

AEA/CDC Summer Evaluation Institute

Offering 31: What Counts as Credible Evidence in Contemporary Evaluation Practice: Moving Beyond the Debates

Description: This workshop is designed to explore one of the most fundamental issues facing evaluators today, and the 4th step in CDC's Framework for Program Evaluation, what counts as credible evidence in contemporary evaluation practice? Many thorny debates about what counts as evidence have occurred in recent years, but few have sorted out the issues in a way that directly informs contemporary evaluation and evidence-based practice. Participants will come away from this workshop with an understanding of the philosophical, theoretical, methodological, political, and ethical dimensions of gathering credible evidence and will apply these dimensions to fundamental evaluation choices we encounter in applied settings.

Audience: Attendees should have a basic background in evaluation

Stewart I. Donaldson, Ph.D. is Professor and Chair of Psychology, Director of the Institute of Organizational and Program Evaluation Research, and Dean of the School of Behavioral and Organizational Sciences, Claremont Graduate University. He has conducted numerous evaluations, developed one of the largest university-based evaluation training programs, published numerous evaluation articles and chapters, and his recent books include *Program Theory-Driven Evaluation Science: Strategies and Applications* (2007), *Applied Psychology: New Frontiers and Rewarding Careers* (2006; with D. Berger & K. Pezdek), *Evaluating Social Programs and Problems: Visions for the New Millennium* (2003; with M. Scriven), *Social Psychology and Policy/Program Evaluation* (forthcoming; with M. Mark & B. Campbell), and *What Counts as Credible Evidence in Evaluation and Evidence-Based Practice?* (forthcoming; with C. Christie & M. Mark). He is co-founder of the Southern California Evaluation Association and is on the Editorial Boards of the *American Journal of Evaluation* and *New Directions for Evaluation*.

Offered (Two Rotations of the Same Content - Do not register for both):

- Monday, June 23, 2:30 – 4:00 PM
- Tuesday, June 24, 2:30 – 4:00 PM

Donaldson, S. I. (in press). In search of the blueprint for an evidence-based global society. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* Newbury Park, CA: Sage.

INTRODUCTION

The Blueprint for an Evidence-Based Global Society

1

IN SEARCH OF THE BLUEPRINT FOR AN EVIDENCE-BASED GLOBAL SOCIETY

Stewart I. Donaldson

As we near the end of the first decade of the 21st century, approximately 1.5 billion members of our vast global community live in pain and misery, and another 2.5 billion are also shut out of the “good life”—just simply getting by, day by day (a total of 4 billion people; Gallup, 2007). On a brighter note, there are now approximately 2 billion people on planet earth, or 33% of the world’s population, who have a favorable standard of living and report high levels of health and well-being on a regular basis (Gallup, 2007). This picture is obviously not good enough for those of us who devote our lives and careers to conducting applied research and evaluation, in efforts to prevent and ameliorate human misery and suffering, and to promote social justice and human betterment.

Basic research is typically driven by scientists’ personal interests and curiosity, and its purpose is to advance knowledge. In contrast, the purpose of applied research is to understand how to prevent or solve practical problems that affect “real” people, organizations, communities, and societies across the globe. Some applied research is descriptive and helps advance our understanding of practical problems and their potential solutions, while other efforts are evaluative and improve or determine the effectiveness of actions (e.g., programs and policies) to prevent and solve practical problems. Donaldson and Christie (2006) described some of the major differences between basic research and applied research and evaluation, such as the origin of the study questions, the purposes for which study information is gathered, and the settings in which the work is conducted. For the purposes of this volume, we are going to focus on what counts as credible evidence in applied research and evaluation. That is, most of the scientific work we will be discussing is problem based or solution oriented, and conducted in “real world” settings as opposed to highly controlled, traditional scientific laboratories.

The Rise and Fall of the Experimenting Society

In 1969, one of the legendary figures in the history of applied research and evaluation, Donald T. Campbell, gave us great hope and set what we now call the “applied research and evaluation community” on a course for discovering a utopia he called the “Experimenting Society” (Campbell, 1991). His vision for this utopia involved rational decision making by politicians based on hard-headed tests of bold social programs designed to improve society. The hard-headed tests he envisioned were called randomized experiments and were focused on maximizing bias control in an effort to provide unambiguous causal inferences about the effects of social reforms. This ideal society would broadly implement social reforms demonstrated to be highly effective by experimental research and evaluation, with the goal of moving at least more, if not most, of the population toward the “good life.”

Some of the most important methodological breakthroughs in the history of applied research and evaluation seemed to occur during this movement toward the Experimenting Society (e.g., Campbell, 1991; Campbell & Stanley, 1963; Cook & Campbell, 1979). For example, detailed understanding of threats to validity, multiple types of validity, bias control, and the implementation of rigorous experimental and quasi-experimental designs in “real world” or field settings were advanced during this era.

However, the progress and momentum of the movement were not sustained. By the early 1980s, it was clear that Campbell’s vision would be crushed by the realities of programs, initiatives, and societal reforms. Shadish, Cook, and Leviton (1991) reported that information or evidence judged to be poor by experimental scientific standards, was often considered acceptable by key decision makers including managers, politicians, and policy makers. Further, they argued that rigorous experimental evaluations did not yield credible evidence in a timely and useful manner, thus inspiring the field to develop new tools, methods, and evaluation approaches. The practice of applied research and evaluation today has moved way beyond the sole reliance on experimentation and traditional social science research methods (Donaldson, 2007; Donaldson & Lipsey, 2006; Donaldson & Scriven, 2003).

An Evidence-Based Global Society

Shades of Campbell’s great hopes for evidence-based decision making can be seen in much of the applied research and evaluation discourse today. However, while the modern discussion remains focused on the importance of the production and use of credible evidence, it is not limited to evidence derived from experimentation. The new vision for a utopia seems to require broadening Campbell’s vision from an “experimenting” to an “evidence-based society.” This ideal society would certainly include evidence from experimentation under its purview, but would also include a wide range of evidence derived from other applied research and evaluation designs and approaches. Many of these newer approaches have been developed in the past two decades and no longer rely primarily on the traditional social science experimental paradigm (see Alkin, 2004; Donaldson & Scriven, 2003).

The momentum for developing an evidence-based global society seems to be at an all-time peak. No longer is applied research and evaluation primarily concentrated in Washington, D.C., or federal governments more broadly. Rather, organizations of all types, shapes, and sizes are commissioning applied research and professional evaluations in pursuit of evidence-based decision making at an accelerating rate.

One striking indicator of the demand for evidence and the growth of applied research and evaluation is the popularity of professional societies across the globe. For example, in 1980 there were only three national and regional evaluation societies. This number almost doubled (five) by 1990. Just one decade later (by 2000), there were more than 50 professional evaluation societies, and today there are more than 75 with a formal international cooperation network to build evaluation capacity across the globe (Donaldson, 2007; Donaldson & Christie, 2006). Members of most of these societies meet regularly to learn and discuss how best to conduct applied research and evaluation in order to yield credible evidence for promoting human betterment.

Another window into the promise of an evidence-based society and the accelerating demand for credible evidence, is the recent proliferation of evidence-based discussions and applications. For example, these discussions and applications are now prevalent throughout the fields of health care and medicine (Sackett, 2000; Sackett, Rosenberg, Gray, & Haynes, 1996), mental health (Norcross, Beutler, & Levant, 2005), management (Pfeffer & Sutton, 2006), executive coaching (Stober & Grant, 2006), career development (Preskill & Donaldson, 2008), public policy (Pawson, 2006), and education (Chapter 5, this volume) just to name a few. In fact, a cursory search on Google yields many more applications of evidence-based practice. A sample of the results of a recent search illustrates these diverse applications:

- ◆ Evidence-based medicine
- ◆ Evidence-based mental health
- ◆ Evidence-based management
- ◆ Evidence-based decision making
- ◆ Evidence-based education
- ◆ Evidence-based coaching
- ◆ Evidence-based social services
- ◆ Evidence-based policing
- ◆ Evidence-based conservation
- ◆ Evidence-based dentistry

- ◆ Evidence-based policy
- ◆ Evidence-based thinking about health care
- ◆ Evidence-based occupational therapy
- ◆ Evidence-based prevention science
- ◆ Evidence-based dermatology
- ◆ Evidence-based gambling treatment
- ◆ Evidence-based sex education
- ◆ Evidence-based needle exchange programs
- ◆ Evidence-based prices
- ◆ Evidence-based education help desk

One might even consider this interesting new phenomenon across the disciplines to be expressed in the following formula:

Mom + The Flag + Apple Pie = Evidence-Based Practice

Or it might be expressed as

In God We Trust—*All Others Must Have Credible Evidence*

The main point here is the movement toward evidence-based decision making now appears highly valued across the globe, multidisciplinary in scope, and supported by an ever-increasing number of practical applications.

But wait—while there appears to be strong consensus that evidence is our “magic bullet” and a highly valued commodity in the fight against social problems, there ironically appears to be much less agreement, even heated disagreements, about what counts as evidence.

Unfortunately, seeking truth or agreement about what constitutes credible evidence does not seem to be an easy matter in many fields. Even in periods of relative calm and consensus in the development of a discipline, innovations occur and worldviews change in ways that destabilize. We may be living in such a destabilizing period now in the profession and discipline of applied research and evaluation. That is, despite unprecedented growth and success on many fronts, the field is in considerable turmoil over its very foundation—what counts as credible evidence. Furthermore, contemporary applied research and evaluation practice rests firmly on the foundation of providing credible evidence. If that foundation is

shaky, or built on sand, studies wobble, sway in the wind, and ultimately provide little value, and can even mislead or harm.

Recent Debates About Evidence

Before exploring this potentially destructive strife and dilemma in more detail, let's briefly look at the recent history of debates about applied research and evaluation. The great quantitative-qualitative debate captured and occupied the field throughout the late 1970s and 1980s (see Reichhardt & Rallis, 1994). This rather lengthy battle also became known as the "Paradigm Wars," which seemed to quiet down a bit by the turn of the century (Mark, 2003).

In 2001, Donaldson and Scriven (2003) invited a diverse group of applied researchers and evaluators to provide their visions for a desired future. The heat generated at this symposium suggested that whatever truce or peace had been achieved remained an uneasy one (Mark, 2003). For example, Yvonna Lincoln and Donna Mertons envisioned a desirable future based on constructivist philosophy, and Mertons seemed to suggest that the traditional quantitative social science paradigm, specifically randomized experiments, were an immoral methodology (Mark, 2003). Thomas Cook responded with a description of applied research and evaluation in his world, which primarily involved randomized and quasi-experimental designs, as normative and highly valued by scientists, funders, stakeholders, and policy makers alike. Two illustrative observations by Mark (2003) highlighting differences expressed in the discussion were (1) "I have heard some quantitatively oriented evaluators disparage participatory and empowerment approaches as technically wanting and as less than evaluation," and (2) "It can, however, seem more ironic when evaluators who espouse inclusion, empowerment, and participation would like to exclude, disempower, and see no participation by evaluators who hold different views." While the symposium concluded with some productive discussions about embracing diversity and integration as ways to move forward, it was clear there were lingering differences and concerns about what constitutes quality applied research, evaluation, and credible evidence.

Donaldson and Christie (2005) noted that the uneasy peace seemed to revert back to overt conflict in late 2003. The trigger event occurred when the U.S. Department of Education's Institute of Education Sciences declared a rather wholesale commitment to privileging experimental and some types of quasi-experimental designs over other methods in applied research and evaluation funding competitions. At the 2003 Annual Meeting of the American Evaluation Association (AEA), prominent applied researchers and evaluators discussed this event as a move back to the "Dark Ages" (Donaldson & Christie, 2005). The leadership of the American Evaluation Association developed a policy statement opposing these efforts to privilege randomized control trials in education evaluation funding competitions:

AEA Statement:

American Evaluation Association Response to U. S. Department of Education

**Notice of Proposed Priority, Federal Register RIN 1890-ZA00, November 4,
2003**

“Scientifically Based Evaluation Methods”

The American Evaluation Association applauds the effort to promote high quality in the U.S. Secretary of Education’s proposed priority for evaluating educational programs using scientifically based methods. We, too, have worked to encourage competent practice through our Guiding Principles for Evaluators (1994), Standards for Program Evaluation (1994), professional training, and annual conferences. However, we believe the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions. We would like to help avoid the political, ethical, and financial disaster that could well attend implementation of the proposed priority.

(1) Studies capable of determining causality. Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. In medicine, causality has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary’s proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs which are sometimes more feasible and equally valid.

RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than designs sensitive to local culture and conditions and open to unanticipated causal factors.

RCTs should sometimes be ruled out for reasons of ethics. For example, assigning experimental subjects to educationally inferior or medically unproven treatments, or denying control group subjects access to important instructional opportunities or critical medical intervention, is not ethically acceptable even when RCT results might be enlightening. Such studies would not be approved by Institutional Review Boards overseeing the protection of human subjects in accordance with federal statute.

In some cases, data sources are insufficient for RCTs. Pilot, experimental, and exploratory education, health, and social programs are often small enough in scale to preclude use of RCTs as an evaluation methodology, however important it may be to examine causality prior to wider implementation.

(2) Methods capable of demonstrating scientific rigor. For at least a decade, evaluators publicly debated whether newer inquiry methods were sufficiently rigorous. This issue was settled long ago. Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific. To discourage a repertoire of methods would force evaluators backward. We strongly disagree that the methodological “benefits of the proposed priority justify the costs.”

(3) Studies capable of supporting appropriate policy and program decisions. We also strongly disagree that “this regulatory action does not unduly interfere with State, local, and tribal governments in the exercise of their governmental functions.” As provision and support of programs are governmental functions so, too, is determining program effectiveness. Sound policy decisions benefit from data illustrating not only causality but also conditionality. Fettering evaluators with unnecessary and unreasonable constraints would deny information needed by policy-makers.

While we agree with the intent of ensuring that federally sponsored programs be “evaluated using scientifically based research . . . to determine the effectiveness of a project intervention,” we do not agree that “evaluation methods using an experimental design are best for determining project effectiveness.” We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely.

Donaldson and Christie (2005) documented an important response to the AEA Statement from an influential group of senior members. This group opposed the AEA Statement, and did not feel they were appropriately consulted as active, long-term members of the association. Their response became known as “The Not AEA Statement.”

The Not AEA Statement:

(Posted on EvalTalk on December 3, 2003)

AEA members:

The statement below has been sent to the Department of Education in response to its proposal that “scientifically based evaluation methods” for assessing the effectiveness of educational interventions be defined as randomized experiments when they are feasible and as quasi-experimental or single-subject designs when they are not.

This statement is intended to support the Department’s definition and associated preference for the use of such designs for outcome evaluation when they are applicable. It is also intended to provide a counterpoint to the statement submitted by the AEA leadership as the Association’s position on this matter. The generalized opposition to use of experimental and quasi-experimental methods evinced in the AEA statement is unjustified, speciously argued, and represents neither the methodological norms in the evaluation field nor the views of the large segment of the AEA membership with significant experience conducting experimental and quasi-experimental evaluations of program effects.

We encourage all AEA members to communicate their views on this matter to the Department of Education and invite you to endorse the statement below in that communication if it is more representative of your views than the official AEA statement. (Comments can be sent to the Dept of Ed through Dec. 4 at comments@ed.gov with “Evaluation” in the subject line of the message).

This statement is in response to the Secretary's request for comment on the proposed priority on Scientifically Based Evaluation Methods. We offer the following observations in support of this priority.

The proposed priority identifies random assignment experimental designs as the methodological standard for what constitutes scientifically based evaluation methods for determining whether an intervention produces meaningful effects on students, teachers, parents, and others. The priority also recognizes that there are cases when random assignment is not feasible and, in such cases, identifies quasi-experimental designs and single-subject designs as alternatives that may be justified by the circumstances of particular evaluations.

This interpretation of what constitutes scientifically based evaluation strategies for assessing program effects is consistent with the presentations in the major textbooks in evaluation and with widely recognized methodological standards in the social and medical sciences. Randomized controlled trials have been essential to understanding what works, what does not work, and what is harmful among interventions in many other areas of public policy including health and medicine, mental health, criminal justice, employment, and welfare. Furthermore, attempts to draw conclusions about intervention effects based on nonrandomized trials have often led to misleading results in these fields and there is no reason to expect this to be untrue in the social and education fields. This is demonstrated, for example, by the results of randomized trials of facilitated communication for autistic children and prison visits for juvenile offenders, which reversed the conclusions of nonexperimental studies of these interventions.

Randomized trials in the social sector are more frequent and feasible than many critics acknowledge and their number is increasing. The Campbell Collaboration of Social, Psychological, Educational, and Criminological Trials Register includes nearly 13,000 such trials, and the development of this register is still in its youth.

At the same time, we recognize that randomized trials are not feasible or ethical at times. In such circumstances, quasi-experimental or other designs may be appropriate alternatives, as the proposed priority allows. However, it has been possible to configure practical and ethical experimental designs in such complex and sensitive areas of study as pregnancy prevention programs, police handling of domestic violence, and prevention of substance abuse. It is similarly possible to design randomized trials or strong quasi-experiments to be ethical and feasible for many educational programs. In such cases, we believe the Secretary's proposed priority gives proper guidance for attaining high methodological standards and we believe the nation's children deserve to have educational programs of demonstrated effectiveness as determined by the most scientifically credible methods available.

The individuals who have signed below in support of this statement are current or former members of the American Evaluation Association (AEA). Included among us are individuals who have been closely associated with that organization since its inception and who have served as AEA presidents, Board members, and journal editors. We wish to make clear that the statement submitted by AEA in response to this proposed priority does not represent our views and we regret that a statement representing the organization was proffered without prior review and comment by its members. We believe that the proposed priority will dramatically increase the amount of valid information for guiding the improvement of education throughout the nation. We appreciate the opportunity to comment on a matter of this importance and support the Department's initiative.

The subsequent exchanges about these statements on the AEA's electronic bulletin board, EvalTalk, seemed to generate much more heat than light and begged for more elaboration on the issues. As a result, Claremont Graduate University hosted and webcasted a debate for the applied research and evaluation community in 2004. The debate was between Mark Lipsey and Michael Scriven, and it attempted to sort out the issues at stake and to search for a common ground.

Donaldson and Christie (2005) concluded

somewhat surprisingly, that Lipsey and Scriven agreed that randomized control trials (RCTs) are the best method currently available for assessing program impact (causal effects of a program), and that determining program impact is a main requirement of contemporary program evaluation. However, Scriven argued that there are very few situations where RCTs can be successfully implemented in educational program evaluation, and that there are now good alternative designs for determining program effects. Lipsey disagreed and remained very skeptical of Scriven's claim that sound alternative methods exist for determining program effects, and challenged Scriven to provide specific examples. (P. 77)

What Counts as Credible Evidence?

In 2006, the debate about whether randomized control trials (RCTs) should be considered the gold standard for producing credible evidence in applied research and evaluation remained front and center across the applied research landscape. 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.

So 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 an experimental and nonexperimental perspective 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 us with 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 quasi-experiments 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 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 that 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

- Alkin, M. C. (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage Publications.
- Campbell, D. T. (1991). Methods for the experimenting society. *American Journal of Evaluation, 12*(3), 223-260.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research*. Chicago: Rand McNally.

Campbell Collaboration website: www.campbellcollaboration.org

Cochrane Collaboration website: www.cochrane.org

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

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. Mahwah, NJ: Lawrence Erlbaum.

Donaldson, S. I., & Christie, C. A. (2005). The 2004 Claremont Debate: Lipsey versus Scriven. Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? *Journal of Multidisciplinary Evaluation*, 3, 60–77.

Donaldson, S. I., & Christie, C. A. (2006). Emerging career opportunities in the transdiscipline of evaluation science. In S. I. Donaldson, D. E. Berger, & K. Pezdek (Eds.), *Applied psychology: New frontiers and rewarding careers*. Mahwah, NJ: Lawrence Erlbaum.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The handbook of evaluation: Policies, programs, and practices* (pp. 56–75). London: Sage Publications.

Donaldson, S. I., & Scriven, M. (2003). *Evaluating social programs and problems: Visions for the new millennium*. Mahwah, NJ: Lawrence Erlbaum.

Gallup. (2007). *The state of global well-being 2007*. New York: Gallup Press.

Joint Committee on Standards for Education Evaluation (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Newbury Park, CA: Sage Publications.

Julnes, G., & Rog, D. J. (2007). *Informing federal policies on evaluation methodology: building the evidence base for method choice in government-sponsored evaluations* (New Directions for Evaluation, No. 113). San Francisco: Jossey-Bass.

Mark, M. M. (2003). Toward an integrative view of the theory and practice of program and policy evaluation. In S. I. Donaldson & M. Scriven (2003.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 109–141). Mahwah, NJ: Lawrence Erlbaum.

Norcross, J. C., Beutler, L. E., & Levant, R. F. (2005). *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions*. Washington, DC: American Psychological Association.

- Pawson, R. (2006). *Evidence-based policy: A realist perspective*. London: Sage Publications.
- Pfeffer, J., & Sutton, R. I. (2006). *Hard facts, dangerous truths, and total nonsense: Profiting from evidence-based management*. Boston: Harvard Business School Press.
- Preskill, H., & Donaldson, S. I. (2008). Improving the evidence base for career development programs: Making use of the evaluation profession and positive psychology movement. *Advances in Developing Human Resources*, 10(1), 104–121.
- Reichhardt, C., & Rallis, C. S. (1994). *The qualitative-quantitative debate: New perspectives* (New Directions for Program Evaluation, No.87). San Francisco: Jossey-Bass.
- Sackett, D. L. (2000). *Evidence-based medicine: How to practice and teach EBM*. Edinburgh/New York: Churchill Livingstone.
- Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., & Haynes, R. B. (1996). Evidence-based medicine: What it is and what it isn't. *British Medical Journal*, 312, 71–72.
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage Publications.
- Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (1994). Guiding principles for evaluators: New directions for program evaluation #66. San Francisco: Jossey-Bass.
- Stober, D. R., & Grant, A. M. (2006). *Evidence-based coaching handbook: Putting best practice to work for your clients*. Hoboken, NJ: Wiley.
- 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.
- What Works Clearinghouse website: <http://ies.ed.gov/ncee/wwc>