This section presents interviews with evaluators whose work can illustrate, in a specific evaluation study, the application of different models, theories, and principles described in evaluation literature. These interviews should also exemplify the choices that evaluators make in the course of planning and conducting and evaluation. Entries in this section will begin with a brief description of the evaluation to provide a foundation for the interview. The focus will be on the dialogue between the section editor and the evaluator to learn more about the choices made, the models followed, the constraints and opportunities that emerged as the study progressed, and the evaluator's assessment of the study itself. The section editor will provide commentary as needed to point out important attributes of the evaluation itself and to highlight connections between the evaluator's comments and relevant literature or issues in the field of evaluation.

Readers who wish to suggest an evaluator or evaluation for Exemplars may contact Christina Christie, section editor, at Tina.Christie@cgu.edu.

Exemplars as Case Studies: Reflections on the Links Between Theory, Practice, and Context

JODY L. FITZPATRICK

Available online 1 October 2004

ABSTRACT

Responding to the call for more research on the practice of evaluation and its link to theory, this article uses the interviews conducted with exemplary evaluators or exemplars in the *American Journal of Evaluation* as cases to describe and analyze the practice of exemplary evaluators. Dimensions of analysis included context, purpose, advance organizers, breadth and depth of stakeholder involvement, method proclivity, and use of results. Results showed that on these cases, evaluators differed on purpose, advance organizers, and choices on breadth or depth of stakeholder involvement. Determining merit and worth was a frequent and primary purpose for many. All used mixed methods, but varied in the confidence they placed in different methods. Instrumental use was identified in each evaluation. Some evaluators' practice was congruent with theories; others were more atheoretical. Differences were often nuanced, influenced by both context and evaluator proclivity.

Jody Fitzpatrick • Graduate School of Public Affairs, University of Colorado, Colorado Springs, CO 80907, USA; Tel. (1) 303-393-8567; Email: jfitzpat@carbon.cudenver.edu

American Journal of Evaluation, Vol. 25, No. 4, 2004, pp. 541–559. All rights of reproduction in any form reserved.
ISSN: 1098-2140 © 2004 Published by Elsevier Inc. on behalf of American Evaluation Association.

Starting in 1997, I conducted a series of interviews with prominent evaluators focusing on a single evaluation they had conducted. These columns, published in this journal, were intended to help readers “learn more about the choices an evaluator makes in the course of planning and conducting an evaluation” (Fitzpatrick, 1997, p. 239). The idea for the pieces emerged from my observation that, with some exceptions, articles in evaluation were either research articles focusing on results and methodology or essays on how evaluation should be practiced with brief examples. I wanted to learn more about how exemplary evaluators actually practiced their craft—the roles they played, the choices they made in framing the evaluation, the involvement of and negotiations with stakeholders, the methods and approaches they used, and the ways in which they reported results and attempted to facilitate use. The early interviews, in a section entitled *Exemplary Evaluations*, focused on evaluations that had won awards, such as the American Evaluation Association Award for an Outstanding Evaluation. Later interviews, with a name change to *Exemplars* at the suggestion of Mel Mark, then the new Editor of *AJE*, were conducted with prominent evaluators, often theorist-practitioners, and focused on an evaluation of their choice. (See Table 1 for a list of evaluators interviewed and a brief description of their evaluation.) Typically, these evaluators chose work that they felt reflected their theories and approach to practice (see Fetterman & Fitzpatrick, 2000 for an exception.) My reflections here on these columns are prompted by the transition to a new editor of *Exemplars*, Christina Christie.

As the interviews progressed, I began to notice several themes emerging. In many ways, these evaluators were more similar in their approach to evaluation than one might have supposed. Yet, they differed on subtle, but potentially important, factors. Further, evaluators whom others in the field might consider dramatically different were, in fact, quite similar on certain dimensions. Conversely, evaluators one might assume to be quite similar showed differences in practice and approach. Christina Christie’s research on the connection between theory and practice (Christie, 2003) prompted me to consider the themes revealed in the interviews further. Lois-ellin Datta (2003), in fact, observed the potential for triangulation between Christie’s quantitative work and my more qualitative case studies. Datta’s goal, and mine here, is to bring together different sources of information to “build up a useful empirical mosaic for further modeling of the paths between theory and practice” (Datta, 2003, p. 39). The purpose of this piece is to reflect on some of the relationships between theory, practice, and context that can be observed in the *Exemplars* case studies.

DIMENSIONS OF EVALUATION THEORY AND PRACTICE

Christie’s work began with the three dimensions of the taxonomy developed by Alkin and House (1992) to classify evaluation theories: methods, values, and use. She asked eight evaluation theorists, selected to represent a wide range of theories on the taxonomy’s dimensions, to develop statements illustrating their particular theory in terms of each of these three dimensions. With these statements, Christie developed a 38-item, Likert-scale instrument and surveyed the eight evaluation theorists and 138 evaluators conducting evaluations in California’s Healthy Start program. Theorists and practitioners were asked to respond to the items in terms of how they actually conduct evaluations. Her multidimensional scaling analyses of the practices reported by her eight evaluation theorists revealed two primary dimensions: scope of stakeholder involvement and method proclivity (Christie, 2003). Scope of stakeholder involvement was measured by nine items tapping stakeholders’ involvement in planning and conducting

TABLE 1.
Description of the Cases

<i>Evaluator</i>	<i>Description of Evaluation</i>
James Riccio	Evaluated the experimental California welfare reform program, GAIN, at the request of the California legislature. Studied variations in implementation and outcomes in six diverse counties.
Marsha Mueller	Evaluated the use and effect of Minnesota's Early Childhood Education (ECFE) program on low-income families in 14 school districts through training staff in evaluation and data collection methods. Studied changes in families' knowledge and behavior.
Debra Rog	Evaluated the Homeless Families Program designed to provide services-enriched housing and systems change in nine cities across the country. Collected information on characteristics and needs of homeless families, systems and services available, program delivery, and outcomes with families and systems.
David Fetterman	Evaluated the Stanford Teacher Education Program (STEP). Studied program activities, teaching quality, student reaction and practice in the field, and other teacher education programs to judge program quality.
Gary Henry	Developed a system for monitoring school performance in Georgia. Using existing data and research, surveys of citizens, and input from stakeholders, developed school "report cards" for monitoring performance in elementary, middle, and high schools.
Jennifer Greene	Evaluated the Natural Resources Leadership Program, a program designed to provide leaders who have different perspectives on environmental issues with new approaches to resolving environmental conflicts. Studied program activities, changes in participants, and use of the training in resolving conflicts.
Leonard Bickman	Evaluated continuum of care mental health systems at Ft. Bragg and Stark County. Studied program implementation and quality as well as changes in children's mental functioning and families.
Stewart Donaldson	Evaluated four programs implemented in 40 organizations. Focus was on the Winning New Jobs evaluation designed to help laid-off workers regain employment. Studied program theory and actions and changes in participants.

the evaluation. Method proclivity was assessed using six items on a variety of methodological issues, but with a focus on methodological flexibility and preferences for qualitative and quantitative methods. While Christie found that her sample of practicing evaluators from California's Healthy Start program differed rather widely on these two dimensions, all eight theorists, who were chosen to reflect a wide range of evaluation theories, reported more similar practices. In fact, all eight were on the positive side of stakeholder involvement and five of the eight were quite similar in method proclivity, being less inclined to favor a particular methodology that included predetermined research steps. In other words, the theorists she studied reported rather similar practices, compared to practicing evaluators, in involving stakeholders and most demonstrated a flexible, responsive approach to methodology.

This review of the *Exemplars* columns was influenced by Christie's work, [Stufflebeam's \(2001\)](#) review of evaluation models, and commentaries on Christie's research ([Datta, 2003](#); [House, 2003](#); [King, 2003](#)). Stufflebeam's identification and analysis of 22 evaluation models,

or approaches, does not link these models with practice, but is useful in reminding us that evaluation theories differ in the purpose(s) they advocate for evaluation. Alkin (2003), too, has noted the centrality of purpose in contrasting evaluation theories. In his graduate course with evaluation students, each is assigned to study a particular theorist. A central element of their assignment is to describe what a theorist views as the ultimate and the subsidiary purposes of evaluation. King notes that "program evaluation means different things to different people" (2003, p. 58) and emphasizes how diverse those meanings can be in a transdiscipline, like evaluation, where practitioners come from many different fields of study with different research traditions and approaches.

Many in evaluation (Fitzpatrick, Sanders, & Worthen, 2003; Stufflebeam, 2003) emphasize its traditional purpose: determining the merit or worth of a program or policy. However, Scriven (1996) and Henry (2000) have observed that, in practice, evaluations are more frequently designed for program improvement and often do not serve the ultimate purpose of judging merit or worth. Mark, Henry, and Julnes (2000) have identified three additional purposes that may more accurately describe evaluation in practice and will be used in the analysis of the *Exemplars* cases. Along with the purpose of determining merit and worth, they include program or organizational improvement (formative evaluation), oversight and compliance (an important aspect of accountability), and knowledge development (research). These purposes will be used as one factor to describe the connection between theory and practice in these case studies.

Stufflebeam (2001) also identifies advance organizers as an important criterion for differentiating evaluation theories. Although he does not define advance organizers in his monograph, they appear to entail the stimulus or heuristic devices used in different evaluation approaches to frame the evaluation. Many prescriptive evaluation writings implicitly or explicitly recommend advance organizers in suggesting appropriate methods to plan an evaluation. These can include identifying a key decision-maker or user and the decisions he or she will make; considering a variety of stakeholders' needs; determining program goals and objectives; specifying or developing program theories; reading scholarly research and evaluation literature on similar programs; obtaining the opinions of content experts; empowering or advocating for clients, staff, the public, or other less powerful groups; stimulating a democratic dialogue; or encouraging organizational development through learning. These and other methods serve as the foundation for determining the *focus* of the evaluation. For many evaluation theories, the advance organizers are the central concept to theory and, thus, are important to examine in the *Exemplar* cases to better understand the links between theory and practice.

House (2003) and King (2003) call our attention to the importance of context, as does Henry (2001) in his introduction to Stufflebeam's volume. Commenting on the similarities in theorists' practice seen in Christie's data, House writes, "Practices are highly constrained by contextual and structural factors. That is, time, costs, and necessary interactions with clients push evaluators in similar directions, whatever their orientations" (2003, p. 53). He notes that, of course, theorists differ from practitioners, theorists "spend time searching for anomalies, problems, and new concepts. . . [they] earn their keep by creating ideas that are not common" (2003, p. 56). But, Christie finds that theorists' self-reports of their own practices are, in fact, more similar than those of practitioners. House is, perhaps, simply reminding us of theorists' *writings*. Is it context that influences the similarity in theorists' practice? Mark et al. (2000) and Mark (2002) have argued for a contingent theory of evaluation to better reflect the influences of context. From Stake's early writings on responsive evaluation (Stake, 1975), we know that evaluators should be flexible and adapt to the context of the program and stakeholders. But, we know little of how this is done. This case study of exemplary

evaluators provides an opportunity to examine some of those contextual factors. One of the elements of context that will be examined in more depth is the internal or external role of the evaluator.

Finally, I consider these cases in terms of the evaluators' involvement of stakeholders. House (2003) notes that Christie's dimension of stakeholder involvement fails to differentiate between breadth and depth of stakeholder involvement. This distinction, House argues, explains why Christie finds that he and Fetterman represent opposite anchors on stakeholder involvement. House's approach to stakeholder involvement is to insure breadth by including many different stakeholders, with the evaluator acting to protect the interests of less powerful stakeholders and to insure against bias, while Fetterman is more concerned with intensive involvement of one or two stakeholder groups, typically clients or program staff, for purposes of empowerment. These dimensions of breadth and depth of stakeholder involvement are helpful in analyzing the evaluation practices of these exemplary evaluators.

THE EXEMPLARS AS CASE STUDIES

Many writers have encouraged evaluators to direct the methodological skills that they typically apply to their evaluation projects to study evaluation practice itself (Cousins & Earl, 1999; Shadish, Cook, & Leviton, 1991; Worthen, 1990). Alkin argues for more descriptive theories of evaluation, describing actual *practices*, instead of the predominant prescriptive theories that explain how evaluations *should* be conducted (Alkin, 1991, 2003; Alkin & Ellett, 1985). Christie's research on the links between theory and practice is an important step in that direction. But her research, as Datta (2003) notes, uses a "high church quantitative methodology" (p. 43). Having theorists respond to the same survey items allows Christie to make interesting and useful comparisons across the eight theorists and a sample of practitioners. However, it fails to give us much of a picture of how these theorists actually practice evaluation.

The exemplar case studies, in contrast, are not nearly so standardized. Each focuses not only on one evaluator, but on one evaluation. Questions and probes differ. As such, my comparisons and reflections are limited by the nature of the case examined. Nevertheless, the limitations of these cases are also their strength: the focus on a single evaluation permits us to analyze these evaluators' practice in greater depth. In the interview, the evaluator is describing and reflecting on that one evaluation. The focus is on his or her practice in that particular case. In each case we gain an understanding of some of the important decisions or issues that the evaluator struggled with, the choices he or she made, and the factors that influenced his or her conduct of the evaluation. Perhaps of equal importance, we also gain a sense for the things they do not consider important, the issues they do not concern themselves with or even raise as issues.

As with Christie's selection of eight theorists, the evaluators studied here are not representative of all evaluators or even all exemplary evaluators. They were selected, over time, because they represented people with different approaches to evaluation, but the sample is purposive and, as such, is a case study of a select group of evaluators.

Each of the interviews treated here as a case involved several steps: an initial conversation between the evaluator and me to discuss the process and to identify a case (with the exception of the first three cases, which focused on evaluations that had won awards); my analysis of evaluation reports, articles, and other documents concerning the evaluation; the interview itself; and our review, discussion, revision, and editing of the transcript of the interview. The central portion of the case, the interview, was grounded in my review of evaluation reports and articles emerging from the evaluation. The core interview questions focused on critical

stages in any evaluation and were similar across interviews: the impetus for the evaluation, actions undertaken to plan and focus the evaluation, involvement of stakeholders, the choice and development of methods for data collection, interpretation of results, and dissemination and use. However, I added questions to each interview to explore issues I found intriguing from my review of the evaluation reports and my knowledge of the evaluator's theory. These questions were intended to gain a greater understanding of how these exemplary evaluators applied their theory or how they approached unusual aspects of their evaluation. Finally, during the interview, my review of the reports and documents allowed me to adapt and probe to get a better picture of the evaluator's actions, to obtain specific examples of a general theme, or to learn more about the evaluator's rationale or thoughts on a particular action.

Ideally, the cases could have been made more complete through interviews with key stakeholders involved in the evaluation; however, akin to the constraints that affect evaluations themselves, time and resources did not permit such investigations. Nevertheless, the cases, conducted by the same interviewer with a similar structure, provide useful qualitative information to learn more about evaluation practice and its links to context and theory. The focus on one evaluation, rather than the evaluator's practice in general, permits us to move beyond prescriptive information to richer detail on these exemplary evaluators' actual practice. Finally, some of these evaluators would certainly be considered evaluation theorists; others would not. As such, the focus here is not on whether these evaluators' practice reflects their own theory, though I will occasionally comment on this issue as relevant. Instead, the focus is on using the cases to describe what these evaluators do and, in so doing, to explore common and distinct elements in the evaluation practices of exemplary evaluators and, as appropriate, link them to existing evaluation theories.

The dimensions used to analyze these studies emerged both from the evaluation literature and from my analysis of the cases. They include context, purpose, advance organizers, breadth and depth of stakeholder involvement, method proclivity, and use of results.

SOME RESULTS AND OBSERVATIONS

Context

The evaluations selected by these evaluators shared one common and important characteristic that is not true of most evaluations: All were multiyear projects. The shortest was 30 months in length. Several evaluations took place over five or six years. While a couple had relatively small to mid-sized budgets, most were well funded and had extensive resources at hand through trained staff, money for travel and communication, and time. But differences emerge in organization, programmatic content and sphere, and the political prominence of the program.

Differences in the organizational base of the evaluation were obvious and had potentially important implications. Those conducted by Mueller, Greene, and Fetterman made use of fewer personnel, often the evaluator and a few others, and these evaluations were not based in large, established evaluation or research centers. In contrast, while varied in scope and program context, the evaluations conducted by Riccio, Rog, Henry, Bickman, and Donaldson required extensive staff and were based in research/evaluation/policy centers with established norms of practice. In regard to primary sphere of influence, some evaluated programs with national focus and included federal decision-makers (Rog, Bickman). Greene's evaluation had a multistate, or regional focus. But, the most common focus was on a single state (Donaldson, Henry, Riccio, Mueller). Some (Riccio, Henry) addressed very high-profile issues (state welfare reform and

school report cards) that would receive prominent attention in the media and among elected officials and high-level administrators.

Many important contextual elements can be identified, but I will describe one here. Henry's development of a performance monitoring system, or report card, on education for Georgia was influenced by several environmental factors in Georgia in the 1990s. The Georgia lottery monies provided a large and relatively secure funding stream for educational programs and the evaluation of such programs. Georgia Governor Zell Miller began his term with a positive focus on improving education in Georgia and, as a part of that effort, established the Council for School Performance. Henry's organization, the Applied Research Center in the Andrew Young School of Policy Studies, received the competitive award to carry out the Council's work. This positive, resource-rich environment, with policy makers interested in data on results, provided a fertile ground for good evaluation work. However, this case illustrates the nuances of context that can be informative. It is, for example, interesting to note that the work was carried out by a School of Policy Studies, not a state department of education or a university-based department of education. A fuller study of context would examine the factors influencing this choice and the ramifications of this choice on the development and implementation of performance measures.

Role of evaluators. One element of context that has been discussed frequently in the literature is the relationships of the evaluator to the organization or program, specifically whether the evaluator is internal or external to the organization whose program is evaluated (Fitzpatrick et al., 2003; Love, 1991; Mathison, 1999). In these case studies, all but one of the evaluators served in external roles. However, the cases reveal that the distinctions between internal and external evaluators often blur in practice as some have argued (King & Stevahn, 2002; Sonnichsen, 2000). Many of the evaluators in the "external" role (Greene, Rog, Henry, Mueller) had conducted other evaluations with their primary client and that experience influenced the conduct of the current evaluation. Greene, for example, commented, "I attribute some of the smoothness of the initial process to the fact that we had worked with some of the key developers of the institute before. They liked our prior work and expected something similar to what we proposed. They already had a familiarity with how we worked" (2001, p. 86). Her relationship with the program developers helped her to better understand their information needs and to form a cooperative, team-like relationship with them: "We had a partnership, a collegiality with the stakeholders. We weren't members of the program team, but we weren't outsiders" (2001, p. 85). Henry and Mueller also had on-going relationships with the programs or agencies they evaluated; Riccio was beginning such a relationship.

A more traditional external role was evidenced in Bickman's work with his primary client, the U.S. Department of the Army, who supplied the initial funding for the evaluation of systems of care for families at Ft. Bragg. He had not worked with the Department previously and was shocked at their preference for determining programs' success through examining process and costs. Bickman argued vigorously for measuring mental health outcomes in children, the ultimate program goal. Much infighting ensued, over outcomes, sample size, and other methodological issues. Describing these interactions, Bickman states, "It was a battle with the Army throughout the project to maintain the integrity of the design. In the end, however, they were very supportive of us because they now trusted our independence and our integrity. We actually received a rather large contract from them to conduct additional analyses of the data at the termination of the evaluation" (2002, p. 69). This quote typifies the argument for external evaluators—their independence, their willingness to take risks. However, one might argue that Bickman's role also demonstrates the drawbacks to the "pure" external role. Unlike Greene,

Donaldson, and Rog, Bickman distanced himself from people managing and implementing the program. As such, if the purpose of the evaluation had been more formative, he might have encountered difficulties in getting program staff to implement recommended changes.

Fetterman was the only evaluator serving in an "internal" role. A member of the education faculty at Stanford University, he evaluated Stanford's teacher education program (STEP) for the president of the university. However, here, too, we see a blurring of the role. Fetterman noted that he had some familiarity with the teacher education program he evaluated, but, while a faculty member in the School of Education, he was not a part of STEP. In fact, their offices were in different buildings. He described the STEP faculty members as colleagues, but did not work closely with them or their program. He viewed his internal, but not too internal, position as advantageous in establishing his credibility with Stanford administrators and providing him with important knowledge about the history of the program, its status, and the culture of the campus. However, the problems that can arise for an internal evaluator are demonstrated in Fetterman's tales of the evaluation. His final recommendations included involving STEP students in his own program on educational policy. Further, evaluating the teaching performance of his colleagues, which was a major focus of the evaluation, raised much political turmoil. While STEP faculty would likely have been troubled by such feedback from an external evaluator, the difficult relations that ensued can hardly have facilitated the recommended cooperation between STEP and Fetterman's program. The suspicion of bias would have been too great.

Purpose

The traditional definition of evaluation, determining merit and worth of a program, was a central purpose for five out of the eight case studies (Riccio, Fetterman, Greene, Bickman, Donaldson). For four of these (Bickman, Riccio, Fetterman, and Greene), determining merit and worth was a primary purpose. Bickman's evaluation of the system of care model for mental health services delivered to children and adolescents is, perhaps, a classic example of determining merit and worth. He notes, "She [Lenore Behar] was the driving force to develop a demonstration to test this system of care. She developed the term 'continuum of care' and she wanted to test it to prove it worked" (Bickman, 2002, p. 71). Riccio's primary stakeholders were members and staff of the California legislature who wanted to know if GAIN, their welfare reform system, reduced dependency on welfare. Fetterman's evaluation of the Stanford teacher education program was designed to help the new president of Stanford: "His primary concern was with the quality of what we [the teacher education program] were doing" (Fetterman, 2000, p. 246). [Note his use of the term "we", reflecting his internal role and identification with the program.] Greene describes the primary purpose of her evaluation as "the development of a good understanding of, and to make some judgments about, the quality and effectiveness of the leadership institute. . . . We hoped our work might inform the institute, but were probably more oriented toward developing a good description of how this institute was implemented, the ways it succeeded, and how it fell short" (2001, p. 86). Greene's discussion of purpose included describing and understanding the program in order to judge its quality, but Fetterman, Riccio, and Bickman, using different approaches, also grounded their evaluations in descriptive work. None of these evaluations were "black box" studies of outcomes, but, in each case, the description and understanding were secondary to judging merit and worth. In these cases the traditional evaluation purpose of determining merit and worth still played an important role in practice.

Further, the purpose of the evaluation had a strong impact on the evaluators' approach to the evaluation and their thinking about their study. For example, Greene observes, "Because this program dealt with contentious environmental issues, the personal stances of the evaluation team members on these issues were ever present. Our job, however, was not to engage in the substance of the environmental disputes, but rather to understand how this program was advancing a different process for resolving them" (Greene, 2001, p. 94). While Greene's language here differs from that that might be used by Riccio or Bickman, Greene illustrates that the purpose of the evaluation influenced her views of the evaluation object and her role. Her role was to determine the merit or worth of this process; her team chose to state their views regarding the merits of the approach more strongly in the final report than during the process. Although at least part of her team's reluctance to address these value conflicts came from not wanting to intrude on program people's "turf", her team's decision to hold back and observe the program model at work, as the developers intended it, is akin to Bickman's and Riccio's approaches. Bickman worked with stakeholders to develop iterations of program theory to understand what his descriptive data was showing, but only later commented on his own views regarding quality. Riccio, too, was watching and describing the counties' differing implementations. He, in fact, admits surprise at the final results. The approach that he and others who study welfare reform thought would work best did not.

Of course, purpose was heavily influenced by context as several evaluators note. In explaining his decision not to use an empowerment approach in the evaluation that served as the focus of our interview, Fetterman emphasizes, "The choice [of approach] depends on the situation you're in. . . . You can't just be a purist or dogmatic about your approach. The bottom line is: What's the purpose of the evaluation?" (2000, p. 242). Bickman comments, "I did not make a decision on that focus. . . . That's the disadvantage of program evaluation. You're given a program, a focus" (2002, p. 71). However, while purpose influenced evaluator's choices, it did not dictate those choices. All four evaluators took actions that they viewed as congruent with the purpose of determining merit and worth, but that were, in fact, quite different. Thus, Bickman argued forcefully to study the program's ultimate goals, changes in children's mental health, because he believes merit and worth cannot be determined solely by describing process and costs as the funders preferred. Fetterman, instead, argued for the importance of immersing himself and his staff in the program to understand the students' perspectives and chose not to directly study students' teaching performance. As noted above, Greene emphasized understanding the values and approach of the program. She did study program impact, but with a different approach than Bickman's. Her interpretation of merit and worth focused on identifying the potential for impact.

Although determining merit and worth was the most common *primary* purpose in these evaluations, every evaluation, with the exception of Bickman's, had some concern with improving the current program or organization. For some, this purpose was primary. Rog, Donaldson, and Mueller, with quite different approaches, all conducted evaluations that primarily served these formative purposes. Rog's evaluation of service-enriched housing programs for homeless families was undertaken when there was little information on homeless families. The evaluation served as both a needs assessment and, then, a formative program evaluation. Noting that there was little understanding or theory about homeless families and their needs, the purpose of her evaluation was "to develop bases of knowledge and, then, in turn use this knowledge to help guide program efforts" (Rog, 1999, p. 568). Donaldson's multiyear evaluation of four programs that were part of the Work and Health Initiative in California was very focused on program planning and improvement in the first few years. Both he and Rog ultimately ad-

dressed merit and worth, but their primary evaluation efforts were concerned with program development and improvement. Mueller's focus was on organizational development and capacity building through training and involving program staff in many phases of the evaluation, but specifically in data collection and analysis. The intent of these efforts, however, was to improve program delivery; she comments, "Staff participation in the evaluation helps family educators think critically about families and about what they're doing" (Mueller, 1998, p. 87).

Two cases were concerned with the purpose of accountability. Accountability was a primary purpose in Henry's development of performance measures of public schools in Georgia. The major goal of the system was "to put data in a format that school systems, educators, and parents could use and that they felt was a quick and easy way to get information" (Henry, 2000, p. 108). His purpose was not accountability in the traditional, narrow sense of reporting performance to those with oversight responsibilities, but in a broader sense of informing many stakeholders – not only policy-makers and high-level officials, but also teachers, parents, and the public – about the performance of schools. His methods were then influenced by this broader purpose. He worked to involve these many stakeholder groups and to design "report cards" that could be understood and used by parents and teachers. Fetterman noted a secondary purpose of accountability in reporting to his primary stakeholder, the president of the university, on what the teacher education program was doing. For others (Bickman, Donaldson, Rog, Greene, Mueller, Riccio), it is likely that the foundations or legislatures (state and federal) that had funded the programs were interested in accountability issues, but merit/worth and program improvement were more dominant themes.

Finally, like accountability, each evaluation could serve the purpose of knowledge development, but the emphasis in this analysis is on how evaluators characterized their purposes and how these purposes influenced their practice. In this context, several evaluators emphasized knowledge development, but in no case was it primary, which would place it in the realm of research. Knowledge development, however, played an important role in the evaluations of Bickman, Rog, and Donaldson. All three of these evaluators are proponents of theory-based approaches, and, as such, their comments and perspectives on the role of knowledge development in their evaluations were consistent with their theoretical approaches. Knowledge development was a particularly strong purpose in Bickman's evaluation, and he is explicit about that in the interview. In response to my question about purpose, he replies: "I saw it as primarily an evaluation project. It was a policy study, but we embedded several research questions in the evaluation. . . . Every good evaluation has the potential to be good research and to have policy relevance. . . . It's a waste not to see the research opportunities as part of an evaluation" (2002, pp. 69–70). But, his emphasis on policy and research differentiate his evaluation from those of the others. His focus was less on the particular program he was evaluating and more on the theory and policy implications. He states: "I'm not really testing the Ft. Bragg Demonstration. I'm testing the concept of a continuum of care" (Bickman, 2002, p. 77). Not surprisingly, then, his primary stakeholders were not those who make decisions or value judgments about the Ft. Bragg or the Stark County programs, but those who influenced children's mental health programs. He observes, "I have a lot of audiences – basic researchers in psychopathology, clinicians, and policy people" (Bickman, 2002, p. 78).

In contrast, Rog and Donaldson note how their results can contribute to knowledge about homeless families or programs for the unemployed, respectively, but their primary emphasis was with the program. For example, Donaldson expresses mixed feelings about his foundation's decision to emphasize replication over adaptation. Although he recognized

that replicating the program, without changes, was important for developing knowledge and program theory, he leaned toward adapting the program to the needs of California workers, quite different than those in Michigan. His comments demonstrated his primary emphasis on formative evaluation in that case. Similarly, Rog notes the dearth of knowledge and theories on homeless families and believes her evaluation will help remedy this problem, but her major allegiance was to informing the program managers and staff who were developing and adapting the current program. Bickman is vocal in his desire to improve children's mental health, but, in the conduct of his evaluation, saw policy change as the means toward achieving that goal.

Advance Organizers

These exemplary evaluators used an array of different advance organizers to frame their evaluations. For most, these advance organizers reflect their theories or approaches to evaluation; however, for others their approach was less directly linked to their own theories. Henry, Riccio, and Rog relied more on the needs of stakeholders as a primary mechanism than did others. Each of these evaluators worked extensively with stakeholder groups to learn of their views and information needs and to adapt the study to address these. Henry and Riccio both had prior experience in the content area of their evaluations, performance monitoring in education and welfare reform, and these experiences also influenced their approach, but they saw differences in the new context and made stakeholders' needs their focus. Rog saw her role as developmental and formative in an area in which there was not much information. Though she drew on her background in theory-based evaluation, it was not her fundamental advance organizer in this evaluation because of the absence of information and experience – hers and the stakeholders – with homeless families. Instead, she relied heavily on stakeholder input, from foundation and program people, with a view toward developing and adapting program practice and theory.

For others, their theory and experience was their primary advance organizer. Bickman's and Donaldson's evaluations were founded on the theory-based approach, though in different ways. Donaldson worked actively with managers and staff of each program to develop program theories. For programs where little theory existed, the theory-development phase appeared to be valuable to program people and the theory provided Donaldson with the foundation for the data collection phase of his evaluation. With Bickman's study, the theory of a continuum of care was already developed, but Bickman worked with the primary contractor and program people to develop different iterations of the model as they observed the program. The theories served as a guide for his description of program implement and quality. It also may have influenced his opinion on the need to measure ultimate outcomes to properly assess merit and worth. He emphasizes this point in the interview, "I did not know if the continuum of care would affect child and family outcomes, but I did know that was what the program claimed it would accomplish. If you're claiming to do policy-relevant research, you must look at what happens to people" (Bickman, 2002, p. 71).

Like Bickman and Donaldson, many of Greene's advance organizers emerged from her own theories of evaluation practice. Her writings emphasize the importance of a reflective methodology and an attention to values, and these principles can be seen as important advance organizers in her evaluation. However, both content expertise and program theory also served as advance organizers for her evaluation team and, ultimately, in determining the merit and worth of the program they evaluated. The content expertise of her colleagues (in leadership and

conflict resolution) was critical to the evaluation team. Program theory as an organizer emerged in her discussion of inconsistencies between program actions and its theory or philosophy of achieving change.

Mueller brought to her evaluation her experiences and views on the positive, capacity-building effects of staff involvement in evaluation on both the current and subsequent evaluations and on the program itself. Staff involvement, then, was a dominant advance organizer for her work. Her focus on staff involvement and capacity building as primary advance organizers is illustrated in contrasting her with Rog. Rog very actively involved staff and managers in the planning of the evaluation, but because her emphasis was not on capacity building, it was not a significant advance organizer. Rog initially proposed using case managers to conduct interviews with clients, but "the case managers were stretched beyond the maximum with the program" (1999, p. 569). Although Rog's advance organizer was not capacity building, she was concerned with staff's views as stakeholders and, as such, was more active in visiting sites and observing program actions than Mueller. The difference here is Mueller's advance organizer of capacity building made the training of staff a priority on a relatively tight budget. Rog's advance organizer of stakeholder needs made site visits a priority to directly learn of those needs by talking with and observing clients, staff, and managers at the sites.

As noted, Fetterman selected his *Exemplars* evaluation to demonstrate his belief that empowerment approaches are not always appropriate. Thus, he does not work to empower staff through evaluation. However, other advance organizers from the empowerment approach are evidenced in his decision to focus on discrepancies between program intentions (theories of action) and actions (theories of practice) (Fetterman, 2003).

Breadth and Depth of Stakeholder Involvement

Just as Christie's survey data showed that theorists were rather similar in their positive level of stakeholder involvement, these case studies illustrate that stakeholder involvement is a central component in these exemplary evaluators' practice. Donaldson, Henry, Rog, and Riccio involved many different stakeholder groups in the conduct of their evaluations, though Riccio's interview suggests slightly less depth of involvement than that of Donaldson, Henry, and Rog. Scope and context influenced the nature of that involvement. Donaldson's and Rog's evaluations took place at many sites and involved many potential stakeholders. In this complex setting, one could choose limited involvement of site-based stakeholders, as Bickman did, or establish formal structures to involve stakeholders from many sites. Donaldson and Rog chose the latter strategy, involving stakeholders in some depth in planning and implementation.

Donaldson's involvement came about partly through his work with stakeholders in developing program theory, but also through his desire to provide information for program improvement. Rog's stories illustrated the depth of involvement of program managers and staff in her evaluation and her deference to their needs. She recounts a case manager's response to a data collection strategy: The "case manager said, 'You're not capturing what I do!' with the listing of activities that we had. So, I said, 'Help me figure out how to understand what you do. What types of work do you do with families?' She and I spoke for a while . . . [and revised the form]. [M]any case managers [then] noted, 'That's more of what I do' " (Rog, 1999, p. 564). As illustrated in this quote, Rog's work was very personal and hands-on for a complex, multisite evaluation. In another telling anecdote, she describes disagreements across sites on whether to ask about domestic violence incidents in interviews with

clients. She suggests project directors debate this issue in a special session and describes the result: "With all the project directors together, the Portland project director was persuasive. The sites finally reached consensus, having all the domestic violence questions included in the family assessment interview" (Rog, 1999, p. 565). Note how Rog listened to their discussion and went with their decision. Donaldson not only involved many different stakeholders, but turned the tables to have program people give annual feedback on the evaluation and foundation.

Henry and Riccio also demonstrated a breadth of stakeholder involvement. Riccio worked with a wide array of groups beyond his primary audiences of legislators and state department people, including representatives of welfare rights organizations and child advocacy groups, schools, and other agencies that provided services to welfare clients. He and his staff briefed these groups and invited feedback on everything from workplans, designs, and data collection strategies to findings, draft reports, and final reports. Henry established separate groups of key stakeholders – principals, superintendents, teachers, and community members – to help identify indicators and plan the performance monitoring system. While they sometimes pulled these groups together, he illustrates his sensitivity to the differential power of groups and the effect of this power on group dynamics by having groups meet separately "to reduce the influence of pre-existing power relationships on the deliberations" (Henry, 2000, p. 109).

Others were less concerned with immediate stakeholder involvement. As noted earlier, Bickman's primary audiences were people influencing mental health policy. Program staff and clients or their representatives were not actively involved as stakeholders. He notes, "We kept our interactions with the treatment facility to a minimum" (Bickman, 2001, p. 71). Similarly, Fetterman was concerned with meeting the information needs of his key stakeholder, the president of Stanford, and met with him occasionally. He was interested in understanding the students' experiences and developed relationships with teachers in the field, but did not actively involve them or program staff – faculty teaching the program – in the planning of the evaluation. Draft reports were made available to faculty, but, perhaps because they were not extensively involved earlier, their responses were not positive.

Greene and Mueller were more concerned with depth of stakeholder involvement than breadth. The key stakeholder group for both consisted of those close to the program—program staff, developers, and implementers. Mueller's focus on depth of stakeholder involvement came through her training of staff to collect evaluation data; however, because she visited only a few sites, her interactions with staff seemed to be concerned more with the evaluation than with the program. In contrast, Greene worked "to establish and nurture an ongoing evaluative conversation with these key staff [program developers]" (Greene, 2001, p. 87). Context may play a role here. The program Greene evaluated had fewer sites and staff, three sites with two to three developers and implementers per site, than those addressed by Mueller, Rog, and Donaldson. This smaller size, which Greene may prefer, enabled her team to establish closer relationships with stakeholders. But, she also chose to involve fewer stakeholders. Commenting on the program developers, implementers, and other stakeholders, she notes, "They [developers and implementers] were the ones with whom we developed the collegial relationship. We were mindful of the Foundation and its philosophy and that did influence, to some degree, the way we framed the evaluation. We did, in the larger vision, see the program participants and their organizations as stakeholders; however, we didn't communicate with them throughout the process" (Greene, 2001, p. 86). Finally, while Greene involved her key stakeholder group in depth, some of the important interactions with this group were not in person as were those conducted

by Rog, Donaldson, and Henry and, thus, differed in style. For example, in commenting on these stakeholders' involvement with methodology, she notes "Each year we updated our evaluation design and shared it with them for their review. These key stakeholders were invited to comment on all of the instruments. We sent them drafts" (2001, p. 90). Similarly, evaluation reports were more likely to be written than oral. She comments: "I think it was a very good communication flow. Those we knew and worked with most closely read the reports. If they had questions, they would ask" (2001, p. 93). In the context of a smaller group, where closer relationships could develop, this format may have worked, but personal briefings and meetings to review results as established by Donaldson, Riccio, and Rog were necessary to involve key stakeholder groups in these larger projects. Again, context influences actions.

Method Proclivity

While differences in methodological preferences emerged, it is important to note that these evaluators *all* used a mix of qualitative and quantitative methods. Differences arose in the manner in which they talked about their results and the methods they found most influential. These differences reflect their different phenomenological approaches to research. Thus, Greene and Fetterman, whose theories reflect a constructivist view, placed more faith in their qualitative results. While both made use of quantitative surveys in their evaluations, they viewed their observations of the programs they evaluated as central to their methodology and were more involved with actually seeing what happened and experiencing the program personally than were other evaluators.

In his interview, Fetterman focuses on his immersion into the culture of the teacher education program. To evaluate the program, he felt he needed to understand the students' perspectives. He and his evaluation staff "spent every moment in the summer with STEP students—from seven in the morning until noon at the public school where they practiced teaching and in Stanford classes until seven at night, and then sorting data at night like students doing their homework" (Fetterman, 2000, p. 247). Fetterman's observation is clearly grounded in his training in anthropology. Greene's observation was also central to her methodology. She notes, "In terms of the broader goal of developing an understanding of the program and its accomplishments, there is no substitute for being there" (Greene, 2001, p. 88). However, her team members were more clearly observers than participants. Fetterman chatted with teachers about students' performance in the classroom. Greene and her two evaluation colleagues used a one-page observation guide; they then discussed their different observations. She viewed her choice of qualitative methods as necessary to be consistent with the values of the stakeholder and the program.

Unlike Fetterman, but like Bickman, Greene studied the final outcomes of the program she evaluated and viewed her mini-case studies of student use of the leadership training, and the impact of that use, as a critical part of the evaluation. Although Greene and Bickman were among the few to study ultimate program outcomes, their approaches were very different. Greene's mini-case studies sampled those students who were most likely to have made an impact in order to "learn whether, at its best, the institute had the potential to make a difference" (2001, p. 88). Each case study was, then, quite labor intensive, interviewing participants and others to see what happened in the field and to describe the participant's role in those events. Like Greene, Bickman selected his second site, Stark County, at least partly because the site "was proud of their program and was nationally recognized as a leader in the field" (2002, p. 77). He then used a more quantitative approach to study children and their families.

Bickman and Riccio and, to a lesser extent, Rog and Donaldson relied heavily on a quantitative base. Random assignment to groups is central to Bickman's and Riccio's study of program impact (or lack thereof). However, each of these postpositivist evaluators made extensive use of qualitative methods through interviews, existing data, and logs or checklists to describe program activities and gain an understanding of outcomes. Riccio's evaluation staff observed meetings between staff and clients. Commenting on the qualitative data, Riccio says "The field research complemented the survey data in important ways. It helped us understand better what we thought we were measuring with the staff survey and allowed us to illustrate some of the important differences in staff practices across the counties" (1997, p. 247.) But, he states, "using randomly assigned control groups helped us enormously in establishing our credibility" (1997, p. 249). While grounded in an experimental approach, he argues for the need for flexibility: "This was a long-term study and we had to remain flexible. We modified our workplan several times" (1997, p. 246) and gives examples of a variety of changes in methodology and approach.

Donaldson and Rog, also using mixed methods, had a slightly less quantitative emphasis. Rog relied on client interviews and existing data and, while her proclivity was quantitative, other comments reveal the importance she gave to qualitative data. Note the mix of methodological approaches in this remark: "We wanted to give each person the opportunity to have her voice, but we also wanted some standardization so that we could have more reliable information" (Rog, 1999, p. 569). Like Riccio, she believed her policy audiences, federal and state legislators, "would attach more credibility to quantitative data", but observes that the qualitative data "added 'color' and depth to the quantitative findings and, I think, increased our credibility to consumer and provider audiences" (Rog, 1999, p. 570). She saw the qualitative data as critical in demonstrating the vulnerability of the homeless women parents served by the program. Her mixed feelings about methodology are reflected in her final comments. When asked what she would change, she first expresses a desire for more experimental control to understand the role of services; however, she then notes that the pre-post changes in residential stability were so dramatic that, in the face of no other external efforts, they had to be considered program effects: "We were going outside the textbook to argue that, but the multisite nature of the study strengthened our ability to feel confident that the change was not due to selection factors, history, and so forth" (1999, p. 574). Note that her emphasis is on establishing causal relationships, but she is comfortable doing that with descriptive data, given the large changes demonstrated in this evaluation.

Donaldson's approach was similar to that of Rog, but with shades of Fetterman and Greene's emphases on observation. He went to different sites and observed different needs among participants across sites. He became very familiar with the details and actions of programs at various sites. He states, "I can't emphasize how important it is to pull away from the conceptual jargon and go see what's happening. I'm a strong believer in trying to observe the program in action and doing it in a way that you're not changing the way the program is delivered when you are not present. You want to get a realistic view of what's happening" (Donaldson, 2002, p. 353). While he personally observed programs in action and made use of structured observations in the formal evaluation, his primary data were quantitative, specifically, participant surveys and structured telephone interviews to determine employment status. His quantitative grounding is evident in the language he uses to describe developing program theory: "You're trying to consider whether the paths leading into this outcome have any chance at all for accounting for a significant portion of the variance in the desired outcome" (Donaldson, 2002, p. 353). He considers his most informative data to be the comparison between his

program numbers and those of the Michigan experimental study. Although he appears to regret the failure to adapt the Michigan model to California's workers' needs, he is pleased with the ability to compare his data with that obtained from an experimental study. Nevertheless, he made use of observation and interviews and states, "It's the interplay between the qualitative and quantitative that is really informative. In increasing our depth of understanding, it's the qualitative things that push us forward" (Donaldson, 2002, p. 360).

Henry's development of performance measures for schools was an evaluation, but not so much a program evaluation. It was an on-going performance measure and, as such, made use of existing, quantitative data. Nevertheless, Henry's learning experience in developing the measure provides insight into an evaluator's own changing views of both the meaning of evaluation and methodological proclivities. Henry notes that his experience in developing the performance measures led him to:

a broader view of evaluation and the activities that constitute an evaluation. . . . For this project I'm not as interested in teasing out whether this program worked or not. I'm not as interested in the causal attribution issue. We're trying to give solid information to people concerned with schools so they can make judgments about whether they are achieving the results they feel are important. If I can provide a good description based on outcome information, which can be digested in very little time with very little effort by parents, I've succeeded. . . . It was only through the process that I realized that my highest-order goal was a clear, low-cost description. Methods instructors who said that the highest-order goal is always a causal analysis were just wrong (2000, p. 117).

Use of Results

These case studies demonstrate that direct, instrumental use of evaluation results does occur. In fact, evidence of direct, instrumental use was demonstrated in every interview. In many projects – Greene, Donaldson, Rog, Mueller – program staff changed their activities because of evaluative feedback. In others – Riccio, Bickman, Fetterman – program administrators, funders, and elected officials changed their decisions about subsequent programs. Henry's performance management system did not address a particular stakeholder, but workshops for teachers stimulated use of the results to change teaching strategies. All these evaluators worked tirelessly to get their results used. Greene engaged in ongoing evaluative conversations with program people to prompt reflection. Riccio, Donaldson, and Rog held frequent meetings with stakeholder groups to present results and obtain feedback, though Donaldson observes that they should have killed fewer trees and made even more use of oral reports. Henry worked with the press to raise awareness of the school performance measures and even took personal phone calls from everyone from parents to superintendents. Mueller's on-site evaluators' use of stimulated response interviews prompted them to consider parents' learning and changes in different ways. In spite of the turmoil accompanying Fetterman's internal evaluation, the new director of the teacher education program adopted many of his specific recommendations. Differences in practice across these evaluators were largely those of dissemination by paper or electronic means versus personal interaction. These differences are worth exploring further, but the important link between theory and practice exhibited on this dimension in these cases is that these diverse evaluators all valued use and worked successfully to accomplish it.

SUMMARY AND DISCUSSION

These case studies complement the recent empirical work by [Christie \(2003\)](#). Just as Christie found considerable similarity across the evaluation theorists she examined, the practices of the exemplary evaluators explored in these cases were reasonably similar on Christie's dimensions of stakeholder involvement and method proclivity. Here, however, we gain a greater understanding of their differences. Six of the eight exemplary evaluators studied worked extensively to achieve stakeholder involvement in these evaluations. Two of these (Greene, Mueller) were concerned with achieving a depth of involvement from one or two groups. The remaining four (Riccio, Donaldson, Rog, and Henry) worked to achieve input and involvement from a larger number of stakeholders, taking their focus on breadth of involvement. On method proclivity, all eight evaluators used a mixed-method, adaptive approach to their evaluations. Further, their use of a mix of qualitative and quantitative methods was not gratuitous. They found value in each approach. However, they differed in the types of data they found most informative or persuasive in these evaluations. These preferences were sometimes small (Rog vs. Riccio) and other times great (Greene vs. Bickman). This result is not surprising. In fact, the strong and solid reliance on mixed methods should be more surprising. These cases illustrate that the exemplary evaluators who have been interviewed do have their methodological preferences, but they use their less-preferred methods and pay attention to these results as well.

Stufflebeam's dimensions of purpose and advance organizers showed stronger differences, including some surprises. Determining merit and worth was a common purpose that transcended methodological proclivity and influenced evaluators' approaches. Program or organizational improvement, however, was an important secondary purpose. Accountability and knowledge development emerged less frequently as primary or important secondary purposes, but were important purposes in a couple of the evaluations. Advance organizers can be grouped into two large categories: (1) stakeholders' needs and (2) organizers based on the evaluator's typical approach and/or training and education. Of course, overlaps occur here. Those who attended to stakeholders' needs as a primary advance organizer (Rog, Riccio, Henry) may typically use this approach and I am simply not aware of that. Also, those attending to stakeholders' needs were certainly influenced by their own training and experience, but a first and primary advance organizer for these evaluators appeared to be stakeholders' needs. Conversely, those using their typical approach also considered stakeholders' needs, but the influence of their theories, training and experience, and values were more evident in their evaluations than in those for whom stakeholders' values and needs were a prominent advance organizer.

Finally, the dimensions of use and role provided helpful information on these exemplary evaluators' practice of evaluation in these cases. All were concerned with use and, in fact, often achieved great instrumental use. This, in itself, is remarkable. But, it is not surprising given the amount of effort expended on this issue. Again, though, we see differences in how use was achieved that deserve further examination. These cases suggest, at least, that instrumental use can be achieved if the evaluator's role is congruent with the context. Greene, in a smaller context, was able to achieve some formative use through collegiality and evaluative conversations. Riccio, in a massive, lengthy evaluation, was able to achieve use through many, many presentations, discussions, and briefings with different, important stakeholder groups that he sought out. All but Fetterman served in an external role, but these cases show the fallacy of using the internal-external dimension in a simplistic way. Most of the external evaluators had

on-going relationships with their program that blurred the internal-external distinction, but probably mirrors much evaluation practice.

My reflections on these cases are tentative. However, I have completed my tenure as Editor of *Exemplars* and, in so doing, wanted to examine these case studies to see what we might learn. As noted, Christina Christie will be succeeding me as Editor of the *Exemplars* section of *AJE* and, through her interviews and mine, we hope to collaborate to learn more about how exemplary evaluators or theorists actually practice evaluation, how they draw from their theories and experience, and how they adapt to context to achieve exemplary evaluations. One of the wonderful things I learned from these columns is that exemplary evaluators bring not only skill, but also energy, joy, and insight to their work. I encourage you to go back and read the original interviews. The best parts are the stories – from Greene’s struggles with different values on environmental policy and leadership, to Bickman’s move from a hedonist to a skeptical optimist, to Rog’s case manager’s surprise that she [Rog] DID it! [changed the form], to Fetterman’s enthusiastic immersion into the role of student in Stanford’s teacher education program – each and every one were thinking, working, and learning to do their best to honestly evaluate a program and have an impact. Let me close with thanking them for their time, their stories, and their work.

REFERENCES

- Alkin, M. C. (1991). Evaluation theory development II. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*. Chicago: University of Chicago Press.
- Alkin, M. C. (2003). Evaluation theory and practice: Insights and new directions. In C. A. Christie (Ed.), *New directions for evaluation, No. 97. The practice-theory relationship in evaluation* (pp. 81–89). San Francisco: Jossey-Bass.
- Alkin, M. C., & Ellett, F. S., Jr. (1985). Evaluation models and their development. In H. J. Walberg & G. D. Haertel (Eds.), *International encyclopedia of education*. New York: Pergamon.
- Alkin, M. C., & House, E. (1992). Evaluation of programs. In M. C. Alkin (Ed.), *Encyclopedia of educational research*. New York: Macmillan.
- Bickman, L., & Fitzpatrick, J. L. (2002). Evaluation of the Fort Bragg and Stark County systems of care for children and adolescents. A conversation with Leonard Bickman. *American Journal of Evaluation, 23*, 69–80.
- Christie, C. A. (2003). What guides evaluation? A study of evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *New directions for evaluation, No. 97. The practice-theory relationship in evaluation* (pp. 7–35). San Francisco: Jossey-Bass.
- Cousins, J. B., & Earl, L. M. (1999). When the boat gets missed: Response to M.F. Smith. *American Journal of Evaluation, 20*, 309–317.
- Datta, L. (2003). Important questions, intriguing method, incomplete answers. In C. A. Christie (Ed.), *New directions for evaluation, No. 97. The practice-theory relationship in evaluation* (pp. 37–46). San Francisco: Jossey-Bass.
- Donaldson, S. I., Gooler, L. E., & Fitzpatrick, J. L. (2002). Theory-driven evaluation of the Work and Health Initiative: A focus on Winning New Jobs. Dialogue with Stewart Donaldson. *American Journal of Evaluation, 23*, 341–365.
- Fetterman, D. (2003). Fetterman-House: A process use distinction and a theory. In C. A. Christie (Ed.), *New directions for evaluation, No. 97. The practice-theory relationship in evaluation* (pp. 47–52). San Francisco: Jossey-Bass.
- Fetterman, D., & Fitzpatrick, J. L. (2000). Summary of the STEP evaluation. Dialogue with David Fetterman. *American Journal of Evaluation, 21*, 239–259.
- Fitzpatrick, J. L. (1997). Exemplary evaluations. *Evaluation Practice, 18*, 239–240.

- Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2003). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Pearson.
- Greene, J. C., & Fitzpatrick, J. L. (2001). Evaluation of the Natural Resources Leadership Program, 1995 through 1998. Conversation with Jennifer C. Greene. *American Journal of Evaluation*, 22, 81–96.
- Henry, G. T., & Fitzpatrick, J. L. (2000). The Georgia Council for School Performance and its performance monitoring system. Dialogue with Gary Henry. *American Journal of Evaluation*, 21, 105–117.
- Henry, G. T. (2000). Why not use? In V. J. Caracelli & H. Preskill (Eds.), *New directions for evaluation*, No. 88. *The expanding scope of evaluation use* (pp. 85–98). San Francisco: Jossey-Bass.
- House, E. R. (2003). Stakeholder bias. In C. A. Christie (Ed.), *New directions for evaluation*, No. 97. *The practice-theory relationship in evaluation* (pp. 53–56). San Francisco: Jossey-Bass.
- King, J. A. (2003). The challenge of studying evaluation theory. In C. A. Christie (Ed.), *New directions for evaluation*, No. 97. *The practice-theory relationship in evaluation* (pp. 57–67). San Francisco: Jossey-Bass.
- King, J. A., & Stevahn, L. (2002). Three frameworks for considering evaluator role. In K. E. Ryan & T. A. Schwandt (Eds.), *Exploring evaluator role and identity*. Greenwich, CT: Information Age Publishing.
- Love, A. J. (1991). *Internal evaluation: Building organizations from within*. Newbury Park, CA: Sage.
- Mark, M. M. (2002). Toward a comprehensive view of the theory and practice of program and policy evaluation. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium*. Mahwah, NJ: Erlbaum.
- Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey-Bass.
- Mathison, S. (1999). Rights, responsibilities, and duties: A comparison of ethics for internal and external evaluators. In J. L. Fitzpatrick & M. Morris (Eds.), *New directions for evaluation*, No. 82. *Current and emerging challenges in evaluation* (pp. 25–34). San Francisco: Jossey Bass.
- Mueller, M. R., & Fitzpatrick, J. L. (1998). The evaluation of Minnesota's Early Childhood Family Education Program. Dialogue with Marsha Mueller. *American Journal of Evaluation*, 19, 80–99.
- Riccio, J., & Fitzpatrick, J. L. (1997). MDRC's evaluation of GAIN: A summary. Dialogue with James Riccio. *Evaluation Practice*, 18, 239–252.
- Rog, D. J., & Fitzpatrick, J. L. (1999). The evaluation of the Homeless Families Program. Dialogue with Debra J. Rog. *American Journal of Evaluation*, 20, 558–575.
- Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, 17, 151–162.
- Shadish, W. R., Jr., Cook, T. D., & Leviton, L. D. (1991). *Foundations of program evaluation*. Thousand Oaks, CA: Sage.
- Sonnichsen, R. C. (2000). *High impact internal evaluation*. Thousand Oaks, CA: Sage.
- Stake, R. E. (1975). *Evaluating the arts in education: A responsive approach*. Columbus, OH: Merrill.
- Stufflebeam, D. L. (2001). *New directions for evaluation*, No. 89. *Evaluation models*. San Francisco: Jossey-Bass.
- Worthen, B. R. (1990). Program evaluation. In H. J. Walberg & G. D. Haertel (Eds.), *International encyclopedia of educational evaluation*. New York: Pergamon.