1

Examining the Backbone of Contemporary Evaluation Practice

Credible and Actionable Evidence

Stewart I. Donaldson

The demand for rigorous and influential evaluations, and thus credible and actionable evidence, is at an all-time high across the globe. The most recent surge of activity has expanded well beyond the evaluation of traditional, large-scale government programs. Evaluations are now being conducted on a wide range of problems, programs, policies, practices, products, personnel, organizations, proposals, and the like across a diverse range of community, organizational, government, and international settings (Donaldson, 2013). While a systematic review of the notable developments related to evaluation practice since the first volume of this book in 2009 is beyond the scope of this chapter, I have selected a few of these developments to set the stage for this second edition.

Both the number and size of existing professional associations for practicing evaluators continue to grow rapidly. The largest national society, the American Evaluation Association (AEA), has grown to nearly 8,000 members, and the most recent annual meeting in Washington, D.C., set a new record with more than 3,500 delegates in attendance despite a U.S. government shutdown. A global grassroots movement to strengthen civil society's evaluation capacity worldwide, EvalPartners, has now identified more than 150 Voluntary Organizations of Professional Evaluators (VOPEs) boosting an aggregate total membership of more than 34,000 (Rugh & Segone, 2013). These VOPEs are not only focused on improving the supply of rigorous and influential evaluations, many are addressing the demand side by advocating for policies and systems that enable high-quality evaluation practice (Rugh & Segone, 2013). Furthermore, EvalPartners, the global movement to strengthen national evaluation capacities, announced that 2015 has been declared as the International Year of Evaluation (EvalYear; http://mymande.org/evalyear/Declaring_2015_as_the_International_Year_of_Evaluation).

This robust expansion of evaluation professional meetings and activities has been accompanied by increased opportunities for evaluation training and professional development. Recent research shows that universities across the globe are providing more evaluation degrees, certificates, courses, and professional development opportunities than ever before (LaVelle & Donaldson, in press). As the profession continues to mature, many practitioners are participating in VOPEs and annual professional meetings, engaging in professional development activities, and collaborating with one another in an effort to learn about emerging evaluation practices. Universities and VOPEs now offer evaluation practitioners a wide range of resources for improving practice, such as the latest books and journals, regular convenings, and a range of professional development opportunities, guiding principles, evaluation competencies, and evaluation standards.

In addition, there has been a rapid expansion of free online evaluation resources and professional development opportunities. For example, EvalPartners supports the My Monitoring & Evaluation project (My M&E; http://mymande.org), a website containing a massive number of evaluation resources designed to foster knowledge sharing and networking among practicing evaluators and evaluation students worldwide. This resource is a repository of free books and manuals, evaluation toolkits, webinars with leading evaluation experts, job announcements, training opportunities, and e-learning programs and certificates. The most recent free e-learning program on development evaluation had almost 13,000 participants from 172 countries (Segone & Donaldson, under review). The resources now available from the evaluation profession can greatly enhance a practitioner's ability to provide rigorous and influential evaluations. Before we begin our careful examination of the backbone of contemporary evaluation practice and credible and actionable evidence, I will briefly introduce several other important aspects of the profession of evaluation, namely evaluation theory, evaluation design and methods, and research on evaluation.

Evaluation Theory

Practitioners working in contemporary evaluation practice can benefit greatly from understanding how to use theory to enhance their practice. Donaldson and Lipsey (2006) have spelled out in some detail the different roles that different types of theory can play to improve contemporary evaluation practice. One of these theory forms is *evaluation theory*, which is largely prescriptive theory that "offers a set of rules, prescriptions, prohibitions, and guiding frameworks that specify what a good or proper evaluation is and how evaluation should be done" (Alkin, 2012). Evaluation theories are thus theories of evaluation practice that address such enduring themes as how to understand the nature of what we evaluate, how to assign value to programs and their performance, how to construct knowledge, and how to use the knowledge generated by evaluation (e.g., Alkin, 2012; Donaldson 2007; Donaldson & Scriven, 2003; Shadish, Cook, & Leviton, 1991).

In 1997, the president of the AEA, William Shadish, emphasized the vast importance of teaching practitioners how to benefit from and use evaluation theory to improve practice. His presidential address was entitled "Evaluation Theory Is Who We Are" and emphasized the following:

> All evaluators should know evaluation theory because it is central to our professional identity. It is what we talk about more than anything else, it seems to give rise to our most trenchant debates, it gives us the language we use for talking to ourselves and others, and perhaps most important, it is what makes us different from other professions. Especially in the latter regards, it is in our own self-interest to be explicit about this message, and to make evaluation theory the very core of our identity. Every profession needs a unique knowledge base. For us, evaluation theory is that knowledge base. (Shadish, 1998, p. 1)

Evaluation theories can also help us understand our quest as practitioners to gather credible and actionable evidence. They often take a stand on what counts as credible and actionable evidence in practice. However, evaluation theories today are rather diverse, and some are at odds with one another (see Mertens & Wilson, 2012). Understanding these differences between theories of practice is one way to help us understand disagreements about what counts as credible and actionable evidence.

In professional practice, it is vitally important that we are clear about our assumptions and purposes for conducting evaluation. Evaluation theory can help us make those decisions and help us understand why other evaluators might make different decisions in practice or criticize the decisions we have made about gathering credible and actionable evidence. In summary, being well-versed in contemporary theories of evaluation practice can enhance our ability to make sound choices about gathering evidence to answer key evaluation questions.

Program-theory–driven evaluation science is one of many examples of a theory of evaluation practice (Donaldson, 2007; Donaldson & Crano, 2011). This evaluation approach attempts to incorporate many of the hard-won lessons of evaluation practice over the past 30 years and to provide an evolving, integrative, and contingency-based theory of practice. Program-theory–driven evaluation science offers practitioners the following concise, three-step approach to practice:

1. Developing program impact theory
2. Formulating and prioritizing evaluation questions
3. Answering evaluation questions

Simply stated, evaluators work with stakeholders to develop a common understanding of how a program is presumed to solve the problem(s) of interest; to formulate and prioritize key evaluation questions; and then to decide how best to gather credible evidence to answer those questions within practical, time, and resource constraints.

This practical program evaluation approach is essentially method neutral within the broad domain of social science and evaluation methodology. The focus on the development of program theory and evaluation questions frees evaluators initially from having to presuppose the use of one evaluation design or another. The choice of the evaluation design and methods used to gather credible and actionable evidence is made in collaboration with the relevant stakeholders and is not solely decided by the evaluation team. The decisions about how best to go about collecting credible and actionable evidence to answer the key evaluation questions are typically thought to be contingent on the nature of the questions to be answered and the context of the setting. Stakeholders are provided with a wide range of choices for gathering credible and actionable evidence, which reinforces the idea that neither quantitative nor qualitative nor mixed-method designs are necessarily superior or applicable in every applied research and evaluation context (e.g., Chen, 1997). Whether an evaluator uses case studies, observational methods, structured or unstructured interviews, online or telephone survey research, a quasi-experiment, or a randomized controlled trial (RCT) to answer the key evaluation questions is dependent on discussions with relevant stakeholders about what would constitute credible and actionable evidence in this context and what is feasible given the practical, time, and financial constraints (Donaldson, 2007; Donaldson & Crano, 2011; Donaldson & Lipsey, 2006).

This practical approach for gathering credible and actionable evidence is highly consistent with the profession's guiding principles, evaluation standards, and other mainstream approaches to practical program evaluation (Chen, 2005; Chen, Donaldson, & Mark, 2011; Donaldson, 2007; Rossi, Lipsey, & Freeman, 2004; Weiss, 1998; Yarbrough, Shula, Hopson, & Caruthers, 2011). One of the best examples to date of program-theory–driven evaluation science in action is embodied in the Centers for Disease Control and Prevention's (2012) six-step Program Evaluation Framework. This framework is not only conceptually well developed and instructive for evaluation practitioners, it also has been widely adopted for evaluating federally funded public health programs throughout the United States. One of the six key steps in this framework is Step 4: Gather Credible Evidence. Step 4 is defined in the following way:

> Compiling information that stakeholders perceive as trustworthy and relevant for answering their questions. Such evidence can be experimental or observational, qualitative or quantitative, or it can include a mixture of methods. Adequate data might be available and easily accessed, or it might need to be defined and new data collected. Whether a body of evidence is credible to stakeholders might depend on such factors as how the questions were posed, sources of information, conditions of data collection, reliability of measurement, validity of interpretations, and quality control procedures.

Program-theory–driven evaluation science is just one of many forms of evaluation theory available today to help guide evaluation practice (see Mertens & Wilson, 2012, for a description of wide range of evaluation theories). It is summarized here to illustrate how evaluation theories offer guidance in terms of how to gather credible and actionable evidence in contemporary practice. It clearly specifies that there is not a universal answer to the question of what counts as credible and actionable evidence. Rather, it suggests the answer to this question in any particular evaluation context is contingent on the evaluation questions and choices made by the relevant stakeholders in the light of practical, time, and resource constraints. Other popular evaluation theories and approaches used to guide contemporary evaluation practice include utilization-focused evaluation (Patton, 2012), participatory evaluation (Cousins & Chouinard, 2012), empowerment evaluation (Fetterman, in press), experimental evaluation research (Bickman & Reich, Chapter 5; Henry, Chapter 4), the science of valuing (Scriven, 2013), realist evaluation (Mark, Henry, & Julnes, 2000; Pawson & Tilley, 1997), culturally responsive evaluation (Hood, Hopson, Obeidat, & Frierson, in press), feminist evaluation (Brisolara, Seigart, & SenGupta, 2014), transformative evaluation (Mertens, 2009), equity-focused evaluation (Bamberger & Segone, 2012), real-world evaluation (Bamberger, Rugh, & Mabry, 2006), values-engaged evaluation (Greene, 2005), and developmental evaluation and systems thinking (Patton, 2011), among many others (see Alkin, 2012; Mertens & Wilson, 2012).

Design and Methods

The decisions made in practice about evaluation design and methods can often be traced back to evaluation theory or at least a practitioner's assumptions and views about what constitutes good evaluation practice. Christie and Fleischer

(Chapter 2) discuss how assumptions about social inquiry and scientific paradigms seem to color views about which designs and methods provide the most credible and actionable evidence. What should be clear from the chapters in this volume is that contemporary practitioners now have a wide range of designs and methods to choose from when they are charged to gather credible and actionable evidence. The discussions throughout this volume provide more details about the strengths and limitations of these various designs and methods. These discussions illuminate ways that practitioners might use this knowledge to make informed decisions about which designs and methods to employ in practice.

Research on Evaluation

Theories of evaluation practice tend to be based more on evaluator experience than on systematic evidence of their effectiveness. That is, unlike social science theories used to help program and policy design, evaluation theories remain largely prescriptive and unverified. There has been a recent surge of interest in developing an evidence base to complement theory for guiding how best to practice evaluation (Cousins & Chouinard, 2012; Donaldson, 2007; Henry & Mark, 2003; Mark, 2003, 2007).

Although research on evaluation is an emerging area and a limited source of help for practitioners at the present time, there are now important works we can point to as exemplars for how research can improve the way we practice in the future. For example, there is a long tradition of research illuminating how to conduct evaluations so they are useful and have influence (Cousins, 2007; Cousins & Chouinard, 2012). Other recent studies examine the links between evaluation theory and practice (Alkin & Christie, 2005; Christie, 2003; Fitzpatrick, 2004), the development of evaluation practice competencies (Ghere, King, Stevahn, & Minnema, 2006), strategies for managing evaluation anxiety (Donaldson, Gooler, & Scriven, 2002) and improving the relationships between evaluators and stakeholders (Donaldson, 2001; Campbell & Mark, 2006), and the like. Furthermore, the AEA has recently supported the development of a new Topic Interest Group charged with expanding the evidence base for practice by promoting much more research on evaluation. All of these examples underscore the point that research on evaluation holds great promise for advancing our understanding of how best to practice evaluation in contemporary times in general and, more specifically, how best to gather credible and actionable evidence.

Debates About Credible and Actionable Evidence

Now that you have a brief overview of contemporary evaluation practice, it is time to focus on one the most fundamental issues facing evaluation practitioners today: How do evaluators gather credible and actionable evidence to answer the wide range of evaluation questions they face across diverse and highly variable contexts? Most would agree that the backbone of any empirical evaluation is the quality of the evidence that supports the evaluative conclusions. Throughout the history of professional evaluation, evaluation theorists, scholars, and practitioners have debated vigorously about what constitutes high-quality evaluation evidence. We will begin our journey into the depths of this fundamental issue by exploring the debates that set the stage for the first volume on *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* (Donaldson, Christie, & Mark, 2009).

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 hardheaded tests of bold social programs designed to improve society. The hardheaded tests he envisioned were called *randomized experiments* and 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, 2013; Donaldson & Crano, 2011; 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 and actionable 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 Mertens & Wilson, 2012).

The promise of an evidence-based society and the accelerating demand for credible and actionable evidence has led to 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 (Gersten & Hitchcock, 2009) 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 Google 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 + Warm 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 that 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 credible and actionable evidence. Unfortunately, seeking truth or agreement about what constitutes credible and actionable 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 and actionable evidence. Furthermore, contemporary evaluation practice rests firmly on the foundation of providing credible and actionable 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 become 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 Mertens envisioned a desirable future based on constructivist philosophy, and Mertens seemed to suggest that the traditional quantitative social science paradigm, specifically randomized experiments, was quite limited for evaluation practice (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" (p. 189). 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 AEA, prominent applied researchers and evaluators discussed this event as a move back to the "Dark Ages" (Donaldson & Christie, 2005). The leadership of the AEA developed a policy statement opposing these efforts to privilege randomized controlled 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 controlled 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

(Continued)

(Continued)

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, December 3, 2003; available at http://bama.ua.edu/archives/evaltalk.html)

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

(Continued)

(Continued)

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 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).

There have also been a plethora of disputes and debates about credible and actionable evidence outside of the United States. For example, the European Evaluation Society (EES, 2007) issued a statement in response to strong pressure from some interests advocating for "scientific" and "rigorous" impact of development aid, where this is defined as primarily involving RCTs:

> EES deplores one perspective currently being strongly advocated: that the best or only rigorous and scientific way of doing so is through randomised controlled trials (RCTs). In contrast, the EES supports multi-method approaches to IE [impact evaluation and assessment] and does not consider any single method such as RCTs as first choice or as the "gold standard."

This new statement briefly discusses the rationale for this perspective and lists examples of publications that consider a number of alternative approaches for establishing impact.

EES Statement:

The importance of a methodologically diverse approach to impact evaluation—specifically with respect to development aid and development interventions.

December 2007

(Continued)

(Continued)

The European Evaluation Society (EES), consistent with its mission to promote the "theory, practice and utilization of high quality evaluation," notes the current interest in improving impact evaluation and assessment (IE) with respect to development and development aid. EES however deplores one perspective currently being strongly advocated: that the best or only rigorous and scientific way of doing so is through randomised controlled trials (RCTs).

In contrast, the EES supports multi-method approaches to IE and does not consider any single method such as RCTs as first choice or as the "gold standard":

- The literature clearly documents how all methods and approaches have strengths and limitations and that there are a wide range of scientific, evidence-based, rigorous approaches to evaluation that have been used in varying contexts for assessing impact.
- IE is complex, particularly of multi-dimensional interventions such as many forms of development (e.g., capacity building, Global Budget Support, sectoral development) and consequently requires the use of a variety of different methods that can take into account rather than dismiss this inherent complexity.
- Evaluation standards and principles from across Europe and other parts of the world do not favor a specific approach or group of approaches—although they may require that the evaluator give reasons for selecting a particular evaluation design or combination.

RCTs represent one possible approach for establishing impact, that may be suitable in some situations, e.g.:

- With simple interventions where a linear relationship can be established between the intervention and an expected outcome that can be clearly defined;
- Where it is possible and where it makes sense to "control" for context and other intervening factors (e.g., where contexts are sufficiently comparable);
- When it can be anticipated that programmes under both experimental and control conditions can be expected to remain static (e.g., not attempt to make changes or improvements), often for a considerable period of time;
- Where it is possible and ethically appropriate to engage in randomization and to ensure the integrity of the differences between the experimental and control conditions.

Even in these circumstances, it would be "good practice" not to rely on one method but rather combine RCTs with other methods—and to triangulate the results obtained.

As with any other method, an RCT approach also has considerable limitations that may limit its applicability and ability to contribute to policy, e.g.:

- RCT designs are acknowledged even by many of its proponents to be weak in external validity (or generalisability), as well as in identifying the actual mechanisms that may be responsible for differences in outcomes between the experimental and control situations;
- "Scaling up," across-the-board implementation based upon the results of a limited and closely controlled pilot situation, can be appropriate for those interventions (e.g., drug trials) where the conditions of implementation would be the same as in the trial, but this is rarely the case for most socio-economic interventions where policy or program "fidelity" cannot be taken for granted;
- An RCT approach is rarely appropriate in complex situations where an outcome arises from interaction of multiple factors and interventions, and where it makes little sense to "control" for these other factors. In a development context, as for most complex policy interventions, outcomes are the result of multiple factors interacting simultaneously, rather than of a single "cause";
- RCTs are limited in their ability to deal with emergent and/or unintended and unanticipated outcomes as is increasingly recognized in complexity and systems research—many positive benefits of development interventions will often be related rather than identical to those anticipated at the policy/program design stage;
- RCTs generally are less suited than other approaches in identifying what works for whom and under what circumstances. Identifying what mechanisms lead to an identified change is particularly important given the varying contexts under which development typically takes place and is essential for making evidence-based improvements.

We also note that RCTs are based upon a successionist (sometimes referred to as "factual") model of causality that neglects the links between intervention and impact and ignores other well-understood scientific means of establishing causality, e.g.:

- Both the natural and social sciences (e.g., physics, astronomy, economics) recognize other forms of causality, such as generative (sometimes referred to as "physical") causality that involve identifying the

(Continued)

(Continued)

underlying processes that lead to a change. An important variant of generative causality is known as the modus operandi that involves tracing the "signature," where one can trace an observable chain of events that links to the impact.

- Other forms of causality recognize simultaneous and/or alternative causal strands, e.g., acknowledging that some factors may be necessary but not sufficient to bring about a given result, or that an intervention could work through one or more causal paths. In non-linear relationships, sometimes a small additional effort can serve as a "tipping point" and have a disproportionately large effect.
- Some research literature questions whether simple "causality" (vs. "contribution" or "reasonable attribution") is always the right approach, given the complexity of factors that necessarily interact in contemporary policy—many of them in specific contexts.

EES also notes that in the context of the Paris Declaration, it is appropriate for the international evaluation community to work together in supporting the enhancement of development partner capacity to undertake IE. Mandating a specific approach could undermine the spirit of the Paris Declaration and as the literature on evaluation utilization has demonstrated, limit buy-in and support for evaluation and for subsequent action.

In conclusion, EES welcomes the increased attention and funding for improving IE, provided that this takes a multi-method approach drawing from the rich diversity of existing frameworks and one that engages both the developed and developing world. We would be pleased to join with others in participating in this endeavour. (European Evaluation Society, 2007)

What Counts as Credible Evidence?

In 2006, the debate about whether 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), and 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.edu/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). The volume based on this symposium, *What Counts as Credible Evidence in Applied Research and Evaluation,* was published in 2009.

The Quest for Credible and Actionable Evidence

SAGE staff contacted us in 2012, suggesting that there was great interest in a second edition. After consulting with many of the original authors and exploring the interest of a couple of new authors, we decided to move forward with the revision and to broaden the scope to *Credible and Actionable Evidence: The Foundations for Rigorous and Influential Evaluations.* The original authors were asked to revise their chapters to take into account new developments in the field and in their thinking as well as this somewhat broader focus. Robin Miller and Eleanor Chelimsky were invited to contribute new chapters.

Our search for a deeper and more complete understanding of credible and actionable 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 the debates about credible and actionable evidence. Ontological, epistemological, and methodological assumptions that anchor views about the nature of credible and actionable evidence are explored. The third and final chapter in Part I is a new chapter by Robin Miller on the individual psychological processes that can affect judgments of credibility. Miller underscores that potential users often rely on various peripheral cues or heuristics to judge the credibility of evaluation evidence and that there are important factors beyond evaluation design and methods that determine if evaluation evidence is considered credible and actionable. These introductory chapters broaden the discussion and preview the positions expressed about credible and actionable evidence in the subsequent sections of the book.

Part II will explore the role of randomized experiments in producing credible and actionable evidence. Gary Henry (Chapter 4) and Leonard Bickman and Stephanie Reich (Chapter 5) explore the value of randomized controlled trials and quasi-experimental designs, while Michael Scriven (Chapter 6) questions the value they add to everyday evaluation practice. Part III explores the value of other evaluation designs and methods for producing credible and actionable evidence with chapters from Sharon Rallis (Chapter 7), Sandra Mathison (Chapter 8), and Eleanor Chelimsky (Chapter 9). Part IV provides general perspectives on credible and actionable evidence with chapters from Jennifer Greene (Chapter 10), George Julnes and Debra Rog (Chapter 11), and Thomas Schwandt (Chapter 12). Finally, Melvin Mark (Chapter 13) ends Part IV and the book with a closing chapter that reviews the central themes about credible and actionable evidence presented throughout the book and a sketch of a broader framework surrounding judgements of credibility and actionability and provides a set of recommendations for improving evaluation practice based on the framework and insights provided by the chapter authors. The chapters in this volume were written to encourage and inspire you to reflect deeply about ways to gather credible and actionable evidence in your evaluation practice and to help you provide rigorous and influential evaluations for the promotion of social betterment worldwide.

References

Alkin, M. C. (Ed.). (2012). *Evaluation roots: A wider perspective of theorists' views and influences*. Thousand Oaks, CA: SAGE.

Alkin, M. C., & Christie, C. A. (2005). Theorists' models in action [Entire issue]. *New Directions for Evaluation, 106*.

Bamberger, M. J., Rugh, J., & Mabry, L. S. (2006). *Real world evaluation: Working under budget, time, data, and political constraints*. Newbury Park, CA: SAGE.

Bamberger, M., & Segone, M. (2012). *How to design and manage equity-focused evaluations*. New York, NY: UNICEF.

Brisolara, S., Seigart, D., & SenGupta, S. (2014). *Feminist evaluation and research: Theory and practice*. New York, NY: Guildford.

Campbell, B., & Mark, M. M. (2006). Toward more effective stakeholder dialogue: Applying theories of negotiation to policy and program evaluation. *Journal of Applied Social Psychology, 36*(12), 2834–2863.

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, IL: Rand McNally.

Centers for Disease Control and Prevention. (2012). *Framework for program evaluation in public health*. MMWR, 48 (No. RR-11). CDC Evaluation Working Group.

Chen, H. T. (1997). Applying mixed methods under the framework of theory-driven evaluations. In J. Greene & V. Caracelli (Eds.), Advances in mixed methods evaluation: The challenge and benefits of integrating diverse paradigms. *New Directions for Evaluation, 74*, 61–72.

Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: SAGE.

Chen, H. T., Donaldson, S. I., & Mark, M. M. (Eds.). (2011). Advancing validity in outcome evaluation: Theory and practice [Entire issue]. *New Directions for Evaluation, 130*.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), The practice–theory relationship in evaluation. *New Directions for Evaluation, 97,* 1–35.

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

Cousins, J. B. (Ed.). (2007). Process use in theory, research, and practice [Entire issue]. *New Directions for Evaluation, 116*.

Cousins, J. B., & Chouinard, J. A. (Eds.). (2012). *Participatory evaluation up close: A integration of research-based knowledge*. Greenwich, CT: Information Age.

Donaldson, S. I. (2001). Overcoming our negative reputation: Evaluation becomes known as a helping profession. *American Journal of Evaluation, 22*(3), 355–361.

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

Donaldson, S. I. (2013). *The future of evaluation in society: A tribute to Michael Scriven*. Greenwich, CT: Information Age.

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., & Mark, M. M. (2009). *What counts as credible evidence in applied research and evaluation practice?* Newbury Park, CA: SAGE.

Donaldson, S. I., & Crano, W. C. (2011). Theory-driven evaluation science and applied social psychology: Exploring the intersection. In M. M. Mark, S. I. Donaldson, & B. Campbell (Eds.), *Social psychology and evaluation* (pp. 141–160). New York, NY: Guilford.

Donaldson, S. I., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation, 23*(3), 261–273.

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

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

European Evaluation Society. (2007). *EES Statement: The importance of a methodologically diverse approach to impact evaluation—specifically with respect to development aid and development interventions.* Retrieved February 7, 2008, from http://www.europeanevaluation.org/download/?id=1969403.

Fetterman, D. (in press). *Empowerment evaluation*. Newbury Park, CA: SAGE.

Fitzpatrick, J. L. (2004). Exemplars as case studies: Reflections on the links between theory, practice, and context. *American Journal of Evaluation, 25*(4), 541–559.

Gersten, R., & Hitchcock, J. (2009). What is credible evidence in education? The role of the What Works Clearinghouse in informing the process. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 78–95). Newbury Park, CA: SAGE.

Ghere, G., King, J. A., Stevahn, L., & Minnema, J. (2006). A professional development unit for reflecting on program evaluator competencies. *American Journal of Evaluation, 27*(1), 108–123.

Greene, J. (2005). A value-engaged approach for evaluating the Bunche–Da Vinci Learning Academy. In M. C. Alkin, & C. A. Christie (Eds.), Theorists' models in action, *New Directions for Evaluation*, *106*, 27–45.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.), The practice–theory relationship in evaluation. *New Directions for Evaluation, 97*, 69–80.

Hood, S., Hopson, R. K., Obeidat, K., & Frierson, H. (in press). *Continuing the journey to reposition culture and cultural context in evaluation theory and practice*. Greenwich, CT: Information Age.

Julnes, G., & Rog, D. J. (Eds.). (2007). Informing federal policies on evaluation methodology: Building the evidence base for method choice in government-sponsored evaluations [Entire issue]. *New Directions for Evaluation, 113*.

LaVelle, J., & Donaldson, S. I. (in press). The state of evaluation education and training. In J. W. Altschuld & M. Engle (Eds.), Accreditation, certification, and credentialing: Whither goes the American Evaluation Association. *New Directions for Evaluation*.

Mark, M. M. (2003). Toward an integrated view of the theory and practice of program and policy evaluation. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 183–204). Mahwah, NJ: Erlbaum.

Mark, M. M. (2007). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111–134). New York, NY: Guilford Press.

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

Mertens, D. (2009). *Transformative research and evaluation*. New York, NY: Guildford.

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice*. New York, NY: Guildford.

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.

Patton M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: Guildford.

Patton, M. Q. (2012). *Essentials of utilization focused evaluation*. London, England: SAGE.

Pawson, R. (2006). *Evidence-based policy: A realist perspective*. Thousand Oaks, CA: SAGE.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: SAGE.

Pfeffer, J., & Sutton, R. I. (2006). *Hard facts, dangerous truths, and total nonsense: Profiting from evidence-based management.* Boston, MA: 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. S., & Rallis, S. F. (1994). The qualitative–quantitative debate: New perspectives [Entire issue]. *New Directions for Program Evaluation, 61.*

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: SAGE.

Rugh, J., & Segone, M. (Eds.). (2013). *Voluntary organizations for professional evaluation: Learning from Africa, Americas, Asia, Australasia, Europe and Middle East*. UNICEF, IOCE and EvalPartners.

Sackett, D. L. (2000). *Evidence-based medicine: How to practice and teach EBM*. New York, NY: 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.

Scriven, M. (2013). The foundation and future of evaluation. In S. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven*. Greenwich, CT: Information Age.

Segone, M., & Donaldson, S. I. (in press). *What have we learned in using social media in the international development context? The case of My M&E, the platform of a global partnership to strengthen national evaluation capacities.* Under review.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1–19.

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

Stober, D. R., & Grant, A. M. (2006). *Evidence-based coaching handbook: Putting best practice to work for your clients*. Hoboken, NJ: Wiley.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Yarbrough, D. B., Shula, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users.* Newbury Park, CA: SAGE.