluation. They clearly articulate why RCTs are superior to other evaluation designs for determining causation and impact, and alert us to the high cost of making a wrong decision about causality. They specify numerous threats to validity that must be considered in applied research and evaluation, and provide a thorough analysis of both the strengths and limitations of RCTs. In the end, they conclude that, "For determining causality, in many but not all circumstances, the randomized design is the worst form of design except all the others that have been tried."

One popular approach for determining if evidence from applied research and evaluation is credible for decision-making has been to establish what might be thought of as "supreme courts" of credible evidence. These groups establish evidence standards and identify studies that provide the strongest evidence for decision and policy making. For example, the Cochrane Collaboration is known as the reliable source for evidence on the effects of health care interventions, and it aims to improve health care decision making globally (www.cochrane.org). The Campbell Collaboration strives to provide decision makers with evidence-based information to empower them to make well-informed decisions about the effects of interventions in the social, behavioral, and educational arenas (www.campbellcollaboration.org). In Chapter 5, Russell Gersten and John Hitchcock describe the role of the What Works Clearinghouse (WWC) in informing decision makers and being the "trusted source of scientific evidence in education" (http://ies.ed.gov/ncee/wwc). They discuss in some detail how the Clearinghouse defines and determines credible evidence for the effectiveness of a wide range of educational programs and interventions. It is important to note that well-implemented RCTs are typically required to meet the highest standards in most of these evidence collaborations and clearinghouses, and applied research and evaluations that do not use RCTs or strong quasi-experimental designs do not make it through the evidence screens or meet credible evidence standards.

George Julnes and Debra Rog discuss their new work on informing method choice in applied research and evaluation in Chapter 6. Their pragmatic approach suggests that for evidence to be useful, it not only needs to be credible but "actionable" as well, deemed both adequate and appropriate for guiding actions in targeted real-world contexts. They argue that evidence can be credible in the context studied but of questionable relevance for guiding actions in other contexts. They provide a framework to address the controversy over method choice and review areas where there is at least some consensus, in particular with regard to the key factors that make one method more or less suitable than others for particular situations. The contexts and contingencies under which RCTs and quasi-experimental designs are most likely to thrive in providing credible and actionable evidence are described. They conclude by suggesting their approach to the debate about evidence, focusing on the specific application of methods and designs in applied research and evaluation, promises to develop a "fairer" playing field in the debate about credible evidence than one that is based solely on ideological instead of pragmatic grounds.

Nonexperimental Approaches

Part III includes five chapters that explore nonexperimental approaches for building credible evidence in applied research and evaluation. Michael Scriven (Chapter 7) first takes a strong stand against the "current mythology that scientific claims of causation or good evidence require evidence from RCTs." He asserts, "to insist that we use an experimental approach is simply bigotry, not pragmatic, and not logical—in short a dogmatic approach that is an affront to scientific method. And to wave banners proclaiming that anything less will mean unreliable results or unscientific practice is simply absurd." Next, he provides a detailed analysis of alternative ways to determine causation in applied research and evaluation, and discusses several alternative methods for determining policy and program impact including the general elimination methodology or algorithm (GEM). He ends with a proposal for marriage of warring parties, complete with a prenuptial agreement that he believes would provide a win-win solution to the "causal wars," with major positive side effects for those in need around the world.

In Chapter 8, Jennifer Greene outlines the political, organizational, and sociocultural assumptions and stances that comprise the current context for the demand for credible evidence. She quotes Stronach, Piper, and Piper (2004), "The positivists can't believe their luck, they've lost all the arguments of the last 30 years and they've still won the war," to illuminate that the worldview underlying the current demand for credible evidence is a form of conservative post-positivism, or in many ways like a kind of neo-positivism. She laments that "many of us thought we'd seen the last of this obsolete way of thinking about the causes and meaning of human activity, as it was a consensual casualty of the great quantitative-qualitative debate." She goes on to describe the ambitions and politics behind priorities and approaches privileging methods and designs like RCTs, and the problems with efforts to promote one master epistemology and the interests of the elite, which she concludes is radically undemocratic. Greene offers us an alternative view on credible evidence that meaningfully honors complexity, and modestly views evidence as "inkling" in contrast to "proof." She describes how credible evidence can provide us a window into the messy complexity of human experience; needs to account for history, culture, and context; respects differences in perspective and values; and opens the potential for democratic inclusion and the legitimization of multiple voices.

Sharon Rallis describes qualitative pathways for building credible evidence in Chapter 9. She emphasizes throughout her chapter that probity, goodness or absolute moral correctness, is as important as rigor in determining what counts as credible evidence in applied research and evaluation. It is also important to her that scientific knowledge be recognized as a social construct, and that credible evidence is what the relevant communities of discourse and practice accept as valid, reliable, and trustworthy. A wide range of examples focused on reported experiences rather than outcomes are provided, and offered as a form of credible evidence to help improve policy and programming and to better serve the people involved. Rallis argues that these qualitative experiences provide credible evidence that is the real basis of scientific inquiry.

In Chapter 10, Sandra Mathison explores the credibility of image-based applied research and evaluation. She asserts that the credibility of evidence is contingent on experience, perception, and social convention. Mathison introduces the notion of an anarchist epistemology, the notion that every idea, however new or absurd, may improve knowledge of the social world. She asserts that credible evidence is not the province of only certain methods (e.g., RCTs), and can't be expressed in only one way (e.g., statistical averages). Qualities of good evidence include relevance, coherence, verisimilitude, justifiability, and contextuality. She concludes by pointing out that it is too simplistic to assert that "seeing is believing," but the fact that our eyes sometimes deceive does not obviate credible knowing from doing and viewing image-based research and evaluation.

Thomas Schwandt provides the final chapter of Part III. He claims that evaluating the merit, worth, and significance of our judgments, actions, policies, and programs requires a variety of evidence generated via both experimental and nonexperimental methods. He asserts in Chapter 11 that RCTs are not always the best choice of study design, and in some situations do not provide more credible evidence than nonrandomized study designs. That is, observational studies often provide credible evidence as well. Schwandt believes that careful thinking about the credibility, relevance, and probative value of evidence in applied research and evaluation will not be advanced in the future by continuing to argue and debate the merits of hierarchies of evidence as a basis for decision making. Rather, he suspects that the field of applied research and evaluation would be better served by working more diligently on developing a practical theory of evidence, one that addressed matters such as the nature of evidence as well as the context and ethics of its use in decision making.

Conclusions

In Chapter 12, the final chapter, Melvin Mark reviews some of the central themes about credible evidence presented throughout the book, and underscores that at this time in our history this is a topic where we do not have consensus. For example, some authors firmly believe that RCTs are needed to have credible and actionable evidence about program effects, while others assume that nonexperimental methods will suffice for that purpose, and yet other authors argue that the question of overall program effects is too complex to answer in a world in which context greatly matters. In an effort to move the field forward in a productive and inclusive manner, Mark provides us with an integrative review of the critical issues raised in the debate, and identifies a few underlying factors that account for much of the diversity in the views about what counts as credible evidence. He concludes by giving us a roadmap for changing the terms of a debate that he believes will help us dramatically improve our understanding of what counts as credible evidence in applied research and evaluation.

The Epilogue by Donaldson supports and expands this roadmap and begins to flesh out a possible blueprint for an evidence-based global society. Together, Mark and Donaldson provide us with hope that the result of this volume will be to inspire new efforts to improve our understanding of deeply entrenched disagreements about evidence, move us toward a common ground where such can be found, enhance the capacity of evaluation practitioners and stakeholders to make sensible decisions rather than draw allegiances to a side of the debate based on superficial considerations, and ultimately provide applied researchers and evaluators with a useful framework for gathering and using credible evidence to improve the plight of humankind across the globe as we move further into the 21st century.

References

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.) (2009). What counts as credible evidence in applied research and evaluation practice? Newbury Park, Sage: CA.

Julnes, G., & Rog, D. J. (Eds.). (2007). *Informing federal policies on evaluation methodology: Building the evidence base for method choice in governmentsponsored evaluations* [Entire issue]. (Vol. 113 of *New Directions for Evaluation* series). San Francisco: Jossey-Bass.

Stronach, I., Piper, H., & Piper, J. (2004). Re-performing crises of representation. In H. Piper & I. Stronach (Eds.), *Educational research, difference and diversity* (pp. 129–154). Aldershot, UK: Ashgate.

For more details and resources see: http://sites.google.com/site/credibleevidence/