Using Christie’s research as an example, the authors describe a variety of forms that a more evidence-based approach to evaluation theory could take and offer some suggestions to help increase the amount and impact of evidence in evaluation theory.

Toward an Agenda for Research on Evaluation

Gary T. Henry, Melvin M. Mark

Evaluation seems to be almost everywhere these days. We read about the findings of large-scale program evaluations in the newspaper, we receive report cards on our neighborhood schools, we allow ourselves to be interviewed for evaluations of conferences we attend. Yet we know remarkably little about how evaluation is being practiced, why it is being practiced, by whom and where it is being practiced, and to what effect. As evaluators, we know a great deal about the evaluations that we conduct, and we may also know a fair amount about those evaluations that we come into contact with professionally, as citizens, or as participants. However, the big picture is less clear. Evaluators are unlikely to be able to say much descriptively about the size, scope, and shape of the evaluation enterprise—things like how many evaluations were completed last year, how much money was spent on evaluation, who did the evaluations, and how they were conducted. More troubling, if evaluators try to tell you anything deeper about the practice of evaluation, what they tell you almost certainly comes from their personal experience, interactions, reading, training, or intuition. The views you hear on the key issues in evaluation—such as what kind of evaluation to do in various circumstances, or when and why evaluation affects important actions—almost certainly are not based on rigorous, systematic evidence. Why not? Because there is a serious shortage of rigorous, systematic evidence that can guide evaluation or that evaluators can use for self-reflection or for improving their next evaluation.

This volume focuses on one new and noteworthy source of systematic evidence about evaluation: in Chapter One, Christina Christie presents her analysis of, first, the similarities or dissimilarities among the views of eight
prominent evaluation theorists, and second, the extent to which a set of practicing evaluators follow the guidance of those eight theorists. We believe that Christie’s research (Chapter One, this volume; Christie, 2001) may reflect, and even better, may help stimulate a revival of a tradition in the field of evaluation that has languished for too long: the collection of systematic evidence about evaluation itself. Starting in the mid-1970s, there was roughly a decade-long “golden age” of research on evaluation, with several scholars paying careful attention to gathering systematic evidence about evaluations. These included some who remain among the most respected figures in the field, such as Weiss, Patton, and Alkin.

In recent years, the tradition of gathering systematic data and explicitly testing one’s assumptions and working hypotheses about evaluation practice and its effects has withered. Prescriptive advice and admonitions about how to do evaluation have been plentiful, filling books, journals, conferences, e-mails, and conversations. But these are generally based on personal experience, observation, and the individual’s sometimes idiosyncratic beliefs and values—and not on carefully gathered evidence that can be described, shared, and critiqued. You can still find the occasional empirical study that provides comparative data about some aspect of evaluation (for example, Rog, 1985; Schula and Cousins, 1997; Preskill and Caracelli, 1997; Holbreck and others, 2000), but such work is rare and is hardly central to most discussion and debate about evaluation. We hope that Christie’s work signals a reawakening of the tradition of empirical, comparative study of evaluation. Toward that end, in this chapter, in addition to commenting on her study and some of its implications, we use Christie’s research as the starting point for a broader discussion of the systematic study of evaluation. Specifically, we focus on two questions: What should the agenda for research on evaluation include? And how can research on evaluation be promoted and realized?

**The Christie Study**

In recent decades, when scholars of evaluation have talked about knowledge accumulation and evaluation (for example, Lipsey, 2001), their focus has been on efforts to learn from evaluations through meta-analysis or research syntheses; little attention has been given to learning systematically about evaluation itself (Mark, 2001). We applaud efforts to achieve cumulative knowledge about social programs and their effects (knowledge from evaluation). But along with others (for example, Shadish, Cook, and Leviton, 1991), we also see the need for cumulative evidence about evaluation itself (knowledge about evaluation) to provide an empirical basis for improving practice and enhancing our understanding of the types of evaluation most likely to move us toward social betterment (Mark, Henry, and Julnes, 2000). Christie’s work is an important start in correcting the relative paucity of systematic research that allows cumulative knowledge about evaluation.
In particular, Christie's research offers a chance to peer into the practice of evaluators who were working under a mandate for local evaluations, with a prescribed framework for conducting those evaluations, and in a diverse array of local school districts that developed their own programs. Her data on evaluation practice are drawn from evaluators of California's Healthy Start program. Healthy Start grantees were required to produce annual evaluations and to allocate at least $10,000 per year to the effort.

Christie found that few of these evaluators appeared to follow closely the major tenets of the eight chosen theorists. Notably, she concluded that these practicing evaluators involved stakeholders less than the eight theorists would, on average. She also concluded that the practicing evaluators were more likely to be concerned with technical aspects of the methods they had chosen and to be more “method-bound” than the eight theorists reported being, on average.

Of course, it could be that the Healthy Start evaluators were following some theory other than those of the eight selected theorists. However, only 10 percent of them mentioned any theory that guided their evaluation, so we think it is improbable that there is such an existing theory to which they were adhering. At the same time, the Healthy Start evaluators turned out to be closer to some of the theorists than others, provoking some interesting speculations and commentary elsewhere in this volume. Still, the responses of the practicing evaluators were extremely varied, which is especially interesting because they all focused on the same program and were subject to the same evaluation guidelines.

Christie's research has some notable strengths. She developed the research with careful attention to a theoretical base, derived from the work of Alkin and House (1992). She engaged the eight theorists in the process of developing her measures, and she created an instrument that was apparently sensitive enough to capture differences in practices. She broke new and important ground by comparing the responses of eight theorists with those of the Healthy Start evaluators. Of course, every research project can be criticized, and Christie's work is not an exception to this rule. Questions could be raised about the specific theorists represented (or not represented), the adequacy of the items to capture all of the key issues on which the theorists may differ, the use of an exploratory technique (MDS) rather than a confirmatory one (confirmatory factor analysis), and the specific sample of practitioners. But we would not criticize the Wright Brothers because they did not have good in-flight food service. Similarly, we prefer not to criticize Christie's work but to congratulate her for getting it off the ground.

Christie suggests that evaluation practice is, to some extent, idiosyncratic. Different theorists and practitioners hold different beliefs, and presumably, engage in different kinds of practice—but on what basis? Where is the evidence that can help guide practicing evaluators or those sponsoring evaluations? As Christie notes, we lack evidence about evaluation that
can be used to guide evaluation. Simply put, we know little about what is being done in the name of evaluation and what works better under which circumstances for what purposes. Christie’s work opens a door for research on evaluation that goes beyond evaluation use. We would like to help push the door open wider. Accordingly, in the next section, we offer some ideas that might help stimulate and support more research on evaluation, expanding on the noteworthy work that Christie has done.

**An Agenda for Research on Evaluation**

In this section, we focus on research that can lead to accumulation of knowledge about evaluation and forgo discussing how to accumulate knowledge from evaluations (Mark, 2001). We offer a list of six types of research on evaluation, along with some examples of how these studies have been or could be conducted. One could argue about the six categories, and whether some should be combined and whether one is a subset of another. We have attempted to justify here our decisions to list the six categories separately. Suffice it to say that we think each is sufficiently important to encourage, and that each can contribute to understanding evaluation theory and practice. In the remainder of this section, we briefly describe these six types of research, which we believe should be included on the agenda of a more research-based field of evaluation.

**Research on Evaluation Outcomes.** Most of the more visible research in the “golden age” focused on the consequences of evaluation, under the rubric of studying evaluation use or utilization. In an important sense, evaluation can be seen as an intervention that may have effects on both the environment in which it is conducted and elsewhere. Research on evaluation outcomes and influence investigates these effects. We advocate new research on evaluation outcomes that examines the influence of evaluation not only on collective decisions such as policies and programs but also on individual-level cognitive processes and interpersonal behaviors (Henry and Mark, n.d.). Moreover, research on evaluation consequences should examine the connections between processes that may (or may not) lead to familiar kinds of use, such as instrumental use. Both evaluation processes and findings should be considered fair game for study. Intended effects (for example, program improvements) as well as unintended consequences (diffusion of the policy into other jurisdictions) should be examined.

A vibrant literature on the consequences of evaluation could be quite informative about the value of evaluation. However, the impetus for research on evaluation practice runs deeper. More evaluators now seem to acknowledge the potential contribution that evaluation can make to social betterment (Lipsey, 2001; Mark, Henry, and Julnes, 2000; Smith, 2001). But we do not have a credible research base to convince others about the extent of evaluation’s contribution to policy and program improvement. Nor, we believe, is there yet a research base adequate to guide evaluators’ efforts to
increase evaluation’s influence—for example, by providing better understanding of the various other factors that contribute to or detract from the use of evaluation findings. Put simply, we need to know much more about which types of evaluation-influenced changes have led to improved social outcomes, how the changes occurred, and under what circumstances similar results can be expected.

**Comparative Research on Evaluation Practice.** In a break from the tradition of research on evaluation use, Christie’s study does not focus on the outcomes of the evaluations, on what they were designed to influence, and on whether or not they were utilized. In this way she reminds us that valuable research on evaluation practice can be carried out that does not emphasize consequences or use. Christie’s research questions center on the extent to which theory guides evaluation practice. She examines this by comparing theorists’ self-reported predilections with the responses of practicing evaluators commenting about a specific evaluation. She has started down a very useful avenue on which we could develop a base of evidence that would allow evaluators to confirm or discard some of the many possible factors that influence how they conduct evaluations. Her research casts doubt on one of the possible influences on evaluation practice: evaluation theory. The Healthy Start evaluators did not seem to rely on any one of the eight “ideal-type” evaluations as represented by the responses of the eight theorists. But to be fair to these theorists and the potential influence of theory, it would have been interesting to have them respond based on the type of evaluation they would have conducted for a local Healthy Start program.

A key question that Christie’s research raises is this: What did influence the choices made by the evaluators? Christie considered the organizational position of the evaluator as one possible explanatory factor. She found differences between the two groups that provoke us to think about these factors more deeply. For example, internal evaluators report relying on stakeholders less than external evaluators do, but more of the internal evaluators were close to Brad Cousins, who is identified with participatory evaluation, and therefore, with high levels of stakeholder participation. Interestingly, Cousins’ responses indicated a more strategic use of stakeholders than some other evaluation theorists known for their advocacy of stakeholder participation, such as Fetterman and House. These results lead us to consider the importance of a contingent theory of evaluation (Mark, 2002; Mark, Henry, and Julnes, 2000). Local circumstances are likely to have influenced the extent to which stakeholders participated in Healthy Start evaluations and the extent to which various types of stakeholders—say, sponsors and program staff—influenced the evaluation (Mohan, Bernstein, and Whitsett, 2002). An enormous difficulty in prescribing an evaluation approach is that it must be flexible and responsive to the program, the sponsor, and other environmental aspects. But what are those contingencies? To what factors are evaluators responsive? Evaluators have lots of personal experiences and hunches but little systematic evidence
about which specific contingencies actually influence their practice substantially.

Christie’s study is but one example of the wide array of possible approaches to research evaluation practice. In general, such research would gather data on actual evaluations and how they are conducted, but could address a range of methods, approaches, and specific questions. Perhaps the most obvious examples include survey studies of evaluators, examining, for instance, the contingencies present in their evaluations and the impact of these on evaluation choices; comparative case studies examining several evaluations as they unfold to learn about decisions made and why; and quantitative analyses based on coding the attributes of evaluations. Some forms of research on evaluation may not be very obvious, however, and so we highlight them in the remainder of our list.

**Metaevaluation.** Metaevaluation is the evaluation of evaluations. It generally focuses on the analysis of an individual evaluation using a set of prescribed criteria and standards, although some examples exist of the comparative metaevaluation of several evaluations (for example, Scott-Little, 2002). Although metaevaluation could be included as a subset of research on evaluation practice, it has such a prominent standing in the field that it deserves special attention. But for all the attention, interest, and advocacy, actual examples of metaevaluation are sparse (for example, Datta, 1999; Grasso, 1999). Stufflebeam (2001) provides a helpful step-by-step description of metaevaluation, illustrated with some of his own metaevaluative work.

Metaevaluations can be informative about the nature of evaluation practice in specific cases. By comparing practice with standards, and sometimes by probing the consequences of evaluation in a specific case, such evaluations can also help evaluators reflect thoughtfully on their work. It may be, however, that the greatest benefit of metaevaluations to the field will come from syntheses across multiple metaevaluations (Gilliam and Zigler, 2000) when there are sufficient numbers of them to allow such comparative analysis.

**Analog Studies.** Analog studies are controlled studies, generally experiments, designed to reflect real-life practice settings while allowing for experimental control in testing some hypothesis about a potential influence on evaluation practice or outcomes. For example, Campbell and Mark (2002) developed an experiment to test the benefits of alternative ways of framing stakeholder dialogue. They developed different ways of engaging people in dialogue, not only from descriptions by evaluators using dialogic methods but also from the literatures on complex reasoning and negotiation. They chose a topic of considerable interest to their student participants (the role of the university in addressing student drinking), found students with diverse views on the topic to engage in dialogue, and in other ways tried to make the experimental setting analogous to actual cases of evaluation practice. Although there are limits to analog
studies, there are also strengths, and we see analog studies as a useful complement to the other kinds of work on the broader agenda of research on evaluation.

**Practice Component Studies.** Practice component studies are studies of the outcomes of specific evaluation activities or procedures. These studies can be done with a single evaluation or across several evaluations. Two studies that fall into this category are Greene’s examination of the effect of stakeholder processes on stakeholder views of the technical adequacy of an evaluation’s methods (1990) and Henry and Dickey’s investigation of stakeholder attitudes about the stakeholder process and its influence on an evaluation (Henry, Dickey, and Areson, 1991).

Practice component studies can produce important information about how a particular process worked, as well as insights into improving or discarding certain procedures. In addition, they offer the opportunity for practicing evaluators to gather systematic evidence about their own practice and produce usable knowledge for the field from their work. This, we believe, is quite important—research on evaluation should not be seen as the exclusive purview of a select few mostly in academic settings. We hope that many evaluators from diverse practice settings will take responsibility for producing research on evaluation. As we discuss later, low-cost, add-on practice component studies could be suggested in the overall evaluation proposal. A case can be made that practice component studies would be valuable to agencies, especially agencies that fund many evaluations, which can use them to refine future requests for proposals.

**Evaluation of Technical Assistance and Training.** Little systematic attention has been given to technical assistance or training, including evaluating its effectiveness. Connor (2001) has produced a snapshot of evaluation training programs that contrasts education, interdisciplinary social science, and policy programs that provide degrees or concentrations in evaluation. Christie speculates that evaluator training (or lack of training) may have influenced the evaluation choices evident in her research. However, there is little systematic evidence on the effectiveness of various types of evaluation training, degree programs, or technical assistance. These efforts should go beyond measures of client satisfaction, though certainly satisfaction is fair game for study.

**Proposals for Stimulating Research on Evaluation**

Increasing the amount of research on evaluation is, like many things, easier said than done. Encouraged as we are by Christie’s research, we believe that little will happen to build momentum in this direction unless evaluators themselves become proactive in promoting such research. In this section we offer a handful of proposals, some quite feasible and some rather ambitious, that could help get us to a healthy, productive, ongoing body of research on evaluation.
Find Opportunities to Piggyback Research on Evaluation onto Actual Evaluations. Some actual evaluations could easily be the venue for valuable research on evaluation—with modest additional resources. We encourage evaluators to propose such add-ons, where reasonable, when they respond to requests for proposals or otherwise negotiate evaluation contracts. We also ask evaluators to encourage funders, especially funders of large-scale evaluations, to develop criteria for judging add-on research components—and even add points in the proposal evaluation process to proposals that include a research component. In some cases, evaluators could look for opportunities to cobble together funding sources to add a valuable study of practices or consequences to an evaluation. For example, the modest funding available from internal university sources might sometimes allow a university-based researcher to add a research component to an evaluation, perhaps one being carried out by someone else.

Work to Develop Grant Programs and Other Funding for Research on Evaluation. Federal and state agencies, as well as foundations, invest significant funds in evaluation with the expectation that the information gained will contribute to change and help them reach their social objectives more efficiently. These funders should be encouraged to allocate some money to research on evaluation, with the goal being to increase the ability of evaluation to contribute to social betterment. Some funding sources already exist. For example, the National Science Foundation’s program on evaluation capacity can support some kinds of research on evaluation. Traditional funding streams in the social and behavioral sciences might support certain kinds of research on evaluation (for example, on the outcomes of evaluation). Some foundations have grant programs that allow research on a profession or field that carries out the social missions they are pursuing. Evaluation can be seen as such a profession, one that will be able to contribute better to the foundations’ agendas if a better knowledge base about the profession is developed.

Create Evaluation Programs in More Universities. Most universities have evaluators on the faculty and in research centers, but relatively few have demonstrated the critical mass (or the will) needed to support and train graduate students in an evaluation program. Having more such programs would have several benefits, one being a likely increase in research on evaluation. Research on evaluation would fit with the reward patterns and interests of many members of evaluation faculty, and it could be a valuable adjunct to the training of graduate students.

Alternative models of evaluation training are possible, including stand-alone departments, programs in more general departments, and interdepartmental training programs. Evaluators who work in university settings could do much to start such programs. Evaluators can also lobby funding agencies to provide external funding in support of evaluation training. If the National Research Council could be convinced to rate evaluation programs in its occasional ratings of university departments, this might influence the
many universities that are interested in increasing their standing (Feller, 2002). Universities might also be influenced by the apparent market for trained evaluators. Whatever else you think about certification or credentialization (Altschuld, 1999), or about adding evaluation training as a criterion in the awarding of evaluation contracts, such steps would probably increase the number and size of evaluation programs—and indirectly increase research on evaluation.

**Increase Recognition of Research on Evaluation.** Collectively, the field could in several ways increase recognition of research on evaluation. Such work could be featured prominently in our journals and conferences. Editors, reviewers, and audiences could ask authors and presenters who advocate a position to highlight the supporting research (Mark, 2002). One of the merits of this issue of *New Directions in Evaluation* is that it highlights trail-blazing research on evaluation. Yet, other forms of recognition could be created. For example, the American Evaluation Association offers several awards, but none specifically focuses on research on evaluation. During the first years of an award’s existence, it might not be given out annually. However, over the long run, an award, along with other ways of highlighting research on evaluation, could help increase both the frequency and quality of such work.

**Adopt the Five Percent Solution: A Federal Accountability, Credibility, and Transparency Act.** Our final proposal is the most ambitious. We recommend that every publicly funded program that expresses goals for social betterment should, by legislative mandate, include 5 percent of the budgeted funds for evaluation. The golden age of evaluation research was built on a similar requirement in a few federal funding streams. Such a requirement would serve us well again in these days when “evidence base” and “science-based evidence” are once more in vogue. Although it could seem self-serving for evaluators to lobby for a mandate for evaluation, we believe a compelling case can be made based on society’s best interests. In a sense, the Government Performance and Results Act (GPRA) can be seen as an antecedent, with the proposed Federal Accountability, Credibility, and Transparency (FACT) act providing resources for the evaluation activities that can complement performance measurement systems and help alleviate their potential dysfunctions (Feller, 2002).

The proposed requirement would not only lead to far more evaluation-based evidence in service of social betterment but also would give evaluators and program administrators greater responsibility for ensuring that society is getting value for the expenditures on evaluation. By increasing the attention on evaluation, its quality and consequences would be more in the spotlight. Administrators and evaluators should answer questions about value, and the 5 percent solution would make this issue more salient. Research on evaluation is one way to help answer questions about the value of evaluation. And if a small part of the 5 percent set aside for evaluation can support research on evaluation, the endeavor could truly flourish.
Although we envision a FACT act that applies across the federal government, intermediate steps could also make a big difference. Evaluation practice, and indirectly, research on evaluation, could be substantially stimulated with initiatives that provide evaluation funding for select agencies. State-level laws would be beneficial and might lead to more widespread adaptation. And the 5 percent figure is of course subject to discussion and change—but we should try to avoid evaluators’ tendency to hoist themselves on their own petard by arguing about the right percentage of funding that should be devoted to their field, leaving the decision fully to others who, in the absence of firm advice, commit little or no funding at all.

Conclusion

Christie’s research on evaluations required by California’s Healthy Start initiative provided the spark for this chapter. Her research finds little evidence that “ideal types” from evaluation theory are carried out in these local evaluations. However, we should not close the book on the potential for evaluation prescriptions to influence practice. We believe that higher-order theories are required to make practice contingent on local circumstances, needs, and conditions. These higher-order theories lie in the middle ground between cookbooks and cafeterias. Cookbook recipes that lock us into certain decisions are unlikely to produce satisfactory evaluations in the complex real world of social programs, but cafeteria lists of options are not particularly helpful either when you need to know which one is best in a particular circumstance. We suspect that all the theorists included in Christie’s study—and most others who would accept such a label—have in mind and practice contingent theories. They may often have preferred modes of operating, but they deviate from them when the circumstances present a compelling need to do so. For example, David Fetterman (2000), widely known as a proponent of empowerment evaluation, has discussed an important evaluation where he chose not to use an empowerment approach. Such examples of contingent decision making by proponents of an evaluation approach are noteworthy, but we need to move beyond individual judgments and develop a better knowledge base that can inform judgments about what kinds of evaluation activities are most likely to have what consequences under which conditions.

Christie pursued the research she describes in this volume as part of her doctoral work and can now press further. We look forward to other reports based on that work and to her future contributions to the literature. In this chapter, we have suggested that the value of Christie’s work is not limited to her specific findings or to the debates and discussions it will foster about various evaluation theorists and their relationship to evaluation practice. Instead, we hope that her research may be a harbinger of a new golden age of research on evaluation. It is likely, however, that that will depend in no small part on the future actions not only of Christina Christie but also of many of the readers of this volume.
References


**GARY T. HENRY** teaches in the Andrew Young School of Policy Studies at Georgia State University, where he also serves as a professor in the Department of Political Science and the Department of Public Administration and Policy Studies.

**MELVIN M. MARK** is a professor of psychology and senior scientist at the Institute of Policy Research and Evaluation at Pennsylvania State University. He is editor of the *American Journal of Evaluation*. 