What Counts as Credible Evidence in Applied Research and Evaluation Practice?

Stewart I. Donaldson
Claremont Graduate University

Christina A. Christie
Claremont Graduate University

Melvin M. Mark
Pennsylvania State University

Editors

SAGE
Los Angeles, London, New Delhi, Singapore

For information:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks,
California 91320
E-mail: order@sagepub.com

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative
Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications
Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

What counts as credible evidence in applied research and evaluation practice? / Stewart I. Donaldson, Christina A. Christie, [and] Melvin M. Mark, editors.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4129-5707-6 (pbk.: acid-free paper)
1. Research—Evaluation. 2. Research—Social aspects. 3. Social sciences—Methodology. 4. Evaluation research (Social action programs) 5. Educational accountability. I. Donaldson, Stewart I. (Stewart Ian) II. Christie, Christina A. III. Mark, Melvin M.

Q180.55.E9W53 2009
001.4′2—dc22 2008017667

This book is printed on acid-free paper.

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Vicki Knight
Editorial Assistant: Lauren Habib
Production Editor: Catherine M. Chilton
Copy Editor: Teresa Herlinger
Typesetter: C&M Digitals (P) Ltd.
Proofreader:
Indexer:
Cover Designer: Candice Harman
Marketing Manager: Stephanie Adams

Epilogue

A Practitioner's Guide for Gathering Credible Evidence in the Evidence-Based Global Society

Stewart I. Donaldson

What do the discussions and debates about credible evidence offer practitioners? I will attempt to answer this question for you by briefly providing some observations about the diversity and changing nature of the enterprise, by offering some lessons from my own applied research and evaluation practice, and by discussing how practitioners might address some of the key issues and challenges of collecting credible evidence raised throughout the chapters in this volume.

Understanding Practice Today

In Chapter 1, I discussed indicators that suggest the demand for credible evidence is at an all-time high across the globe, and that applied research and evaluation practice is booming. The recent surge of activity has expanded well beyond the evaluation of traditional, large-scale, government programs. Evaluation and applied research are being conducted on a much wider range of problems, programs, policies, practices, products, personnel, organizations, proposals, and the like across a diverse range of community, organizational, government, and global settings. Practitioners today are confronted with a shifting landscape of applied research and evaluation practice.

I have suggested that this changing landscape has set the stage for a similar but broader perspective on evidence than D. T. Campbell's (1991) earlier vision of an "experimenting society." This new vision of an "evidence-based global society" promotes the search for and thoughtful reflection about evidence collected from a wide range of designs and methods to answer key applied research and evaluation questions. While the gold standard in the evidence-based society might be thought of as "methodological appropriateness" (Mark, Chapter 12; Schwandt, Chapter 11), it is important to underscore that RCTs will play a very important role and will be considered the gold standard for answering some questions under specific circumstances, as will other research and evaluation approaches, designs, and methods when the conditions are optimal for their use.

It is most likely a healthy sign for the intellectual development of the field if the debates rage on about credible evidence, gold standards, and optimal research and evaluation paradigms and designs. Our hope should be that these discussions in the scholarly literature, perhaps reframed as Mark suggests (Chapter 12), will inspire research and inquiries that someday provide practitioners with more evidence-based guidance about how best to practice applied research and evaluation. But in the meantime, there is critically important work to be done and this natural uncertainty in a young field should not slow us down in our pursuit to help solve the important problems of the day. As we practice applied research and evaluation, let me suggest that we as practitioners draw on the body of knowledge about practice we have now accumulated. Below, I briefly review the value of our knowledge base with respect to evaluation theory, design and methods, the profession, and research on evaluation.

Evaluation Theory

Practitioners working in the new "evidence-based global society" 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 applied research and 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, 2004). 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, 2004; Donaldson & Scriven, 2003; Shadish, Cook, & Leviton, 1991).

In 1997, the president of the American Evaluation Association, 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 evidence. They often take a stand on what counts as credible evidence in practice. However, evaluation theories today are rather diverse and some are at odds with one another (see Donaldson & Scriven, 2003). Understanding these differences between theories of practice can help us understand disagreements about what counts as credible evidence.

In professional practice, it is vitally important that we are clear about our assumptions and purposes for conducting applied research and evaluation. Evaluation theory can help us make those decisions, and help us understand why other applied researchers and evaluators might make different decisions in practice, or criticize the decisions we have made about gathering credible 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 the key applied research and evaluation questions we are being paid to address.

Design and Methods

The decisions made in practice about research and 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) discussed how assumptions about social inquiry and scientific paradigms seem to color views about which designs and methods provide the most credible 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 evidence. The discussions throughout the previous chapters take us to a new level of understanding about the strengths and limitations of these various designs and methods. I will discuss in more detail later the ways that a practitioner can use this knowledge to make decisions about which designs and methods to employ in practice.

The Profession

The profession of applied research and evaluation is maturing. As was discussed in Chapter 1, more practitioners than ever before are participating in annual professional meetings, going to professional development activities, and collaborating with one another in an effort to learn about best practices. The number of professional associations has grown dramatically throughout the world in the past 15 years, and the size of the more established associations has increased substantially. These organizations now offer us as practitioners a wide range of resources for improving our practice such as the latest books and journals, regular convenings, a range of professional development opportunities, guiding principles, and evaluation standards. Donaldson and Christie (2006) describe the emerging transdiscipline and profession in some detail, and provide examples of the broad array of career opportunities that now exist in applied research and evaluation. The resources now available from the profession can greatly enhance a practitioner's ability to gather credible evidence and to provide accurate and useful applied research and evaluations.

Research on Evaluation

Theories of evaluation practice tend to be based more on philosophy and 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 applied research and evaluation (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). 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), improving the relationships between evaluators and stakeholders (Donaldson, 2001; B. Campbell & Mark, 2006), and the like. Along these same lines, Schwandt (Chapter 11, this volume) proposes an agenda of inquiry that might lead to the development of a practical theory of how to gather credible evidence in applied research and evaluation. Furthermore, the American Evaluation Association 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 in contemporary times in general, and more specifically how best to gather credible evidence.

Program Theory–Driven Evaluation Science

I have recently provided a framework and detailed examples of how to gather credible evidence in contemporary, practical program evaluations (Donaldson, 2007). This framework attempts to incorporate many of the hard-won lessons in applied research and 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 methodology. The focus on the development of program theory and evaluation questions frees evaluators initially from having to presuppose use of one evaluation design or another.

The choice of the evaluation design and methods used to gather credible 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 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 evidence, which reinforces the idea that neither quantitative, 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 an RCT to answer the key evaluation questions is dependent on discussions with relevant stakeholders about what would constitute credible evidence in this context, and what is feasible given the practical, time, and financial constraints (Donaldson, 2007; Donaldson & Lipsey, 2006).

This practical approach for gathering credible evidence is highly consistent with the profession's guiding principles, evaluation standards, and other mainstream approaches to practical program evaluation (Chen, 2005; Rossi, Lipsey, & Freeman, 2004; Weiss, 1998). One of the best examples to date of program theory–driven evaluation science in action is embodied in the Centers for Disease Control's (1999) 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 Donaldson & Scriven, 2003). It is summarized here to illustrate how evaluation theories do offer guidance in terms of how to gather credible evidence in contemporary practice. This particular perspective remains open to the

wide range of experimental and nonexperimental views about what counts as credible evidence that have been expressed throughout this volume.

However, it clearly specifies that there is not a universal answer to the question of what counts as credible evidence. Rather, 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.

Some Final Thoughts About the Debates on Credible Evidence

The chapter authors in Section II presented us with some very compelling arguments about the strengths of RCTs and experimental approaches as the route to credible evidence, particularly when the stakes are high for determining "what works." It is laudable that these chapters also explored the limitations and boundary conditions of the experimental approach to gathering credible evidence. The chapters in Section III do an excellent job of extending our understanding of the limitations of the experimental approach, and highlighting the problems of considering the RCT as a universal gold standard for gathering credible evidence. These chapters also provide us with a much deeper understanding of the increasing range of nonexperimental approaches to gathering credible evidence, and the broad array of evaluation questions that can be answered in applied research and evaluation, moving us way beyond the common question of "what works."

However, as Mark points out in Chapter 12, it should be underscored that the chapters in this volume demonstrate that the topic of "what counts as credible evidence" is not one where consensus exists at this time in our history. In fact, while the volume has been in production there has been a plethora of new disputes and debates about credible evidence across the emerging evidence-based global society. For example, the European Evaluation Society (EES, 2007) recently 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 multimethod approaches to IE 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.

Please note the similarities of the debate now going on in Europe and other parts of the world, with the North American version as expressed in the AEA Statement and Not AEA Statement displayed in Chapter 1.

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

December 2007

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.:

(Continued)

(Continued)

- 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 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)

Conclusion: So What Counts as Credible Evidence?

It depends! The chapters in this volume have shown us that the answer to this question depends on characteristics such as

- The question(s) of interest
- The context
- Assumptions made by the evaluators and stakeholders

- The evaluation theory used to guide practice
- Practical, time, and resource constraints

Let these subtle nuances help you answer that question clearly in the context of your own evaluation practice.

Stated another way, I think it is safe to say we are a long way from consensus and a universal answer to the question of what counts as credible evidence in contemporary applied research and evaluation. However, the very rich discussions throughout this volume and the debates across the globe are providing us with ideas for what might go into an early draft of a blueprint for the emerging "evidence-based global society." It now seems clear that disparate views about what constitutes credible evidence are predicated on divergent assumptions, often implicit (Mark, Chapter 12). Practitioners should be well advised to explore these assumptions in great depth with their stakeholders before embarking on their quest to gather credible evidence. A full exploration of the strengths and weaknesses for the range of designs and methods available for any given applied research and evaluation problem seems warranted. A goal to strive for is to fully inform the key stakeholders of the benefits and costs of the wide range of approaches available to answer their key applied research and evaluation questions before they embark with you on the journey of gathering credible evidence (Donaldson, 2007).

A potential unintended positive consequence of disagreements about how to best gather evidence is the interest being sparked to support more serious research and scholarship focused on advancing the understanding of how best to practice applied research and evaluation. Practitioners now have a wealth of resources to help guide their practice such as evaluation theory, a range of professional development opportunities, many more design and method options, and the promise of "research on evaluation" sorting out some of the more intractable problems preventing good practice. This volume has offered you a plethora of ideas and new directions to explore as you continue your pursuits of gathering credible evidence. It is my hope that this volume will inspire you to be much more reflective about your practice and the range of designs and methods that might be feasible to help you answer the pressing applied research and evaluation questions you encounter. In the end, great minds may disagree that there is a best road to travel to get to credible evidence, but they do seem to come together in consensus around the view that applied research and evaluation are critical activities for improving societies and the human condition as we venture into the unknowns of the 21st-century evidence-based global society.

References

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage Publications.

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

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.

Centers for Disease Control and Prevention. (1999). 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* (pp. 61–72). (Vol. 74 of *New Directions for Evaluation* series). San Francisco: Jossey-Bass.

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

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* (pp. 1–35). (Vol. 97 of *New Directions for Evaluation* series). San Francisco: Jossey-Bass.

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

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. (2005). Using program theory–driven evaluation science to crack the Da Vinci code. In M. Alkin & C. Christie (Eds.), *Theorists' models in action* (pp. 65–84). (Volume 106 of *New Directions* for Evaluation series). San Francisco: Jossey-Bass.

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

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* (pp. 243–260). Mahwah, NJ: Erlbaum.

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: Sage Publications.

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.

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.

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.

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* (pp. 69–80). (Vol. 97 of *New Directions for Evaluation* series). San Francisco: Jossey-Bass.

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: Guilford Press.

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

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 Publications.

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