There appear to be an increasing number of calls for more and better research on evaluation. After noting the benefits of having an increased evidence base about evaluation, I argue that there is also some merit in developing a framework of different types of studies of evaluation. I then present a framework of alternative types of research and illustrate it with various examples. The framework offered here includes (1) a classification of different “subjects of study” in research on evaluation and (2) a set of four functional types of studies. Expanding on the framework, tentative suggestions are given about the expected benefits of engaging in different types of research on evaluation.

**RESEARCH ON EVALUATION: A FUNDAMENTAL ISSUE?**

Why include a chapter about research on evaluation in a book addressing fundamental issues in evaluation? In Chapter 1, Smith indicates that fundamental issues recur over time, are interpreted in light of the current context, and are a topic of discussion and debate. From this vantage point, consider how research on evaluation can be seen as constituting a fundamental issue.
Research on evaluation has been a topic of interest for some time. For example, in a volume published in 1975, Bernstein and Freeman assessed the quality of a set of 236 federally funded evaluations on a number of aspects of methodology. Major studies of evaluation use took place roughly 30 years ago (e.g., Patton et al., 1977; Weiss, 1977). Studies of evaluation have continued to appear since then (e.g., Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985; Torres, Preskill, & Piontek, 1997; Christie, 2003), albeit sporadically and with limited cumulativity. Calls for more research on evaluation have occurred both earlier (e.g., Bernstein & Freeman, 1975) and more recently in the field's history (e.g., Alkin, 2003). Interest in research on evaluation ebbs and flows, but it seems inevitably to resurface.

Turning to another characteristic of fundamental issues, views about research on evaluation can vary in light of the current context. Take one noteworthy example. Carol Weiss, through her important studies of evaluation influence (e.g., Weiss, 1977), is one of if not the most notable practitioner of research on evaluation. Nevertheless, she recently wrote:

I'll admit that it is fun to talk to each other about evaluation theory, and having more research about evaluation (Henry & Mark, 2003; King, 2003) can no doubt contribute to our knowledge and skills. But I think it is time we spent more time doing evaluation, teaching people how to do evaluation, advising on evaluations, reviewing and critiquing evaluations, meta-analyzing evaluations, and in general, advancing the practice of evaluation.” (Weiss, 2004, p. 166)

Weiss's recent expression of concern about research on evaluation appears to have been stimulated by recurring requests from graduate students for her to participate in dissertations using survey methods to study evaluation. But perhaps her diminished interest in research on evaluation was also influenced by more general concerns about possible limits of the evidence-based practice movement as practiced (cf. Weiss, Murphy-Graham, & Birkeland, 2005). Views about research on evaluation may well depend on the context within which the idea is considered.

Moreover, research on evaluation is subject to sharp debate among evaluators. Contrast Weiss's recent comment, above, with the view of Mary Alkin (e.g., 2003), who, among others, has been calling recently for more research on evaluation. Alkin says “We must make a start” (p. 88) toward an empirical theory of evaluation based on systematic research. Yet another position has been voiced recently by Donaldson and Lipsey (2006). They suggest that research on evaluation in principle would be useful but that there will never be enough research to be help-ful as a guide to evaluation practice. Again reflecting Smith's analysis of fundamental issues, evaluators debate the value of research on evaluation.

The idea of research on evaluation is not minor. Rather, the potential consequences are considerable. For example, early research on use led to the concept of conceptual use or enlightenment, which remains influential today (Weiss, 1977). Early research on use also led to the notion of the “personal factor,” which became the foundation of Patton's (1997) utilization-focused evaluation. For several other examples of the impact of research on evaluation, see Shadish, Cook, and Leviton (1991) and Smith (1993). Contemporary and future research on evaluation likewise should have important implications.

In short, research on evaluation has long been of interest. Research on evaluation, and discussion of it, recurs and resurfaces in different forms. Views seem to vary at different times and in different contexts. And research on evaluation has potentially important consequences, as illustrated historically by the impact of early research on use on evaluation theory and practice. In short, research on evaluation can be seen as a fundamental issue akin to the others addressed in this book. In another sense, research on evaluation can contribute to our understanding of several other fundamental issues that exist in evaluation (e.g., what are the consequences of alternative evaluation approaches?). Research on evaluation might even help to provide an answer to perhaps our most fundamental practice issue: How should one choose from among the multitude of options available for evaluation?

RESEARCH ON EVALUATION: PREVIOUS CALLS

A key premise of this chapter is that value exists in moving beyond general calls for more and better research on evaluation to a taxonomy of types of research on evaluation. Put differently, this chapter is an attempt to contribute to a conversation about (1) the different categories of research on evaluation that can be conducted and (2) the relative benefits of different kinds of research on evaluation.

The idea that it would be helpful to increase the evidence base about evaluation practices and their consequences is not new. Consider, for example, Shadish et al. (1991), who reviewed a select set of major evaluation theorists. They concluded that the best way to improve advice for evaluation practice would be by moving "Toward More Data-Oriented Evaluation Theory" (p. 477). Shadish et al. indicate that “one of the most important but least frequent ways that evaluation theory has improved over the years is by increasing its empirical content” (p. 478).
And they conclude that evaluation “need[s] empirical study to answer the unanswered questions” that remain (p. 480). Similarly, Smith (1993) noted that “Empirical knowledge about the practice of evaluation is essential for the development of relevant and useful evaluation theories” (p. 237). These and other advocates of research on evaluation suggest that such research can help answer a range of questions. For example, how do we know whether a particular approach to evaluation meets its promises? What evidence, if any, exists for assessing whether a specific evaluation left clients better off? Indeed, do we have good answers to simple descriptive questions, such as what evaluation practice looks like in various areas of application?

Despite the hopes of Shadish et al., Smith, and others, many past and current debates about evaluation appear not to be based on careful and shared empirical evidence. Instead, debates are usually based on untested beliefs, testimonials by advocates, and anecdotes (Smith, 1993). It is probably more accurate to say that today we have an expert-based or evaluation-model-advocate-based evaluation practice rather than an evidence-based evaluation practice.

Against this backdrop, it appears that the number of voices calling for systematic research on evaluation may be increasing in recent years. For instance, Marv Aikin has in several publications advanced the argument Smith (1993), Shadish et al. (1991), and others have made for a “descriptive theory” of evaluation. Roughly, descriptive theory refers to an empirically based assessment of what evaluation looks like, under different conditions, and what kinds of consequences result from various approaches to evaluation (e.g., Alkin, 2003; Smith, 1993). As another example, several contributors to a 2001 special issue of the *American Journal of Evaluation* on the past, present, and future of evaluation issued a call for increased research on evaluation (see Mark’s 2001 summary).

These and other recent calls for more research on evaluation (e.g., Henry & Mark, 2003a) have an important cousin. That is, several studies of various kinds of evaluation (e.g., Christie, 2003; Campbell & Mark, 2006; Agodini & Dynarski, 2004; Petrosino, 2003; Weiss et al., 2005) have appeared in recent years. It is not possible to say whether this represents either a long-term trend or a response to the calls for more research from Shadish et al., Smith, Alkin, and others. Still, it seems a favorable sign, with potentially important benefits.

Possible Benefits of More Research on Evaluation

Calls for more and better evidence to guide evaluation theory and practice appear to make great sense. Shouldn’t evaluation itself be open to systematic inquiry, just as the policies, programs, and practices that we evaluate are (Dahler-Larsen, 2006)? Viewed from this perspective, a growing evidence base about evaluation should help answer questions such as: Which approaches to evaluation, implemented how and under what conditions, actually lead to what sort of improvements? A growing evidence base can also answer simpler questions, such as whether particular evaluative methods are used more in one program or policy area than another (see, e.g., Petrosino, 2003, who assessed the relative frequency of randomized trials in six different areas of interventions targeted at children).

In addition to providing better advice for evaluation practitioners (Shadish et al., 1991), a larger evidence base about evaluation might have several other benefits (Smith, 1993). For instance, increasing the evidence base of evaluation might:

- Improve the terms of debate among evaluators by helping to substitute some degree of empirical evidence for rhetorical style.
- Allow us to document and understand evaluation’s current and past contributions.
- Facilitate appropriate claims about what evaluation can do, perhaps most often by moving evaluators in the direction of modesty.
- Stimulate efforts to improve evaluation practice, in part by identifying circumstances in which evaluation demonstrably fails to meet its promise.
- Increase a sense of professionalism among evaluators by making it clear that evaluation itself is worthy of systematic study.
- Help move the field past generic and relatively abstract standards and guiding principles to more empirically supported guidance about the relative benefits of different evaluation practices.

Possible Benefits of a Framework of Types of Research on Evaluation

Despite the arguments for a more evidence-based evaluation practice, a case can be made that a *global call* for more research on evaluation is not sufficient. This chapter expands on past calls for additional research on evaluation by describing a possible typology that can be used to classify different kinds of research on evaluation.

Why might such a framework be worth developing? A taxonomy or classification system of alternative kinds of research can be of benefit in several ways:
It can help us identify gaps in the developing evidence base.

- It can therefore help guide the design of additional research on evaluation. In other words, the classification system can help discussion of research on evaluation move beyond general discussion of the need for research on evaluation to more specific talk about the benefits of particular kinds of research on evaluation.
- The mere existence of a classification system might also help stimulate further research on evaluation.
- A taxonomy can guide subsequent efforts to synthesize research on evaluation.
- In the preceding ways, a taxonomic system of the kind provided here might be able to help us move toward better answers to the questions about evaluation practice noted by Shadish et al. (1991).
- Thus, a taxonomy may aid in moving debates about evaluation practice away from claims based on rhetoric and perceived expertise rather than sharable evidence.
- Down the road, a taxonomy can serve a knowledge management function that helps guide the use of research on evaluation by others, including prospective evaluation users.
- Efforts to develop and refine a taxonomy of research on evaluation may help us move forward to a better framework of talking about evaluation issues in general.

If we assume, however tentatively, that at least some of these potential benefits would result, then the task of developing a framework of different kinds of research on evaluation seems worthwhile.

FROM GENERAL CALLS TO A FRAMEWORK

In thinking about a taxonomy of research on evaluation, one need is to organize the focus of such research. That is, a taxonomy should help clarify what is studied. A second possible part of a taxonomy would address how the investigation is done, that is, which kind of methods a particular study of evaluation uses. Consider first the focus of inquiry.

What Is Studied

Table 6.1 offers one way of organizing the various foci that research on evaluation might have. The table suggests that there are (at least) four overarching broad categories into which the subjects of research on evaluation will fall: evaluation context, evaluation activities, evaluation consequences, and professional issues. As we shall see, these are not mutually exclusive categories. To the contrary, important studies of evaluation often will examine relationships across these categories.

The categories in Table 6.1 emerged inductively from an attempt to classify real and hypothetical cases of research on evaluation. Interestingly, the categories that emerged resemble, to some degree, other frameworks familiar to evaluators. One correspondence is to logic models, where the current categories of context, activities, and consequences mirror similar categories used in logic models, and with professional issues corresponding to part of the inputs of a logic model. The Table 6.1 categories also correspond in part to the CIPP model of Stufflebeam (2003), with CIPP’s context, input, process, and product corresponding, respectively, to Table 6.1’s evaluation context, professional issues, activi-

<table>
<thead>
<tr>
<th>Concept</th>
<th>Subcategories</th>
<th>Example</th>
<th>Example</th>
<th>Example</th>
<th>Table 6.1 Subjects of Inquiry in Research on Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>implications for evaluation</td>
<td>evaluation approaches as represented by</td>
<td>credentializing evaluators</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaluation activities</td>
<td></td>
<td>theorists</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evaluation consequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Professional issues</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ties, and consequences. These correspondences suggest that the current inductively developed categories reflect meaningful distinctions.

Another category could be added to Table 6.1, one capturing *domain-specific* research questions. Domain-specific questions arise within a specific area of evaluation practice. For instance, one might test, meta-analytically, how well various instruments used in evaluations of early childhood interventions relate to subsequent academic performance. Or a researcher might compare across evaluations to try to assess whether a particular health intervention works in both rural and urban settings. Domain-specific research on evaluation can be extremely early childhood interventions relate to subsequent academic perfor-

**Table 6.1.** Relative to domain-specific research, research on ties, and consequences. These correspondences suggest that the current inductively developed categories reflect meaningful distinctions.

Another category could be added to Table 6.1, one capturing *domain-specific* research questions. Domain-specific questions arise within a specific area of evaluation practice. For instance, one might test, meta-analytically, how well various instruments used in evaluations of early childhood interventions relate to subsequent academic performance. Or a researcher might compare across evaluations to try to assess whether a particular health intervention works in both rural and urban settings. Domain-specific research on evaluation can be extremely important. However, the focus here is on the more general categories in Table 6.1. Relative to domain-specific research, research on Table 6.1’s more general categories should better contribute to *general evaluation theory* and practice.

The first of the categories in Table 6.1 involves *evaluation context*. Several subcategories fall within, reflecting different levels of context (here, as elsewhere in this model, only a subset of the myriad possible subcategories are described). Segerholm (2003), in her analysis of how evaluation plays out within a specific national context, reminds us that evaluation context can be examined at a *societal* level. For example, cross-national comparisons could be made. Alternatively, many studies of context will emphasize an *organizational* level (within a given society). For example, one might study the suggestions offered by Preskill and Torres (e.g., 1998) about which kinds of organizations are most open to incorporating evaluation in their ongoing routines. Background characteristics of a specific evaluation also form part of the context. For example, is the evaluator internal or external, what is the history of evaluation in the local context, what kind of training and experience has the evaluator had, etc.?4

The second category in Table 6.1 refers to *evaluation activities*. Several alternatives exist in terms of the way evaluation activities can be examined within research on evaluation. For instance, one might study a *global evaluation approach*. In a sense, Christie (2003) did this when she used multidimensional scaling to map the relative location of a selected set of evaluation theorists. For example, Christie (2003) found that the global evaluation approaches represented by Bob Boruch and Huey Chen differed on a methodology dimension from approaches represented by Ernie House and Michael Patton (see Christie, 2003, Fig. 1.1, p. 17).

As suggested in Table 6.1, an alternative way to study evaluation activities is to take a *given model* of evaluation, break it down into its *component parts*, and study each part. This approach has been used frequently in studies of therapeutic models in psychology; for example, researchers might identify the distinctive parts of, say, cognitive-behavioral therapy for generalized anxiety disorder, and then conduct studies examining the effects of each part in isolation and in conjunction with each other component. The study of a particular model’s components will likely be infrequent as research on evaluation grows, but it is an alternative worth keeping in mind. Among its benefits is that the componential approach requires you to do the work of identifying what the distinct parts are of a given evaluation model.

A more likely kind of research focus on evaluation activities will address specific *evaluation practices*. The idea here is to examine some specific aspect (or aspects) of the way evaluation is done in detail. (Unlike the previous option, this does not entail taking a particular evaluation model and decomposing it into its component parts.) As an example, Campbell and Mark (2006) studied the effects of alternative ways of structuring stakeholder involvement. In an “analog experiment,” Campbell and Mark examined the consequences of having stakeholder dialogue take place either with each stakeholder feeling accountable only to others who shared his or her beliefs, or alternatively with the stakeholders feeling accountable to a group of people with diverse views on the issue being discussed. In addition, the participants in stakeholder dialogue either did or did not receive instructions drawn from the literature on effective negotiations.

As the Campbell and Mark example suggests, studies of evaluation practices can focus on a wide variety of aspects of the work evaluators do. For example, Torres et al. (1997) focused on how evaluators report on their work. Julian, Reischl, Carrick, and Katrenich (1997) examined the consequences of different levels of participation by stakeholders (though in United Way decision making, rather than evaluation). As another example, a growing number of papers report the consequences of creating a comparison group in one way or another (typically, at random versus through some matching procedure; e.g., Agodini & Dynarsky, 2004). As these examples suggest, an almost limitless number of specific topics could fall under the subcategory of practices.

Third, arguably the most important focus of research on evaluation involves *evaluation’s consequences*. This focus is well represented in classic (e.g., Weiss, 1977) and more contemporary (e.g., Simons, Kushner, Jones, & James, 2003) research on evaluation use. Research on evaluation use (or influence; Henry & Mark, 2003a; Kirkhart, 2000; Mark & Henry, 2004) has been quite important. Continued research is needed, in part because the world has changed since the seminal research on evaluation use, including new arrangements that can alter the likelihood of use (Weiss et al., 2005).

Despite the potential importance of additional research on evaluation use/influence, it is important to remember that evaluation conse-
quences can include outcomes other than classic forms of use. For instance, the consequences of evaluation may involve changes in those who actively participate in the evaluation. Indeed, some approaches to evaluation explicitly express as a key goal the changing of participants in a certain way. For instance, empowerment evaluation has as one of its goals strengthening participants’ sense of empowerment. Research on evaluation seems especially warranted for those evaluation models that advertise such effects.

Quite importantly (though to date, infrequently), research on evaluation can study the consequences of evaluation for actual client outcomes. Henry (2004) reported an attempt to do so in the context of a longitudinal evaluation of Georgia’s pre-K program. Generally speaking, Henry attempted to examine the effects of the evaluation and its use on children’s outcomes at a later point in time. More specifically, an early wave of evaluation had identified benefits when pre-K teachers were credentialed. Evaluation use then occurred, in that policies were changed to provide new incentives for pre-K centers to employ credentialed teachers. Henry subsequently attempted to assess whether this policy change itself resulted in improved student outcomes. If similar studies of evaluation’s consequences become more commonplace, synthesis of many such studies would go a long way toward answering questions about the influence and impact of evaluation itself, and also about the relative influence and impact of alternative approaches to evaluation.

Another subcategory under evaluation consequences in Table 6.1 also refers to the consequences of evaluation on the evaluation context, or on subsequent evaluation activities, or on professional issues. For instance, does an evaluation itself change some aspect of the evaluation context? This possibility is often implied, for example, by evaluators who suggest that their preferred approach to evaluation will in one way or another strengthen democratic processes (Greene, 2006). Do these approaches have such consequences in practice? As another example, today’s evaluation within an organization may lead to changes in the evaluation approach that is used in that organization tomorrow.

Again, the point is that evaluation consequences should not be narrowly construed in terms of traditional forms of use—and not necessarily even in terms of the more important consequences of client outcomes and other participant consequences. Instead, in a given research study, the focus could be on the consequences of evaluation for subsequent evaluation context, or evaluation activities, or views regarding professional issues.

As the preceding discussion suggests, many important studies of evaluation will cross the categories of Table 6.1. Indeed, a persuasive case can be made that the most important studies of evaluation explicitly examine the relationship across two or more categories. For example, Henry (2004) did not simply look alone at evaluation influence. He also focused on a global evaluation approach. That is, he examined the consequences of his pre-K evaluation as a global whole, without parsing out the contribution of various components or specific practices. More importantly, he examined the effect of his global evaluation approach on evaluation outcomes, particularly children’s outcomes. As another example, Petrosino (2003) examined the relationship between program area, as a specific aspect of evaluation context, and the frequency of use of a specific evaluation approach, the RCT. One can imagine many more such intersections across the columns of Table 6.1. For example, studies of the societal component of evaluation context could examine the relationship between types of societies and the evaluation approaches that predominate. One could study the consequences of training for the evaluation practices that evaluators adopt. Many more examples are possible. The key point is that the most important questions for researchers on evaluation may involve trying to assess the influence of one category from Table 6.1 on another.

Turning to the fourth and final column of Table 6.1, professional issues are another possible subject of inquiry. Professional issues that might be studied include evaluation training, evaluation standards, and the credentialization or certification of training programs or evaluators (see, e.g., the survey of evaluators’ opinions about certification by Jones & Worthen, 1999). Other professional issues include the nature and function of evaluation associations, networks of or modes of professional communication (such as listservs; Christie & Azzam, 2004), professional conferences, and the like.

A Case Example of Research on Evaluation

Table 6.1 summarizes the “what” of research on evaluation, in the form of several general options regarding what the focus of such research can be. Before proceeding to discuss the “how”—that is, the ways of doing research on evaluation—consider one case example in some detail—in part because it foreshadows the subsequent discussion.

In an ongoing project, I am attempting to examine the “evaluation portfolio” at a number of organizations. By evaluation portfolio, in general terms I mean the way that evaluation resources are allocated across different kinds of evaluation activities. One way I am looking at evaluation resource allocations is in terms of the “inquiry modes” that Mark, Henry, and Julnes (2000) suggest constitute method families, or clusters of methodological approaches used by evaluators. The four inquiry modes are displayed in the upper half of Table 6.2.
As indicated in that table, **classification** refers to the use of methods to identify categories, such as various types of program clients or different groupings of service delivery settings. **Causal analysis**, in the context of evaluation, refers to the use of methods designed to do such things as estimate the effects of a program and/or identify or test the mechanisms by which a program has its effects. **Description** refers to methods that measure such things as the characteristics of clients, the nature of service delivery, or the standing of clients on important outcome variables, but without identifying new categories or actually investigating causal relations. **Values inquiry** refers to evaluation methods used to probe the values embedded in and related to a program, such as by assessing which of the potential process and outcome variables stakeholders care most about.

Consider a single case from an ongoing research project. The case involves a private U.S. foundation with a regional focus. This foundation makes grants in the area of health and access to health care, distributing roughly $10-$12 million annually. As summarized in the bottom half of Table 6.2, the vast majority of the evaluations funded by this foundation, roughly 95%, are descriptive. Mostly these involve relatively simple performance measurement and account for perhaps 20% of the spending on evaluation. On the other hand, the relatively small pro-

portion of evaluations that involve causal analysis account for a relatively high proportion of spending on evaluation, perhaps 70% (note: data on total spending for evaluation are, to use a “technical” term, rather “iffy”).

Describing the evaluation portfolio of even this one foundation can raise several interesting questions, such as: Is the formulaic approach to most evaluation planning in this foundation appropriate? What are the opportunity costs of the many, relatively cheap, descriptive evaluations? Given the general absence of classification and values inquiry, are these the evaluation “best buys,” at least for this foundation? How good is the planning that leads to selecting the projects for high-cost causal analyses? As I hope these questions suggest, descriptive analysis of evaluation practice can raise important issues for further consideration.

In addition to demonstrating the potential of research on evaluation for learning and for generating further questions for practice, the portfolio analysis, including its use of the Mark et al. (2000) taxonomy of inquiry modes, is presented here for another reason: in the process of reviewing numerous real and hypothetical cases of research on evaluation, I found that this taxonomy of research methods appeared to capture useful distinctions among various studies of evaluation. Although the taxonomy was initially developed to describe different kinds of methods that can be used in *doing evaluation*, I suggest here that it can also profitably be used to describe different kinds of *research on evaluation*. Table 6.3 gives a brief overview.

<table>
<thead>
<tr>
<th>Description</th>
<th>Classification</th>
<th>Causal analysis</th>
<th>Values inquiry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measuring characteristics of clients, services, outcomes</td>
<td>Categorizing (clients, service settings, etc.)</td>
<td>Estimating effects; identifying mechanisms</td>
<td>Identifying value positions of stakeholders and public</td>
</tr>
<tr>
<td>Focus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical methods</td>
<td>Performance measures; observations</td>
<td>Cluster analysis; taxonomies</td>
<td>Experiments; quasi-experiments; case studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portfolio: % of evaluations</td>
<td>95.5%</td>
<td>—</td>
<td>3.5%</td>
</tr>
<tr>
<td>Approx. % of evaluation spending</td>
<td>20%</td>
<td>—</td>
<td>70%</td>
</tr>
</tbody>
</table>

As indicated in that table, **classification** refers to the use of methods to identify categories, such as various types of program clients or different groupings of service delivery settings. **Causal analysis**, in the context of evaluation, refers to the use of methods designed to do such things as estimate the effects of a program and/or identify or test the mechanisms by which a program has its effects. **Description** refers to methods that measure such things as the characteristics of clients, the nature of service delivery, or the standing of clients on important outcome variables, but without identifying new categories or actually investigating causal relations. **Values inquiry** refers to evaluation methods used to probe the values embedded in and related to a program, such as by assessing which of the potential process and outcome variables stakeholders care most about.

Consider a single case from an ongoing research project. The case involves a private U.S. foundation with a regional focus. This foundation makes grants in the area of health and access to health care, distributing roughly $10-$12 million annually. As summarized in the bottom half of Table 6.2, the vast majority of the evaluations funded by this foundation, roughly 95%, are descriptive. Mostly these involve relatively simple performance measurement and account for perhaps 20% of the spending on evaluation. On the other hand, the relatively small pro-

portion of evaluations that involve causal analysis account for a relatively high proportion of spending on evaluation, perhaps 70% (note: data on total spending for evaluation are, to use a “technical” term, rather “iffy”).

Describing the evaluation portfolio of even this one foundation can raise several interesting questions, such as: Is the formulaic approach to most evaluation planning in this foundation appropriate? What are the opportunity costs of the many, relatively cheap, descriptive evaluations? Given the general absence of classification and values inquiry, are these the evaluation “best buys,” at least for this foundation? How good is the planning that leads to selecting the projects for high-cost causal analyses? As I hope these questions suggest, descriptive analysis of evaluation practice can raise important issues for further consideration.

In addition to demonstrating the potential of research on evaluation for learning and for generating further questions for practice, the portfolio analysis, including its use of the Mark et al. (2000) taxonomy of inquiry modes, is presented here for another reason: in the process of reviewing numerous real and hypothetical cases of research on evaluation, I found that this taxonomy of research methods appeared to capture useful distinctions among various studies of evaluation. Although the taxonomy was initially developed to describe different kinds of methods that can be used in *doing evaluation*, I suggest here that it can also profitably be used to describe different kinds of *research on evaluation*. Table 6.3 gives a brief overview.

<table>
<thead>
<tr>
<th>Description</th>
<th>Classification</th>
<th>Causal analysis</th>
<th>Values inquiry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential exemplar: Describing what evaluation actually looks like in various settings</td>
<td>Identification of different types of evaluation practice, contexts</td>
<td>Study of the consequences of evaluation practice</td>
<td>Assessing what stakeholders value re: what evaluation might achieve</td>
</tr>
</tbody>
</table>
Inquiry Modes as Part of a Framework

A range of methods are possible for description in research on evaluation, including observational, questionnaire, and content analysis methods. For example, a variety of studies that survey evaluation practitioners or evaluation users typically involve description. To take but one example, Torres et al. (1997) surveyed a random sample of members of the American Evaluation Association. The survey asked several questions about the communication and reporting of evaluations. For instance, respondents indicate that the most common forms of communication are written final reports. Other surveys have provided descriptive information about such things as evaluators’ views about use (Preskill & Caracelli, 1997) and the characteristics of collaborative or participatory evaluation as practiced (Cousins, Donohue, & Bloom, 1996).

Of course, descriptive methods other than surveys are possible for studying evaluation. For example, Petrosino (2003) recently examined abstracts from reports of outcome evaluations drawn from a set of electronic bibliographic databases. Petrosino used the abstracts to estimate the relative frequency of randomized controlled trials (RCTs) in six different areas of interventions designed for children. He found, for example, that almost 70% of the outcome evaluations of childhood interventions in health care appeared to use an RCT, while only perhaps 15% in education did. The assessment of evaluation portfolios, illustrated in Table 6.2, is another example of description. In scanning the literature, it appears that descriptive studies may currently be the most common form of research on evaluation.

Turning to classification, the Christie example comes very close to being an example. Certainly, if Christie had gone beyond her mapping of evaluation theorists by reporting cluster analysis results that put the various theorists into groups, it would have been classification. Another example of classification, though for the most part conceptually and not empirically driven, is the work by Brad Cousins and his colleagues (e.g., Cousins & Whitmore, 1998) that attempts to define different types of participatory evaluation. A paper by Cousins et al. (1996) that reports a survey of evaluators’ views about their collaborative evaluation practices reminds us that survey data can be put to use in classificatory analyses.

More generally, classification studies of evaluation can seek to identify the different types of evaluation practice, the varying types of contexts in which evaluation takes place, and the categories of evaluation users, evaluators, uses, and so on. Classification studies of evaluation are useful in their own right, in helping us better understand the lay of the land in evaluation practice. Especially given the relative paucity of classification studies, an increase in their use seems desirable. Classification studies also have considerable potential when combined with other kinds of methods. For example, one could combine a classification study of different kinds of participatory evaluation with a causal analysis, leading to conclusions about the differential consequences of different forms of participatory evaluation.

Moving to causal analysis, both the Henry (2004) and the Campbell and Mark (2006) examples fall into this category. The two cases illustrate some (but certainly not all) of the wide variety of methods that can be used for causal analysis. Henry’s study used quasi-experimental methods and statistical controls, while Campbell and Mark used experimental methods (in an analog situation). Yet other methods can be employed, including case studies. An argument can be made that causal investigations are likely in general to be among the most important forms of research on evaluation. This is because, just as program outcomes are typically of keen interest in making bottom-line evaluative judgments, the consequences of evaluation (and of alternative forms of evaluation) are vital for assessing the merit and worth of evaluation and its alternative forms. Given the important benefits of causal analysis as applied to evaluation itself, I briefly return in the “Discussion” section to a few issues regarding this form of research on evaluation.

One of the values of using the Mark et al. (2000) taxonomy is that it reminds us that research on evaluation can include values inquiry. In general terms, this would typically involve systematic inquiry into such questions as: What is it that stakeholders value (positively or negatively) about evaluation? Do different stakeholder groups generally value different possible consequences of evaluation? For example, in recent years various members of the evaluation community have discussed such possible consequences of evaluation as empowerment, creation of a space for dialogue, and facilitation of some preferred model of democracy. Despite the importance of such debates within the evaluation community, by most standards these debates do not rise to the level of systematic inquiry or research. And systematic research on stakeholder values could be quite informative. For example, we could assess how strongly various stakeholders value empowerment, dialogue, and related consequences, relative to previously emphasized potential evaluation consequences such as getting a better answer to a question that might be relevant to decision making. Values inquiry applied to evaluation itself includes the (in some sense modest but in practice perhaps radical) idea that, rather than having evaluators arguing among ourselves about the various alternative goals of evaluation, we should also consider attempting, at least in the occasional research study, to assess what relevant stakeholders think about these alternatives. Of course, this effort could
involve comparing values across different stakeholder groups, different organizations, different circumstances, etc.

Although empirical examples of values inquiry in research on evaluation do not appear to be plentiful, there are some near examples. For instance, in a study of educational research in Australia (most of which was descriptive), one phase involved asking several stakeholder groups about what kind of value they sought in educational research (Holbrook, Bourke, Owen, McKenzie, & Ainley, 2000). Segerholm (2003) also offers suggestions about what values inquiry might look like as applied to evaluation itself. Expanding on Segerholm's analyses, one might undertake a comparative study of the values surrounding evaluation in different nations.

This section and the summary in Table 6.3 offer a few examples of descriptive, causal analysis, classification, and values inquiry in research on evaluation. Space precludes going beyond these few examples and systematically illustrating the intersection between inquiry modes and Table 6.1's subjects of inquiry. I invite readers to consider the combination in different nations.

If used well, a taxonomy of research on evaluation will not constrain the way we think, but will foster thoughtful consideration of the many research questions about evaluation that are worth studying.

DISCUSSION

Drawing on previous calls for research on evaluation, this chapter is in part premised on the idea that a more evidence-based practice of evaluation can help (1) identify and (2) over time strengthen the contribution that evaluation makes. The call for an evidence-based evaluation practice is not new. However, openness to and opportunities for research on evaluation may be growing. For instance, the broader evidence-based movement should be conducive to research on evaluation itself. As one example, the activities of the Campbell Collaboration, which supports meta-analyses of evaluations of social and educational programs and practices, appear to be stimulating interest in research on evaluation use and related topics (e.g., Davies, Nutley, & Walter, 2005). In addition, the seemingly growing interest in and commissioning of evaluation in many contexts may imply more resources for research on evaluation. Some sources of funding appear to be open to supporting research on evaluation directly, such as the National Science Foundation's program on Research and Evaluation on Education in Science and Engineering (though this program appears narrower than one of its predecessors, on


evaluative research and evaluation capacity). Funding for research on evaluation is reportedly growing in Canada (J. Love, personal communication, January, 2006). Further, it appears that the evaluation community itself is growing, including drawing in some people who by virtue of their backgrounds may be interested in conducting research on evaluation itself (e.g., Campbell, Christie, Cousins, Henry, Petrosino). Such circumstances suggest that conditions today may be congenial to at least some increase in research on evaluation.

This, I believe, is good. Nevertheless, I suggest it is also worthwhile at this point to move beyond general calls for more research on evaluation. Thus, the framework presented in this chapter is offered as a supplement to more general calls for research on evaluation. This typology could have several benefits, including helping to stimulate further contributions to an evidence base for evaluation practice and offering a way of classifying, comparing, and synthesizing the findings that result. Having a taxonomy should also facilitate more thoughtful discussions about the relative value of different types of research on evaluation.

In this regard, one should expect that, to the extent research on evaluation grows, it will do so largely as a consequence of the interest of individual investigators and the joint actions of the community of scholars. The direction of future research on evaluation, as in any area of research, may also be influenced by such factors as funding opportunities, the relative feasibility of doing research in different areas, and the receptivity of journals to different topics (on the issue of options for funding for research on evaluation, see Henry & Mark, 2003b). In other words, arguing for the value of a taxonomy of research on evaluation does not entail replacing or devaluing the ordinary processes that arise in any area of research.

Brief Comments on Causal Analyses in Research on Evaluation

It is beyond the scope of this chapter to have a detailed discussion of the opportunities and challenges in conducting causal analyses as part of research on evaluation. However, a few brief comments seem warranted. First, those interested in the empirical study of evaluation may need to be opportunistic and clever in order to strengthen causal inference. Second, when comparing the consequences of different approaches to evaluation, researchers need to be mindful of the possibility of unfair comparisons. For more general discussion of unfair comparisons (e.g., confounds between stakeholder information needs and the type of evaluator selected), see Cooper and Richardson (1986). And third, openness to a range of causal methods is to be encouraged, with multiple methods an aspiration.
Metaevaluation and the Taxonomy

A question that can arise in terms of the present taxonomy involves whether, and how, metaevaluation (the “evaluation of evaluation”) fits. My answer is not a simple one, because there is no single universally agreed-upon methodology for conducting meta-analyses (Cooksey & Caracelli, 2005). If a given metaevaluation consists simply of expert review of the judgments taken in an evaluation, I probably would not label it as research. On the other hand, a meta-analysis might carefully, and with conscious attention to alternative explanations, trace the apparent consequences of an evaluation; like Henry (2004), this would be an example of causal analysis of the consequences of the evaluation approach used in the evaluation. As another example, a comparative meta-analysis might compare across several evaluations and examine the relationship between some aspect of evaluation context and the kinds of evaluation activities that were carried out (akin to Pretrosino, 2003). Again, this would fairly clearly be an example of research on evaluation. As these two examples show, metaevaluation does not intrinsically fit within one cell of the taxonomy, but can include different types of research on evaluation.

Conclusions and Caveats

The taxonomy described here, I believe, is at a different “level of analysis” than past discussions of research on evaluation. To date, discussion about whether and how to develop an evidence base for evaluation practice has either been quite abstract (e.g., general calls for more research on evaluation, or equally general skepticism) or quite specific (e.g., a solitary study of evaluation practice, or a critique of a single study). This chapter has offered a relatively mid-range analysis, with brief attention to comparative case analysis and with the development of a tentative typology of different kinds of research on evaluation. At some point in the evolution of a research-based evaluation practice, we probably need to move beyond abstract arguments for and against research on evaluation, and even to move beyond individual studies of evaluation and their critique. Taxonomies of research on evaluation are one way to move beyond the individual study and the abstract argument. A taxonomy can, for example, help us see where gaps exist in the literature, stimulate research, and aid in the synthesis and use of research on evaluation. A taxonomy can help achieve some of the cumulativity that is missing in evaluation (Mark, 2001, 2003).

Several caveats need to be added to this endeavor to move toward a taxonomy of types of research on evaluation:

• A single study (and, even more, a line of research) may partake of multiple kinds of research on evaluation. For example, Christie used simple descriptive methods to study whether her sample of evaluators reported subscribing to any evaluation theory, and she used (or at least approached) classification methods in her mapping of evaluation theorists’ responses.

• As with almost any classification system, it is likely that some cases will reside at the boundaries between categories and thus will be difficult to classify.

• At best, the current chapter is a start toward a good framework—and I hope it will not be the last word on the subject. At the least, various elaborations and modifications of the framework described here are possible. For example, if the body of research grows and if this taxonomy is used, revisions will be needed, such as finer distinctions.

• In addition, alternatives to the framework described here are possible. Rather than the list of four “inquiry modes,” you probably could use the chapter titles from your favorite research methods textbook. Rather than the “subject of inquiry” categories used here, several alternatives are possible. For instance, one might instead use Shadish et al.’s (1991) five components of evaluation theory: social programming (or, more generally, theory of the evaluand), knowledge, valuing, use, and evaluation practice. Perhaps there is even value in there being multiple complementary taxonomies of research on evaluation.

Despite all these (and, doubtless, other) caveats, I hope that we at least have begun a conversation about alternative forms of research on evaluation, that this conversation will help us move forward in terms of the broad project of creating a larger evidence base to guide evaluation practice (including stimulating thoughtful criticism), and that there also is some positive movement in terms of developing one or more frameworks for classifying the research that results.

One more general caveat: even those of us who advocate the development of more research on evaluation should also of course acknowledge the limits of such an endeavor. Research on evaluation will not be a magic bullet. It will not immediately transform the field. It will not replace all judgment about evaluation practice—but instead should aid such judgment. Research on evaluation, like evaluation itself, will raise questions about generalizability and applicability to specific situations. Research on evaluation will at least some of the time be ignored—even in cases where it could be useful. In short, the various problems and limits we all know about regarding evaluation (and research) itself and its use will also arise in the context of research on evaluation.
At the same time, I find it difficult to conceive of a truly compelling rationale for supporting the systematic study of policies, programs, and practices as part of our evaluation work while simultaneously arguing against the systematic study of evaluation itself. And the broad project of doing research on evaluation can move ahead and contribute even as we disagree. As we do with the very nature of evaluation itself, we may vigorously debate what research on evaluation should look like, what kinds of conclusions it should be directed toward, and so on. But, just as those debates about evaluation currently do not keep us from doing evaluation, I hope that any debates about research on evaluation do not keep us from doing research on evaluation.

At the very least, I am certainly persuaded that the broad project of developing an evidence base about evaluation itself is an endeavor worth engaging in, with some potential to strengthen the contribution that evaluation makes. The evidence base that results may even help move the discussion ahead on a variety of other fundamental issues in evaluation.

ACKNOWLEDGMENTS

This chapter is based on a paper presented at the 2004 annual meeting of the American Evaluation Association. Thanks go to Paul Brandon, Nick Smith, Tina Christie, and Gary Henry for helpful comments on a previous version. This chapter is supported in part by a grant from the National Science Foundation (No. REC-02311859).

NOTES

1. This admittedly contentious statement ignores, of course, that the experts and model advocates presumably base their preferences on some form of evidence from their practice experiences. However, in general such practice-based evidence (assuming it exists—as Ernie House (2003) points out, evaluation theorists sometimes get ahead of practice) is not readily shared, critiqued, detailed, and exposed to counterevidence in ways that we in general expect for research in general, or for evidence-based practice in particular—or for most professional evaluation, for that matter.

2. For example, the word transformative has been bandied about by many evaluators in recent years, and it would be good either to have evidence that evaluation is indeed transformative or instead to embrace the value of more modest consequences of evaluation.

3. Counterarguments can be made against the idea of a more evidence-based evaluation practice, but I do not see these as compelling. A brief listing and rebuttal of several counterarguments is available from the author.

4. Note that, in specifying these different levels of context, I am not suggesting that research on context needs to address all of these levels. As with other parts of the taxonomy, which category or subcategory is represented within a study will depend on the specific research question an investigator is motivated to examine.

5. One kind of descriptive study warrants special note. Many of the post hoc case studies that constitute a good part of the recent research literature on evaluation—in which, for example, the evaluator describes selected aspects of evaluation process and use—do not rise beyond the category of description. Many such case studies of evaluation appear to have been done retrospectively, after an evaluator notices some interesting aspect of the evaluation they are doing. It appears that post hoc case studies, because of their after-the-fact nature and attendant data limitations, often do not include the kind of features that would support reasonably strong conclusions about the evaluation's contributions.

6. This is not considered here as an example of classification, because the fourfold grouping of inquiry modes was not developed empirically during this investigation.

REFERENCES


Practice is the actual doing of evaluation. It involves the hard work of daily professional performance, often under difficult conditions and limited resources. Practice consists of implementing work plans, dealing with diverse political interests, responding to changing field conditions, and managing personnel and resources, all the while attending to shifting client and stakeholder needs and priorities. Practice entails somehow managing to simultaneously balance technical quality, client satisfaction, and available resources. If theory and method reflect the science of evaluation design, practice reflects the art of its performance. While theory and method provide the why and how of evaluation, it is practice that ensures that evaluations get done, with quality and in a manner that contributes to a better society.

Within the evaluation profession, issues of practice often are given less attention and are seen as less glamorous than issues of theory or