

American Journal of Evaluation

<http://aje.sagepub.com>

Evaluation Theory is Who We Are


William R. Shadish

American Journal of Evaluation 1998; 19; 1

DOI: 10.1177/109821409801900102

The online version of this article can be found at:
<http://aje.sagepub.com/cgi/content/abstract/19/1/1>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

On behalf of:

American Evaluation Association

Additional services and information for *American Journal of Evaluation* can be found at:

Email Alerts: <http://aje.sagepub.com/cgi/alerts>

Subscriptions: <http://aje.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

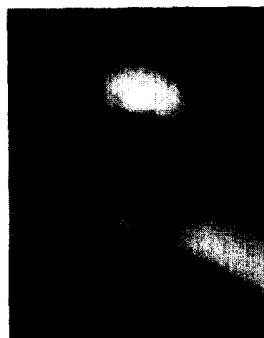
Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Evaluation Theory is Who We Are

WILLIAM R. SHADISH

ABSTRACT

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 regard, it is in our own self-interest to be explicit about this message, and to make evaluation theory the very heart of our identity. Every profession needs a unique knowledge base. For us, evaluation theory is that knowledge base.



William R. Shadish

INTRODUCTION

Perhaps because evaluation is primarily a practice-driven field, evaluation theory is not a topic that causes evaluators' hearts to flutter with excitement. Well, in this article I am going to try to convince you that evaluation theory is not only one of the most exciting topics in our field, but in many important respects, it literally *is* our field. As much or more than any other feature of evaluation, evaluation theory reveals who we really are.

I am using the phrase *evaluation theory* in a very general sense here to refer to a whole host of more or less theoretical writings with evaluation practice as their primary focus. My usage is general in the same way that the phrase "feminist theory" broadly refers to a diverse group of theories that do not necessarily share the *same* feminist theory (e.g., Olesen, 1994), or in the same way that the phrase "atomic theory" refers to any of several different theories of the structure of the atom, from the ones first broached thousands of years ago by Greek philosophers to those used today in fields like modern chemistry. Scriven (1991) gives a similarly broad description of the phrase *evaluation theory* in the fourth edition of his *Evaluation Thesaurus*. He says:

William R. Shadish • Department of Psychology, The University of Memphis, Memphis TN 38152; 901-678-4687 Tel: 901-678-2579; Fax: shadish@mail.psy.memphis.edu.

American Journal of Evaluation, Vol. 19, No. 1, 1998, pp. 1-19.
ISSN: 1098-2140

Copyright © 1998 by JAI Press Inc.
All rights of reproduction in any form reserved.

Some evaluation theories are theories about evaluation in a particular field (e.g., theories of program evaluation such as the discrepancy model ['local theories']). Some are about evaluation in general (e.g., theories about its political role or logical nature ['general theories'])....General theories include a wide range of efforts from the logic of evaluative discourse—general accounts of the nature of evaluation and how evaluations can be justified (axiology)—through metamethodology to sociopolitical theories of its role in particular types of environments, to 'models' that are often simply metaphors for, conceptualizations of, or procedural paradigms for evaluation. (Scriven, 1991, pp. 155-156).

I add to Scriven's observations that parts of evaluation theory are empirically based, but especially given the youth of our field, much more of it is hypothetical, conjectural, and unproven, still waiting to be tested. With few exceptions, evaluation theory is neither concise nor axiomatic; and it is not a single theory but rather a set of diverse theoretical writings held together by the common glue of having evaluation practice as their target.

Scriven's description of evaluation theory also refers to the diverse *topics* covered by evaluation theory, including not only the nature of evaluative discourse but also sociopolitics

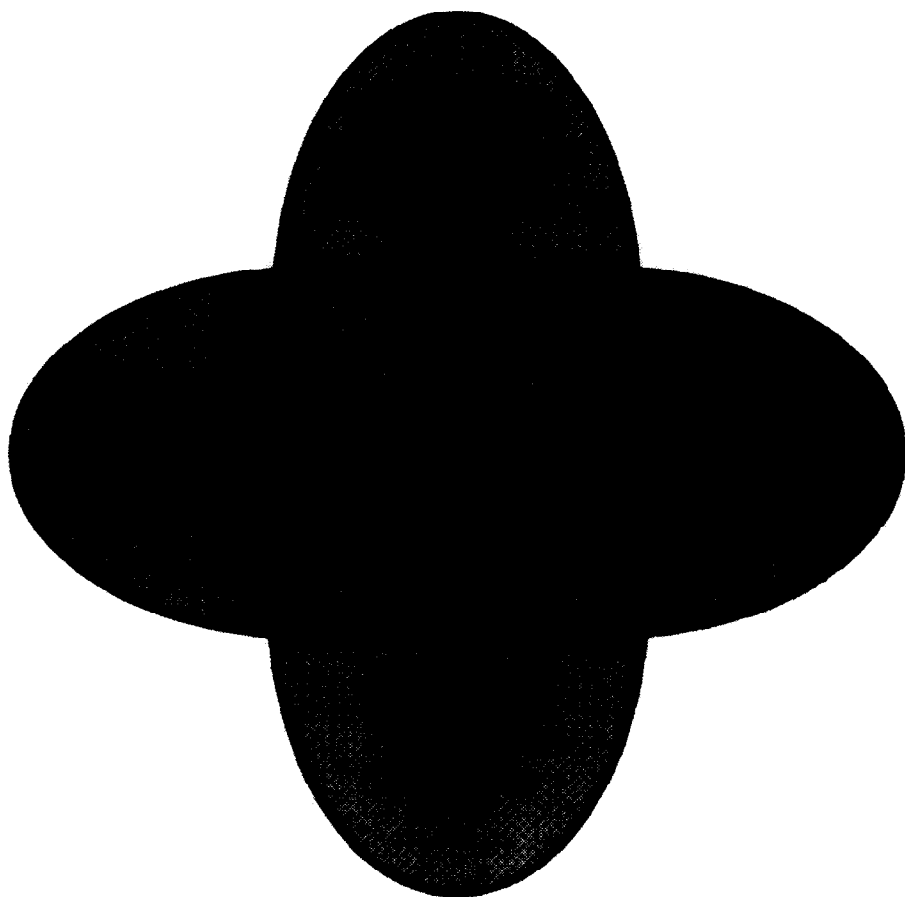


Figure 1.

and methodology. Figure 1 helps to portray this topical diversity¹. In the center of the figure is evaluation practice, the topic that ultimately drives most theoretical discussions, and the only one that is absolutely essential to evaluation theory. Converging on that center are a host of enduring themes concerning valuing, knowledge construction, knowledge use, and the nature of the evaluand (the thing we are evaluating). In various guises, these themes constantly reoccur in discussions of evaluation practice, and therefore they also become the stuff of evaluation theory. Elsewhere we have outlined these themes in more detail, showed why evaluators need to have information about them, and analyzed how well various evaluation theories have responded to that need (Shadish, 1994; Shadish, Cook, & Leviton, 1991).

WHY EVALUATION THEORY IS “WHO WE ARE”

With this as background, I will now examine some of the many ways in which evaluation theory is “who we are”. First, evaluation theory is “who we are” because it provides the language that we use to talk to each other about evaluation. Over the years, evaluation theorists have given us a rich vocabulary that includes such phrases as formative and summative evaluation, instrumental and conceptual use, enlightenment, cluster evaluation, multicultural validity, realist evaluation, and so on. While we all rightly hope that unnecessary jargon can be kept to a minimum, some specialized vocabulary like this is inevitable in any specialized science. The extent of this specialized vocabulary in evaluation can be seen in Scriven’s (1991) massive *Evaluation Thesaurus*; the 1991 fourth edition has almost 400 pages and hundreds if not thousands of entries. Knowing Scriven, the thesaurus is even larger today. Of course, I am not claiming that the thesaurus is itself a theory, only that it is a handy source to use to get to know the vocabulary that has been developed by evaluation theorists over the years. If you want to learn to talk the talk, there is no better source.

Second, evaluation theory is “who we are” in the sense that it encompasses many of the things in our field about which we seem to care most deeply. For example, I would bet that if most AEA members were asked to identify the single most hotly debated, deeply dividing issue in evaluation in the last 20 years, they would identify the quantitative-qualitative debates. To be sure, those debates certainly involved methodology, especially because many evaluators needed to learn about qualitative methods with which they had little real familiarity. Yet those debates were, in essence, mostly about theory—about epistemology and ontology, about what assumptions we make when we construct knowledge, about the nature of many fundamental concepts that we use in our work like causation, generalization, and truth, and about what those theoretical concerns imply for the practice of evaluation. Even on less hotly debated topics such as performance monitoring, we seem to be inclined to discuss theoretically-driven matters like whether performance monitoring corrupts the index that is being monitored. This does not mean that the methodological issues in performance monitoring are not important; it’s just that those matters are often worked out more completely and more rapidly than the underlying theoretical debates, so it is those theoretical debates that tend to receive sustained attention from us.

Third, evaluation theory is “who we are” in the sense that theory has defined the themes of the majority of AEA conferences. Table 1 lists those themes, which I have sorted into three categories: (1) *Evaluation Theory* (marked with a Θ , in boldface italics): themes were primarily theoretical in emphasizing the relationship of evaluation to such topics as social justice, empowerment, public policy, politics, utilization, and the synthesis of quantitative and quali-

TABLE 1
Presidential Themes (and Presidents) of Past American Evaluation Association Conferences

Themes and Presidents

- 1997: "Theory and Practice: Partners in Evaluation", (Θ) (William Shadish)**
 1996: "A Decade of Progress: Looking Back and Looking Forward" (S) (Len Bickman)
 1995: "Evaluation for a New Century: A Global Perspective" (T) (Eleanor Chelimsky)
1994: "Evaluation and Social Justice" (Θ) (Karen Kirkhart)
1993: "Empowerment Evaluation" (Θ) (David Fetterman)
1992: "Synthesizing Evaluation: Perspectives, Practices, and Evidence" (Θ) (David Cordray)
 1991: "New Evaluation Horizons" (T) (Lee Sechrest)
1990: "Evaluation and the Formulation of Public Policy" (Θ) (Yvonna Lincoln)
 1989: "International and Cross-Cultural Perspectives on Evaluation" (T) (Ross Conner)
1988: "Evaluation and Politics" (Θ) (Michael Patton)
1987: "Utilizing Evaluation Results" (Θ) (Bob Covert)
 1986: "What Have We Learned?" (S) (Richard Light)
-

Note: Each theme is categorized as theoretical (Θ), topical (T), or summing up (S). Plausibly theoretical themes are highlighted in boldface italic.

tative approaches, (2) *Summing Up* (marked with an S): efforts to review what we have learned and where we are going, which often included summing up what we know about theory (e.g., 1996); and (3) *Topical* (marked with an T): themes that focused mostly on *things* that we evaluate that pose particular challenges to the field, frequently involving substantially theoretical discussion of those topics. It is admittedly difficult to categorize such themes, and my judgments clearly rely as much on my personal experiences at each of those conferences as they do on subsequent published materials. Nonetheless, it is probably fair to say that more than half of these themes of the last 12 years have been directly on some aspect of evaluation theory, while the others were theoretical in substantial part. To the extent that we define ourselves by the things we talk about when evaluators get together, we talk a lot about evaluation theory.

Fourth, evaluation theory is "who we are" in the sense that it provides evaluators with an identity that is different from the identity of other professionals. For example, even though some evaluators may do randomized experiments, we know very well that doing randomized experiments is not at all the mark of being an evaluator. Such experiments are also done by psychologists, educators, economists, biologists, engineers, and a whole host of other professionals. If we want to be different from them, we can hardly appeal to the experiment. Similarly, the fact that we debate the values that are implicit in our work does not make us unique, for values are also debated by philosophers, medical ethicists, and even politicians. What makes us unique is the peculiar and unique conjunction of issues that bears on evaluation practice in the manner portrayed in Figure 1. Ask yourself this: At what other conference would you find a list of conference themes that matches the particular set of themes seen at AEA over the years—where such diverse themes as social justice, information use, public policy, politics, and epistemology were all seen as so central to the mission of the field that an entire conference could be devoted to each of them? At no other conference do all these themes come together in the particular combination that they do at AEA.

TABLE 2
Ten Questions about Evaluation Theory

<i>Question</i>	
1.	What are the four steps in the logic of evaluation?
2.	Are qualitative evaluations subject to validity criticisms?
3.	What difference does it make whether the program you are evaluating is new or has existed for many years?
4.	What difference does it make whether you are evaluating a large program, a local project within that program, or a small element within that project?
5.	How can you increase the chances that evaluation results will be used in the short-term to change the thing you are evaluating?
6.	What are the disadvantages of focusing on that kind of short-term instrumental use?
7.	What role does causal inference play in evaluation?
8.	Would your answer change if I asked what role causal inference played in making value judgements?
9.	When does a question have leverage?
10.	What is metaevaluation, and when should you do it?

Fifth, evaluation theory is “who we are” in the sense that it provides the face that evaluators present to the outside world. For example, think about how you would answer the question at work when your boss asks you, “Why do we need an evaluator? Can’t the auditors do the job? Can’t I just hire this economist who said she could do the evaluation using econometric models?” The answers that you construct to such questions define how you present yourself as an evaluator. I will bet that your answers will rely on those things that make evaluators unique—for example, our willingness and ability to attack value questions by studying merit and worth, our understanding about how to make results more useful than most other social scientists can do, or the strategies we have developed to help us choose which methods for knowledge construction to use depending on the needs of the evaluation client. All of these unique features of our field are enduring themes in evaluation theory.

Sixth, evaluation theory is “who we are” in the sense that it is the knowledge base that defines the profession. It is what *we* know that other professions *don’t* know. To help make this point, try to answer the ten questions that appear in Table 2. They are questions that I would *not* expect most other professionals to know because they are questions about ideas that are unique to evaluation. I do expect that evaluators should know them. Use the honor system and keep score for yourself.

Of course, you are not really going to be able to grade yourselves very well on this test because I do not have time to show you the answers—indeed, some of these questions do not have just one correct answer². But we can do several interesting qualitative analyses. First, some of you probably did not even recognize the terms in the questions. That is definitely a problem if you think you are an evaluator. The terms and concepts in these questions are part of the fundamental conceptual tools that make evaluation different from other professions. Second, many of you probably did recognize the terms. After all, if you hang around AEA long enough, you can probably talk the talk even if you cannot walk the walk. But if you were unable to answer most of them, then you probably do not know much about evaluation theory. And if you do not know much about evaluation theory, you are not an evaluator. You may be

a great methodologist, a wonderful philosopher, or a very effective program manager. But you are not an evaluator. To be an evaluator, you need to know that knowledge base that makes the field unique. That unique knowledge base is evaluation theory, reflected in the kinds of issues outlined in Table 2. That is why I said at the start of this article, evaluation theory literally is our field. It is who we are. If we do not know that knowledge base, then we are no different from the legions of others who also market themselves as evaluators today.

I realize that my contentions about the central role of evaluation theory in our field—and especially my presumptuousness in suggesting that those who cannot answer the questions in Table 2 are not evaluators—might be unduly provocative. I hope I can fall back on the norm that the President is allowed to be modestly provocative at least once in the Presidential Address. Of course, by the standards of some past AEA Presidential Addresses, my provocations may be mild indeed. Nonetheless, to avoid undue provocation, let me add six caveats. The first caveat is that the ability to recognize and discuss evaluation theory is different from agreeing with any particular evaluation theory. For example, reactions to Michael Scriven's plenary address at the 1997 AEA Annual Conference (see Scriven, this issue) ranged from agreement to sharp dissent. What would be problematic, however, would be to react by saying "I don't see what this has to do with evaluation practice". You may disagree with Scriven about his analysis of the nature and role of valuing in evaluation practice, but if you do not see that his claims are central to our understanding of the field, that is a problem.

The second caveat is that I do not want to make people feel unwelcome in the American Evaluation Association (AEA) if they do not know much about evaluation theory. After all, many people who are new to being an evaluator will not have had the time to learn much about the important theoretical issues in the field. Further, AEA attracts many people who are not—and may not want to be—evaluators, including program managers, policymakers, and researchers whose primary identity is with some other discipline such as psychology or education. For all these members, my hope is simply that the questions in Table 1 give them a better understanding of the central theoretical issues that make evaluators unique. Armed with that knowledge, they can decide for themselves how much of it they wish to learn. Presumably such learning is more important for those who aspire to a career as a professional evaluator.

The third caveat is that I am not claiming it will serve one's professional career better to be an evaluator than to be something else. From a financial point of view, for example, some of the most highly sought after consultants are statisticians who charge up to \$2500 per day; and industrial-organizational psychologists often charge quite a bit more. From an institutionalization point of view, professions like auditing have greater legal precedent supporting their activities than do evaluators, and this often translates into greater support for doing their job. From an access point of view, professions like economics often have a longer history of access and connections with the powerful governmental agencies that we frequently wish to influence than do evaluators. If you know evaluation theory, that may make you an evaluator, but it may not guarantee that anyone will hire you or listen to you.

The fourth caveat, closely related to the previous one, is that I have defined evaluators by what they know; but it is reasonable to define evaluators using the socioeconomic criterion that a person is an evaluator if that person is paid to act in a social role called evaluator, no matter what that person knows. After all, many people have occupied positions with evaluation job descriptions before there ever was an AEA, and before most evaluation theory was written. Who am I to tell them they are not evaluators, especially if I may even have followed their leadership in entering the field to begin with! These socioeconomic criteria are particularly powerful ones, and no approach to defining a field can be successful if it ignores them.

In fact, I would venture to guess that if it came down to a conflict between a theory-driven definition of evaluator and a socioeconomic definition, the latter would win every time. Indeed, the theories would probably even tend to follow the socioeconomics. Fortunately, however, this dichotomy is far more threatening in principle than in practice. After all, part of my point earlier in this article is that evaluation theory has *already* followed the socioeconomics of evaluation practice to a very large degree, so the discrepancy between what we do and what we say we do narrows annually. In fact, my goal is partly to help evaluators to see why narrowing that gap even further is crucial to the advancement of the profession.

Fifth, phrasing the ten questions in Table 2 as if they constituted a test that one passes or fails conveys an all-or-none quality to the decision about whether one is an evaluator, drawing a line in sand that some might argue is too confrontational. Might one better argue for degrees of "evaluatoriness"? Yes, but at the same time we must take seriously Scriven's central evaluative dictum that a judgment of value requires setting some standard of comparison that the evaluand must exceed to be judged good. Applying this dictum metaevaluatively to ourselves, we must confront the issue of how much of this knowledge we need to reach a minimally acceptable level of merit as evaluators. Surely one such minimum is that we need to know more about evaluation than other professions, which in turn argues for giving this test to them, too, to establish some nonevaluator norms. Whether we can take this one step further and justify a cutoff for passing and failing is beyond the scope of my talk today. But I do think our own evaluative logic requires some willingness to tackle this problem in a serious fashion.

Sixth, one can also imagine the possibility that the profession of evaluation might develop legitimate evaluation subspecialties, say, the experimental evaluator, the government evaluator, or the international evaluator. Each of these specialties probably would require more particular knowledge than is reflected in the questions in Table 2, so each would require questions and a test of their own. Yet the value of Table 2 would still remain even if such subspecialties developed, for those ten questions challenge us to ask whether evaluators have anything in common at all, whether evaluation subspecialties can even exist without an overarching evaluation parent specialty that justifies calling them *evaluation* subspecialties. It would be unfortunate for the field if we chose to develop and acknowledge the legitimacy of such subspecialties but neglected the higher order question of what holds these specialties together as a whole—especially if we did so because we could not figure out what we shared in common. I am optimistic about our readiness to know those communalities today, although I admit my optimism may be tempered by the fact that my own work has been aimed in substantial part at finding them (e.g., Shadish et al., 1991). Those communalities are knowledge of the sort reflected by the questions in Table 2, knowledge that cuts across all forms of evaluation, knowledge that is (or should be) an equal opportunity constraint on all of us³.

PROBLEMS WITH EVALUATION THEORY

I hope these six caveats will prevent readers from having negative reactions to the idea of a test for evaluators like that in Table 2 that are based on misunderstandings of its intent. There is, however, another objection that can be raised that is far more substantial, that evaluation theory is not sufficiently well-developed to warrant the central place that I have given it in our field. The point is interesting because it is just as important to understand the problems and limitations of evaluation theory as it is to understand its accomplishments, for those problems and limitations are also part of our identity, defining what we lack in our presentations to each

other and to the world. The first two of these problems concern important theoretical matters that no theorist has addressed particularly well:

- The general failure of most theorists to develop contingent theories of evaluation practice that tell evaluators how to make choices based on the contingencies of the situation.
- The general omission of a consensual vision of the place of evaluation in the world today.

The second two problems concern what we might call evaluation metatheory:

- The lack of a widely-accepted metatheoretical nomenclature that would help us to classify any given theory about evaluation, and to use that classification to understand what a particular theory does and does not claim to do.
- The neglect of a comparative theory of evaluation, one that uses the common meta-theoretical nomenclature to compare and contrast the relative strengths and weaknesses of individual theories.

Our failure to address them adequately is as much a part of who we are as any of our accomplishments might be—we are defined as much by what we do not do as by what we can. So let me describe these problems in somewhat more detail.

The Need for Contingency Theories of Evaluation Practice

We have previously pointed to the need for what I have called contingency theories of evaluation practice (Shadish et al., 1991). All approaches to evaluation involve tradeoffs among the many goals we try to maximize in evaluation (e.g., the goals of use, constructing valid knowledge, valuing, assisting in social change, etc.), so we need to have conceptual tools to help us understand those tradeoffs. By their very nature, those tradeoffs involve contingencies of the form “If A is the case, then B is an implication”. The contingencies in which we are interested in evaluation theory are those with implications that make a difference to evaluation practice. For example, a contingency discussed by Rossi and Freeman (1993) is the stage of program development—whether the program is new or long-established (see question 3 in Table 2 and in the Appendix). This is a contingency that makes a difference to evaluation practice because new programs frequently allow different kinds of questions to be asked and methods to be used than do well-established programs. One is more likely, for instance, to be able to find a suitable comparison group for use in an outcome evaluation of a new program than for one that is so well-established that it has full coverage. Other examples of such contingencies include the concept of question leverage, the structural size of the thing being evaluated, and the employment setting of the evaluator (Shadish et al., 1991).

The value of such contingencies is that they help us to evaluate and choose among competing forms of theoretical and practical advice for how to do evaluation. Nearly always, that competing advice stems from different, often implicit, presumptions about a contingency that can or should drive practice. For example, it helps to understand Scriven’s advice about doing evaluation to see how heavily his early writings depended on examples from product evaluation. Products are almost always a good deal structurally smaller than most social programs, so they can be changed more frequently and cheaply than can larger entities like programs

(see question 4 in Table 2 and in the Appendix); and products have a built-in dissemination system through the marketplace economy that makes issues of evaluation utilization less important to address explicitly than is the case with the use of information about social programs. If you are evaluating something small like a product, your options and the constraints surrounding those options are quite different than if you are evaluating something quite large like a national social program. The contingency involved in this analysis, that is, the structural size of the thing being evaluated, helps show when and where Scriven's analysis succeeds better or worse, and where other analyses might succeed better for the case of social programs. In this way, contingency theories show the potential unity underlying the appearance of great diversity.

These contingencies are important for another reason, as well. In many respects, an understanding of the contingencies of practice is a central feature of the development of a profession. Dreyfus and Dreyfus (1986), for example, noted that the distinction between a novice in a profession versus someone who has become a competent professional is that the novice often knows the requisite skills, facts, competencies, and rules, but the competent professional also understands that those skills and rules are embedded in complex, real world situations, and they are able to sort out the more important from the less important features of situations in making decisions about when, where, and why to apply those skills. The contingencies of evaluation practice are those features of the evaluation practice situation that are important in making decisions about what kind of evaluation to do. Without having a clear sense of those contingencies, evaluators are less capable professionals.

Finally, contingencies are the one conceptual tool we have in the field that give unity to the otherwise apparently disparate recommendations that most evaluators receive from evaluation theorists. Today, for example, one can find evaluation theorists that tell you to do experiments, to empower stakeholders, to make use the primary responsibility of the evaluator, or to make the judging of program merit or worth most important (to name just a small bit of advice one can find from reading any evaluation journal). It was this sort of diversity that led Glass and Ellett (1980) to say words that seem almost as true today: "Evaluation—more than any science—is what people say it is; and people currently are saying it is many different things" (p. 211). This diversity challenges us to make sense of the whole, to give unity to an evaluation theory that reflects so many different issues. Contingency devices are one of the main ways to achieve that unity because they tell us when and why each of the different pieces of advice might be preferred in evaluation practice.

The Need for a Vision of the Role of Evaluation in the World

If what I said in the previous section on contingencies is at least partly true, then it leads to a problem—it leaves evaluators without a simple, concise vision of what the profession of evaluation is all about. The profession would be so much easier to explain to the uninitiated, to market to potential funders, and to legitimize institutionally, if we could articulate such a vision that each evaluator could give as a response to inquiries from the outside. Granted, some evaluators have articulated such visions, but these visions are too scattered and too widely ignored for the good of the field, leaving evaluation without a central driving theme to present to the world. Scriven's vision (that evaluation is about valuing) is probably the nearest we have to a consensus about the matter, in no small part because nearly all evaluation theorists give at least lip service to the notion that evaluation is about merit and worth. Unfortunately, Scriven (e.g., this issue) has embedded his vision in a larger system of evaluation that

includes advice that many evaluators are reluctant to follow (e.g., evaluators usually should not make recommendations, they should often ignore program goals and not talk to managers, they should generally not take exploration of the reasons why something is good to be part of their job, etc.), so most evaluators throw out the baby with the bathwater by failing to implement the core of Scriven's advice about the investigation of valuing (e.g., question 1 of the Appendix). This is particularly unfortunate because Scriven's points about how one can move from data to value statements are applicable to all evaluations, because they would not take major changes in most evaluations to implement, and because they would serve to give the field somewhat more of a consensual vision.

However, Scriven's vision of evaluation as the science of valuing will fail to capture a substantial and important part of what has come to define the practice of evaluation—its focus on influencing the thing being evaluated for the better. In program evaluation, by far the dominant kind of evaluation among members of AEA, this translates into some recognition of evaluator interests in social policy, and in providing information that can in some way be connected to improving the capacity of programs to respond to the social needs that they might influence. In this sense, Scriven's vision is less compelling than, say, Weisner's (1997) vision of evaluation as an arbitrator between governments and marketplaces. However, it is very difficult to identify a concise statement of this sort of more general vision that most evaluators would agree on, that would capture the reality of the field, and perhaps most important, that would capture the imaginations of most evaluators such that they would make it a consensual part of their identity.

What I am looking for in this search for a vision does exist in other fields. In community psychology, for example, there is a consensual agreement among community psychologists about some of the central features that define the field—e.g., a commitment to prevention rather than cure, and a preference for treating communities rather than individuals. Similarly, the profession of auditing has a fairly clear understanding of its central features, such as the crucial importance of the independence of the auditor, or the focus of the auditor on compliance with standards (Chelimsky & Shadish, 1997). In both cases, the vision does not have to preclude diversity in how the vision is achieved; neither community psychologists nor auditors are uniform in their approach to their tasks. But their respective visions do allow them to present some version of a unified face to those outside the profession. Evaluators have yet to achieve this much.

The Need for a Common Metatheoretical Nomenclature

Another central problem in theoretical discourse is our lack of a common metatheoretical language and framework within which we can categorize and debate diverse positions. To appeal to an analogy, psychotherapists distinguish among various approaches to doing therapy such as analytic, dynamic, humanistic, systemic, behavioral, and eclectic therapies; and when a new therapy appears, they can classify it and quickly understand some of its likely strengths and weaknesses. We have too little of that linguistic capacity when it comes to understanding the place of new evaluation theories. We do, of course, share some language in common. For example, when most evaluators hear the word *use* or its cognates (e.g., utilization), they associate that word with the same general ideas and issues about how the process and results of evaluation can be used for various purposes. Continuing this example, most of us now recognize the distinction between instrumental use and conceptual use, and between short-term use and long-term use. These distinctions have

proven to be critical to understanding both theory and practice in evaluation. For instance, those of us who have been around AEA for the last decade remember the friendly disputes between Patton and Weiss about the use of evaluations back in the late 1980s (Patton, 1988; Weiss, 1988a, 1988b). Although the arguments were initially phrased as if use is a single thing, so that the debate could be won by one side or another, the crucial issues proved to hinge on the fact that Patton and Weiss were both right but were talking about two different variations on the general theme of use—Patton being more concerned with short-term instrumental use, and Weiss (at least at the time) more concerned with long-term conceptual use.

For present purposes, however, the point is even simpler. We must have the language to conceptualize such differences before we can debate and resolve them. What is needed, then, is a more general and complete set of terms that we can use to have such debates. For instance, the category system that is represented in Figure 1 reflects the terminology advanced in Shadish, Cook and Leviton (1991)—that all debates in the theory of evaluation practice can be divided into issues of how we value, how we construct knowledge, how evaluations are used, how evaluands function and change, or how practice is best done under the practical constraints we face. Within each of those five general categories, we proposed additional terminology. Regarding valuing, for example, we distinguished between descriptive, prescriptive, and meta-theories of valuing. Regarding use, we distinguished between instrumental, conceptual, and persuasive uses, and between long-term versus short-term uses. Regarding social programs, we distinguished between incremental versus radical change, between changing programs, projects, or elements, and between changing new versus existing interventions. Regarding knowledge construction, we distinguished among different levels of certainty that might be desired, and among different kinds of knowledge (e.g., descriptive, causal, explanatory). Regarding evaluation practice, we distinguished the kinds of questions asked (e.g., about needs, implementation, clients, effects, impacts, costs), about the things done to facilitate use, about whose questions to ask, about the role of the evaluator, and about what methods to use. This metatheoretical nomenclature is not intended to dictate which option is best (e.g., whether a descriptive or a prescriptive approach to valuing is best), even though we may also have taken a position on that issue after outlining the basic categories. Rather, the purpose of all these distinctions in terminology is to allow us to categorize—categorize general approaches to evaluation, any particular evaluation, or particular tactics recommended for implementing evaluation. That is important to the field because things that are categorized as similar are likely to share a host of important characteristics that go beyond those pertaining to category membership. For example, a focus on short-term instrumental use will tend to produce results pertaining to incremental rather than radical change. Chelimsky (this issue) thoughtfully challenges this latter claim, but I think close examination will show that her examples are the exception rather than the rule. Of course examining those exceptions is theoretically crucial, for I can imagine few more important tasks than to understand the circumstances under which we can get important social change quickly by virtue of what evaluators do.

In any case, it is not for me to say whether our particular metatheoretical nomenclature will eventually prove to be the best one; but I do claim that some such framework is absolutely necessary to theoretical progress in evaluation. It is the map of our territory that provides the common referents we need to navigate effectively.

The Need for a Comparative Theory of Evaluation

Evaluators have proven to be quite proficient at proliferating theory. Like psychotherapists constantly inventing new forms of therapy, or at least new labels for old forms of therapy, evaluators excel at inventing new ideas, new approaches, and new labels. In many respects, this is a good thing. We clearly do not yet understand all of the possibilities for practice, and it would be foolhardy and self-defeating to place the field in a straightjacket of strictures about what can and cannot be discussed. Nonetheless, this proliferation has not been accompanied by enough effort to compare and contrast theories with each other, and to learn from our history of previous successes and failures. It is probably only a minor exaggeration to say that our idea of theoretical history frequently seems limited to misunderstandings of logical positivism and caricatures of "traditional quantitative paradigms". We can do better than that.

To consolidate our theoretical gains, we need a comparative theory of evaluation to help us identify real innovation when it happens. That comparative theory should focus primarily on understanding the similarities and differences between different theories of evaluation practice. It would be a good start if we asked that whenever someone presents an approach to evaluation, they should present both its strengths *and* its weaknesses relative to other ways of doing evaluation; and that they present the historically important precedents to the approach they are proposing, along with the important ways in which their approach differs and does not differ from those precedents. In his 1997 revision of *Utilization-Focused Evaluation*, Patton (1997b) made some thoughtful efforts to do some of this kind of analysis of the strengths and weaknesses of his approach. Of course, none of us are particularly good at criticizing ourselves, so there is a limit on the extent to which we are likely to discover all our own weaknesses. But right now most evaluation theorists are making only token efforts to do this kind of comparative work, or no effort at all.

Let me give an example. Fetterman has in the last five years made empowerment evaluation one of the most widely recognized phrases in evaluation practice (e.g., Fetterman, Kaffarian, & Wandersman, 1996). There is much that I admire about Fetterman's accomplishments in this regard, and I very much envy his ability to show the field the value of his approach. However, we also need to have the capacity in evaluation to situate empowerment evaluation in a larger context of similar and dissimilar theories⁴. Of course, this capacity depends crucially on having the common metatheoretical nomenclature that I discussed a few minutes ago, which is another reason why that common framework is so necessary. From the perspective of our framework in *Foundations for Program Evaluation*, for example, Fetterman's work is probably most similar to the previous work of Patton, Wholey, and Stake, and least similar to the work of Scriven, Campbell, and Weiss. This might lead us to explore whether empowerment evaluation offers anything more than what Wholey or Patton has already offered, and if so, what⁵? Moreover, the placement of empowerment evaluation in this comparative framework allows us to make predictions about its likely strengths and weaknesses. For example, we might decide that it would be somewhat more prone to sacrifice the highest standards of accuracy in favor of use and social change than would some other approaches to evaluation like, say, Campbell's. I emphasize that this is not intended to be a negative criticism because all evaluation approaches involve tradeoffs among the many goals that we try to maximize in evaluation. It is merely to assert that it is possible to make these comparative judgments about strengths and weaknesses, and that making those judgments is essential to consolidating our theoretical gains over the last 30 years. To use an analogy, there

is a fine line between evaluation as a United Nations and evaluation as a Tower of Babel. In the former case, we strive to come together in ways that help us understand all our diverse strengths and weaknesses, and how they can be brought together into a whole. In the latter, we go so much our separate ways that we lose the capacity to discuss our relative strengths and weaknesses. The comparative theory of evaluation will help us move toward the former future and avoid the latter.

Theory and Practice: Partners in Evaluation

I started this article by saying that, as much or more than any other feature of our field, evaluation theory is who we are. I want to end by acknowledging that there is one other crucial feature of our identity that I have not yet discussed much, and of course, that is evaluation practice. Any discussion of evaluation theory that does not recognize the crucial role of evaluation practice in defining our field is bound to fail. Figure 1 highlights the centrality of practice by placing it in the center. Evaluation is primarily a practice-driven field, and evaluation theory is substantially driven by both the needs of evaluation practitioners and by their inventiveness. We meant to highlight this symbiotic relationship in the theme chosen for the 1997 AEA annual conference: "Theory and Practice: Partners in Evaluation". In fact, when we were trying to find just the right phrasing for the theme, one colleague suggested "Theory and Practice: *Equal* Partners in Evaluation". I rejected that suggestion because I do not think the relationship is one of equality. Of the two, practice is the *more* equal partner. Without evaluation practice, there would be no evaluation theory.

But by the same token, evaluation practice without evaluation theory can never be a recognized profession. It will simply be too scattered, too ill-defined, and too vulnerable to poaching by the many other people who also claim that they can do evaluative work as well as we can. Without evaluation theory, evaluation practice is little more than a collection of methods and techniques without guiding principles for their application. This is exactly the image of our field that we have fought to defeat over the last 20 years.

Evaluation practice is not just applied social science methodology. Evaluation is a field that raises deeply interesting and challenging intellectual issues, a field that has developed a set of unique conceptualizations about how to deal with those issues. Those issues, and how we cope with them, are the stuff of evaluation theory, and they define *who we are* as a profession.

CONCLUSIONS

In summary, I have tried to offer a very simple message, that all evaluators should be interested in evaluation theory because it is who we are. It is what we talk about more than anything else, it is what 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 regard, it is in our own self-interest to be explicit about this message, and to make evaluation theory the very heart of our identity. Every profession needs a unique knowledge base. For us, evaluation theory is that knowledge base. We need to consolidate it, nurture it, learn it, and teach it. Our future as a profession depends on it.

Author's Note

This article is based on the author's Presidential Address to the 12th Annual Conference of the American Evaluation Association on November 7th, 1997, in San Diego, California.

NOTES

1. This figure was also the logo for the 1997 Annual Conference of the American Evaluation Association.
2. Subsequent to the Presidential Address in which these ten questions first appeared, numerous colleagues asked me to discuss the answers to the questions to facilitate both education and further debate. Consequently, I have added the material in Appendix A to this article for that purpose.
3. Given these observations, a few colleagues have asked about the relationship between the ten question in Table 2 and the kind of examination that would be used to certify evaluators. Although certification exams were not my concern, the potential connection is too obvious to overlook. To the extent that a certification exam for evaluators would reflect a professional knowledge base that was unique to evaluators, it would probably include these kinds of questions. But it would also include many other kinds of questions, especially methodological and ethical ones, that would help determine that the test taker also had sufficient practical knowledge to implement the more general strategies that tend to be the focus of evaluation theory.
4. Of course, Fetterman's work has been subject to critical scrutiny (e.g., Fetterman, 1997; Patton, 1997a; Scriven, 1997), and some of that scrutiny has been comparative in nature. But the focus in these criticisms has generally not been the development of a comprehensive comparative theory of evaluation, but rather has been to comment on that one approach in relative isolation.
5. I might add that I do think it offers something more, but perhaps less than would be thought by those without real familiarity with evaluation's theoretical history.

APPENDIX A

Discussion of Questions in Table 2

The questions in Table 2 were designed as a rhetorical device in my Presidential Address to illustrate the kinds of theoretical issues with which I believe aspiring evaluators should be conversant in order to claim that they know the knowledge base of their profession. Subsequently, however, numerous colleagues have asked me to discuss the answers to these questions for both pedagogical and theoretical reasons. In this appendix, I will provide the rudiments of such a discussion; more extensive discussion can be found in Shadish et al. (1991), in the works of the many theorists that book discusses, and in references I provide in answering the questions below. Of course, the reader is reminded that some of these questions do not have a single correct answer, and that space limitations prevent me from providing here the level of detailed discussion that each deserves and that we have provided elsewhere. The questions vary considerably in difficulty, and in how universally the issues involved would be recognized by most evaluators today. What follows, therefore, is more of an outline of the issues than a correct "answer"; and the expectation should be that evaluators should be able to identify and discuss these issues intelligently.

What are the Four Steps in the Logic of Evaluation?

Here I refer primarily to Scriven's various writings on the topic of the logical sequence of concepts that defines how we try to connect data to value judgments that the evaluand is good or bad, better or worse, passing or failing, or the like. As he outlined the four steps (Scriven, 1980), for example, they included (1) selecting criteria of merit, those things the evaluand must do to be judged good, (2) setting standards of performance on those criteria, comparative or absolute levels that must be exceeded to warrant the appellation "good", (3) gathering data pertaining to the evaluand's performance on the criteria relative to the standards, and (4) integrating the results into a final value judgment. Scriven (this issue) presents a slightly more elaborated version of these steps, but their essence is the same. To the extent that evaluation really is about determining value, some version of this logic ought to be universally applicable to the practice of evaluation. It is, perhaps, the single most underappreciated idea in evaluation theory and practice.

Are Qualitative Evaluations Subject to Validity Criticisms?

Having just reviewed the qualitative literature on this topic (to appear in a forthcoming book), my impression is that the consensus answer of qualitative theorists themselves is probably "yes". That is, more qualitative theorists than not seem to both use the word and endorse some version of its applicability—although some qualitative theorists who reject both the term and any cognates seem to garner attention disproportionate to their representation in their own field. From outside the qualitative camps, the answer also seems to be more uniformly "yes". However, the subtleties required for an intelligent discussion of this question are extensive, of which the following few will illustrate but not exhaust. Even those who reject "validity" will acknowledge they are concerned in their work to "go to considerable pains not to get it all wrong" (Wolcott, 1990, p. 127). Further, within and across those methods qualitative theorists often disagree among themselves, with Maxwell (1992), for example, having quite different views than, say, Lincoln (1990). In addition, qualitative methods often (but not always) aim to produce knowledge of a substantively different kind than other methods, so that particular validity criteria (e.g., the classic internal validity criterion pertaining to descriptive causal inference) may be less frequently pertinent to the interests of qualitative evaluations relative to the applicability of validity criteria for, say, the meaningfulness of observations. Indeed, it would be wrong to assume all qualitative methods are alike, so that different qualitative methods may have different aims that bring different validity criteria to bear. In the end, though, some version of validity as an effort to "go to considerable pains not to get it all wrong" (Wolcott, 1990, p. 127) probably underlies all methods used by all evaluators, quantitative and qualitative.

What Difference Does it Make Whether the Program You are Evaluating is New or Has Existed for Many Years?

Rossi and Freeman (e.g., 1993) long made this distinction central to their approach to evaluation because it has several implications for evaluation practice. For example, brand new programs have not yet had time to work out program conceptualization and implementation problems, so a focus on those kinds of questions is likely to be more useful and more acceptable to program staff than a focus on, say, outcome questions. In addition, less background

information and fewer past evaluations are likely to exist for new programs, so more work will have to be done "from scratch". Well-established programs, on the other hand, may be more ready for outcome evaluation, and they may have a greater wealth of information already available on them. However, long-established programs may also have reached so many of the potential participants that outcome evaluations might be thwarted by difficulty finding appropriate control group participants if a controlled design is used.

What Difference Does it Make Whether you are Evaluating a Large Program, a Local Project within that Program, or a Small Element Within that Project?

This distinction points to an interesting tradeoff between ease and frequency of short-term change on the one hand, and likely impact on the other (Cook, Leviton, & Shadish, 1985; Shadish et al., 1991). Small elements (e.g., an admissions procedure) have natural turnover rates that are much more frequent than for local projects (e.g., an inpatient ward in a hospital), which themselves turnover less often than large programs (e.g., the community mental health center program). Hence the opportunity to change each of them by replacement is more frequent for smaller than larger entities. However, smaller entities are usually likely to have a smaller impact on the overall set of problems to which the program, project, or elements are aimed. All this has implications for the kinds of questions worth asking depending on what kind of use and impact is desired. Of course, the exceptions are as interesting as the rule here, with those small interventions that have large impact being of particular interest—if we could predict them (see Chelimsky, this issue).

How Can You Increase the Chances That Evaluation Results Will Be Used in the Short-term to Change the Thing You are evaluating?

The literature here is extensive (e.g., Shulha & Cousins, 1997; Patton, 1997b; Weiss, this issue), and includes advice to locate a powerful user(s), identify questions of interest to the users, focus on things that the users have sufficient control over to change, discuss exactly what changes the users would make given different kinds of answers that might result from the evaluation, provide interim findings at points when they might be useful, consider reporting results in both traditional and nontraditional formats (e.g., written report and oral briefing), provide brief executive summaries of results, have continued personal contact after the evaluation ends, and lend support to subsequent efforts to foster use of evaluation results (e.g., by testifying about evaluation results in public settings). And more.

What are the Disadvantages of Focusing on that kind of Short-Term Instrumental Use?

There is a risk that the evaluation will focus on less important interventions or questions than might otherwise be the case, and lose the big picture or the long-term outlook about what is important. In part this reflects the tradeoffs discussed regarding the program-project-element distinction because instrumental use is more likely with smaller elements likely to have less impact. It also reflects the fact that the modern industrial societies where much evaluation takes place have often solved the easiest problems, so that those that remain are often difficult to do anything about in the short-term. Those things that can be addressed in the short-term are rarely likely to fall into the set of most difficult problems. Finally, it is rare to find a user who can control options that promise truly powerful or fundamental changes.

What Role Does Causal Inference Play in Evaluation?

The most obvious version of this question concerns the role of outcome evaluation. From an early dependency on outcome evaluation as paradigmatic for the field (e.g., Campbell, 1971), the field realized the value of asking a wide array of other questions depending on contingencies like those discussed previously regarding use (e.g., sometimes the user wants to know about participant needs), program size (e.g., there may be little point in asking an outcome question about a large program if short-term instrumental use of the results is desired), and stage of program development (e.g., new programs may need information about implementation). Thus causal inference of the traditional sort assumed a smaller role in evaluation than in early years. Another version of this question appeals to the distinction between descriptive causal inferences (e.g., does the treatment work) and causal mediation (e.g., why does it work); the latter has a special role in the generalization of treatment effects (Cook & Shadish, 1994), and has enjoyed some recent resurgence in some kinds of theory-driven evaluation.

Would Your Answer Change if I Asked What Role Causal Inference Played in Making Value Judgements?

Most readers probably assumed “evaluation” in the previous question to mean the wide range of activities that fall under the rubric of professional evaluation practice. This question plays on limiting the meaning of the term “evaluation” to the activity of making a value judgment. Some readers might not realize that, even in this limited context, causal inference still plays an important role. Referring back to the answer to the first question about the logic of evaluation, in most applications it is implicit that the thing being evaluated caused the observed performance on the criteria of merit (e.g., that the treatment met recipient needs). If that were not the case, it would be improper to attribute the merit or value to the evaluand; rather, it should be attributed to whatever else actually caused the improvement in the criteria of merit. Thus the central task of evaluation, attributing merit or worth, is frequently causal in substantial part.

When Does a Question Have Leverage?

Cronbach and his colleagues (Cronbach, Ambron, Dornbusch, Hess, Hornik, Phillips, Walker, & Weiner, 1980) used this term to describe questions they thought particularly worth asking because of their potential for high payoff. Such questions have little prior information available, they can be feasibly answered with the resources and in the time available, the answers will probably reduce uncertainty significantly, and the answers are of interest to the policyshaping community. (Actually, they used the term in a slightly more limited way than this, but the spirit of this answer is roughly consistent with their view).

What is Metaevaluation, and When Should you do it?

Metaevaluation is the evaluation of evaluation (Cook & Gruder, 1978; Scriven, 1969), and recommendations vary from doing it for every evaluation to doing in periodically. The general prescription is that metaevaluation can be done using the same general logic (and sometimes methods) for doing the primary evaluation. One might apply the logic of evalua-

tion from the first question, for example, asking what the evaluation would need to do well to be a good evaluation (e.g., would it be useful, true, important?), deciding how well would it do so (e.g., how useful? true by what standards?), measuring the performance of the evaluation in these regards, and then synthesizing results to reach a judgment about the merits of the evaluation. Metaevaluation can be applied at nearly any stage of an evaluation, from evaluating its planned questions and methods, to a mid-evaluation review, to evaluating the completed report by submitting it to independent consultants and critics.

REFERENCES

- Campbell, D. T. (1971). *Methods for the experimenting society*. Paper presented to the Eastern Psychological Association, New York City, and to the American Psychological Association, Washington, D.C.
- Chelimsky, E., & Shadish, W. R. (Eds.). (1997). *Evaluation for the 21st Century: A handbook*. Thousand Oaks, California: Sage Publications.
- Cook, T. D., & Gruder, C. L. (1978). Metaevaluative research. *Evaluation Quarterly*, 2, 5-51.
- Cook, T. D., Leviton, L., & Shadish, W. R. (1985). Program evaluation. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd Ed., pp. 699-777). New York: Random House.
- Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past 15 years. *Annual Review of Psychology*, 45, 545-580.
- Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass.
- Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. NY: The Free Press.
- Fetterman, D. M. (1997). Empowerment evaluation: A response to Patton and Scriven. *Evaluation Practice*, 18, 253-266.
- Fetterman, D. M., Kaftarian, S., & Wandersman, A. (Eds.). (1996). *Empowerment evaluation: Knowledge and tools for self-assessment and accountability*. Thousand Oaks, California: Sage Publications.
- Glass, G. V., & Ellett, F. S. (1980). Evaluation research. *Annual Review of Psychology*, 31, 211-228.
- Kuhn, T. S. (1971). The relations between history and history of science. *Daedalus*, 100, 271-304.
- Lincoln, Y. S. (1990). Campbell's retrospective and a constructivist's perspective. *Harvard Educational Review*, 60, 501-504.
- Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, 62, 279-300.
- Olesen, V. (1994). Feminisms and models of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 158-174). Thousand Oaks, California: Sage Publications.
- Patton, M. Q. (1988). The evaluator's responsibility for utilization. *Evaluation Practice*, 9, 5-24.
- Patton, M. Q. (1997a). Of vacuum cleaners and toolboxes: A response to Fetterman's response. *Evaluation Practice*, 18, 267-270.
- Patton, M. Q. (1997b). *Utilization-Focused Evaluation: The new century text* (3rd Ed.). Thousand Oaks, California: Sage Publications.
- Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th Ed.). Thousand Oaks, California: Sage Publications.
- Scriven, M. (1969). An introduction to meta-evaluation. *Educational Product Report*, 2, 36-38.
- Scriven, M. (1980). *The logic of evaluation*. Inverness, California: Edgepress.
- Scriven, M. S. (1991). *Evaluation thesaurus* (4th Ed.). Thousand Oaks, California: Sage Publications.

- Scriven, M. S. (1997). Comment's on Fetterman's response. *Evaluation Practice*, 18, 271–272.
- Shadish, W. R. (1994). Need-based evaluation theory: What do you need to know to do good evaluation? *Evaluation Practice*, 15, 347–358.
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, California: Sage Publications.
- Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice*, 18, 195–208.
- Weiss, C. H. (1988a). Evaluation for decisions: Is anybody there? Does anybody care? *Evaluation Practice*, 9, 5–20.
- Weiss, C. H. (1988b). If program decisions hinged only on information: A response to Patton. *Evaluation Practice*, 9, 15–28.
- Wiesner, E. (1997). Evaluation, markets, and institutions in the reform agenda of developing countries. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 189–200). Thousand Oaks, California: Sage Publications.
- Wolcott, H. F. (1990). On seeking—and rejecting—validity in qualitative research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 121–152). New York: Teachers College Press.