Metaevaluation in Practice

Selection and Application of Criteria

Leslie J. Cooksy
University of Delaware

Valerie J. Caracelli*
United States Government Accountability Office

ABSTRACT: This paper examines the practice of metaevaluation as indicated by the Metaevaluation standard of the Program Evaluation Standards, as the evaluation of a specific evaluation to inform stakeholders about the evaluation's strengths and weaknesses. The findings from an analysis of eighteen metaevaluations, including a description of the data sources and methods used to come to conclusions about the evaluation and the criteria of quality employed, are reported. A diverse set of practices were identified, ranging from the use of emergent criteria in a narrative review of information about an evaluation to the structured application of the Program Evaluation Standards using a checklist. The paper concludes that the evaluation field does not have a common understanding of metaevaluation practice.

KEYWORDS: *metaevaluation; metaevaluation criteria; metaevaluation practice; evaluation quality*

* The opinions in this paper should not be construed to be the policy or position of the U. S. Government Accountability Office.

Implicitly or explicitly, a concern with evaluation quality drives our discussions of evaluation models, methods, and practices. Explicit statements of what constitutes evaluation quality can be found in The Program Evaluation Standards (PgES) established by the Joint Committee on Standards for Educational Evaluation (1994)[2] and the American Evaluation Association Guiding Principles. While these are intended to be generally applicable, at least in the context of the United States, different evaluation models have different standards of quality (Stufflebeam, 2001b). For example, the meaningful engagement of all stakeholders would be valued under a constructivist approach, but perhaps less emphasized under certain postpositivist approaches (Guba & Lincoln, 1989; Schwandt, 1997). Quality criteria are also indicated in statements that evaluation should be systematic (Scriven, 1991; Shadish, Newman, Scheirer, & Wye, 1995; Weiss, 1998), transparent (Henry, 2001; U.S. Office of Management and Budget, 2002), balanced (Mixed-method Collaboration, 1994; Patton, 1997), relevant (Patton, 1997), culturally competent (Kirkhart, 2004), and so on. Clearly, quality criteria abound; but to what extent do we use them to systematically reflect on the quality of our work, that is, for metaevaluation?

Metaevaluations are systematic reviews of evaluations to determine the quality of their processes and findings (Bickman, 1997; Cook & Gruder, 1978; Greene, 1992; Leeuw & Cooksy, 2005; Lipsey, Crosse, Dunkle, Pollard, & Stobart, 1985; Scriven, 1991). When applied to a single study, metaevaluation may be conducted for formative or summative purposes (Greene, 1992; Joint Committee, 1994). A formative metaevaluation serves to improve the evaluation while it is underway, while a summative metaevaluation of a single study provides information about the strengths and weaknesses of the evaluation to evaluation clients and audiences and to the evaluator (Scriven, 2001; Stufflebeam, 2001a; Worthen, 2001). When applied to multiple evaluations, metaevaluation can be used as one of the first steps to an evaluation synthesis or to assess the evaluation capacity of a group or organization (Cook et al., 1992; Cooksy & Caracelli, 2005; Dickersin & Berlin, 1992; U.S. General Accounting Office, 1992; Wortman, 1994). The number of evaluations evaluated is only one feature of metaevaluation. In her conceptualization of metaevaluation, Bustelo (2002) identified the purpose and timing (relative to the evaluation) of metaevaluation, location of the metaevaluator, and evaluation phase that is metaevaluated as key descriptors of a metaevaluation. Additional dimensions of metaevaluation are the criteria of quality, the procedures for applying the criteria, and the audience for the metaevaluation findings.

Although metaevaluation has been identified as a hallmark of good evaluation for four decades (House, 1987; Joint Committee, 1994; Scriven, 1969, 1975, 1991, 2001; Stufflebeam, 1974, 1978, 2001a), the extent to which it is actually practiced is not clear. This paper explores the practice of metaevaluation in the sense implied by the metaevaluation standard of the PgES, which reads,

> The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, on completion, stakeholders can closely examine its strengths and weaknesses. (Joint Committee, 1994, p. 185)

Focusing on metaevaluations that are in the spirit of the PgES standard means that the research is limited to evaluations of single evaluations. The goal, specifically, is to understand the practice of metaevaluation that is conducted to improve an evaluation in process, reflect systematically on the strengths and weaknesses of an evaluation to enhance one's future evaluation practice, or provide

[2] In the third edition of *The Program Evaluation Standards*, metaevaluation is likely to be elaborated through a separate attribute on Evaluation Accountability.

information about the credibility of the findings to users. This purpose excludes metaevaluations conducted as a precursor to evaluation synthesis or as a method of determining evaluation capacity. This paper is based on an ongoing study of the nature and extent of metaevaluation practice; a later phase of the study will examine multiple-evaluation metaevaluations (Cooksy & Caracelli, 2007).

After a brief overview of our methodology, criteria that have been applied in specific metaevaluations and the methods of their application are presented and discussed. The paper then reflects on issues in the selection and application of quality criteria in single-evaluation metaevaluations and concludes with considerations for metaevaluation practice.

Methodology

We have used several strategies to identify metaevaluations for our study, including a call to EVALTALK (the listserv of the American Evaluation Association), a search of evaluation journals and ERIC, explorations of university Web sites, and individual contacts. These methods have yielded several metaevaluations and a great deal of literature about metaevaluation. However, we have obtained only eighteen metaevaluations that are relevant to this paper in the sense of being (1) evaluations of single evaluations conducted for the formative or summative purposes identified in the PgES and (2) identified specifically by the author as a metaevaluation or metaevaluative audit. The sample does not include metaevaluations conducted to develop an evaluation approach (see, for example, Brandon, 1998; Cleave-Hogg & Byrne, 1988; Curran, 2000; Hanssen, Lawrenz, & Dunet, 2008; Rebelloso, Fernández-Ramírez, Cantón, & Pozo, 2002) on the grounds that they are serving a research and development purpose that goes beyond the PgES intent. We also excluded the metaevaluation articles developed specifically for publication in the *American Journal of Evaluation*'s Metaevaluation Section (see Datta, 1999; Grasso, 1999; Sanders, 1999; Stake & Davis, 1999) because these were solicited primarily for educational and illustrative purposes. Although EVALTALK has an international reach and the journals that were searched included *Evaluation* and the *Canadian Journal of Evaluation*, none of the metaevaluations included in the sample are from outside the United States. (Our larger pool of metaevaluations, which will be the focus of future research, contains metaevaluations of groups of evaluations and includes several from outside the U.S.)

The metaevaluations in our sample come in multiple forms: nine reports of varying length and detail, eight articles, and one book. (A second book elaborates on the context of one of the studies described in an article.) The eighteen metaevaluations span four decades, with one conducted in the 1970s, six in the 1980s, six in the 1990s, and five in the 2000s. The metaevaluations included in our sample are listed in the reference list, marked by an asterisk. Appendix A also provides a description of the metaevaluations, including their evaluands and the variables that are discussed in the findings section below.

The documents in our sample have been coded using a semistructured instrument, which allows for variation in the kind and amount of information. The data collected for each metaevaluation were:

- Client for the metaevaluation (funder, evaluator, other)
- Timing (during and/or after)
- Purpose (formative and/or summative)
- Criteria
- Data sources
- Procedure for applying criteria
- Audience

In addition, we recorded a brief summary of the findings of the metaevaluations. Given the range of ways in which the metaevaluations are

documented, not all of these variables can be fully described. Of the variables recorded, this paper focuses on the criteria and procedures for applying the criteria, and also reports on purpose and data sources.

The limited number of metaevaluations included in the study means that the findings are useful primarily as illustrations of the range of metaevaluative criteria and the various methods with which they are applied. The sample does not support conclusions about the extent to which metaevaluation is practiced or about the frequency with which any specific set of criteria are used. Despite the limitations of the study, these early results may serve a heuristic purpose of stimulating discussion and reflection on metaevaluation practice.

Findings

The metaevaluations identified are varied in almost every way, however some characteristics are dominant. The most dominant feature relates to the purpose of the metaevaluation: Seventeen, all but one, have a summative purpose. In addition to the single metaevaluation with a formative purpose, five of the summative metaevaluations also have a formative role. There is also a clear penchant for external metaevaluators, with only two of the metaevaluations being conducted internally (i.e., by the evaluators themselves). More varied than purpose and location are the criteria and the ways in which the criteria are applied. These are discussed below, followed by an analysis of the association between criteria and methods of application.

Metaevaluation Procedures

To lay the groundwork for the discussion of criteria, this section describes the procedures used in the metaevaluations, specifically the data sources and the approaches to synthesizing the data to come to statements about evaluation quality. We found that almost all the metaevaluations used multiple data sources (see Table 1). We do not include the two internal metaevaluations (Lynch et al., 2003; Stufflebeam & Wingate, 2002) in this group. Since neither stated any specific sources of data, we have coded their data source as personal experience. It is likely that the metaevaluators reviewed reports and other documents and met with the evaluation team and other stakeholders. In contrast with the other metaevaluations, however, we do not know the extent to which these tasks, which are routinely performed as part of an evaluation, were conducted specifically for the metaevaluation. Smith (1999), whose metaevaluation was limited to an intensive review of evaluation reports, is also not included in the count of metaevaluations using multiple data sources.

Evaluation reports were a source for all but one of the sixteen external metaevaluations. The article by Greene, Doughty, Marquart, Ray, & Roberts (1988), actually a summary of two small audits, is the only external example that did not appear to include a report review because the metaevaluations were conducted simultaneously with the production of the reports. In most cases the reports were reviewed in draft form so that the metaevaluation findings could be used to improve the final report. The exceptions are listed in the note to the table.

Table 1
Sources of Data Used in the Metaevaluations (N = 18)

Metaevaluation Data Sources	# of Metaevaluations
Report review*	15
Review of other documents	15
Interviews/meetings with evaluation team	13
Review/reanalysis of original data	10
Interviews/meetings with other stakeholders	6
Replication of data collection activities (e.g., site visits; survey tests)	3
Personal experience	3

*These were mostly draft reports with the metaevaluation findings contributing to improvements to the final report. However, three appeared to focus on reports that had already been finalized (Farrar & House, 1987; McKinley, 1999; Stake, 1986).

In addition to reports, other documents reviewed included original requests for proposals, design papers, data collection instruments, coding schemes, internal memos, and others. Interviews with the evaluation team were less common, but still used in more than half of the metaevaluations. These interviews took the form of either individual interviews (in the case studies by Farrar and House, 1987, and Stake, 1986, for example) or meetings with an evaluation team. An example of the latter is what Kemmis (1997) called a "metaevaluation court." In this case, the metaevaluator convened a meeting with the evaluation team and representatives from the evaluation advisory committee and the client organization "to raise questions about matters which might throw doubt on the dependability of the findings" (p. 1). The metaevaluation court also counts as a meeting with other stakeholders, but in general interviews or meetings with other stakeholders were not a commonly reported source of data for the metaevaluations in our sample.

More than two thirds of the metaevaluations included some examination of evaluation data. Most of these consisted of the kind of review of data that is common in an evaluation audit (discussed further below). Descriptions of reanalysis were rare, limited to House (1987) and House, Glass, McLean, and Walker (1978)—the two metaevaluations that were specifically described as using an expert panel of reviewers. Finally, in some cases, the metaevaluators replicated the data collection activity (Finn, Stevens, Stufflebeam, & Walberg, 1997; Hartmann & Loizides, 2001; House, 1987). Hartmann and Loizides, for example, went through the process that a survey respondent went through in their metaevaluation of the survey component of an evaluation. (We have included House in this group although it is not clear if the observations of classrooms he describes were in advance of or subsequent to classroom data collected by the evaluators.)

The metaevaluators synthesized the data obtained from these various sources to come to conclusions about the quality of the evaluation. We have categorized the methods used for these syntheses as (1) narrative reviews, (2) semistructured reviews, (3) checklists, and (4) evaluation audits. A brief overview of these methods is given here. The procedures and the criteria associated with them are discussed in greater detail in the section that follows. The eight metaevaluations that we have categorized as narrative reviews include those that have identified themselves as using case studies (Farrar & House, 1983; Stake, 1986), expert panels (House, 1987; House et al., 1978), and a "critical friend" approach (Smith, 1999), as well as two that do not name a specific approach, although they do describe their procedures (Burbules, 2000; Greene, 1999; Kemmis, 1997).

Although this is a disparate set, they are linked by the use of very generally stated guidelines rather than explicit criteria and by dependence on the reviewers' experience and expertise. It could be argued that the selection of metaevaluator implies a quality focus, as when House included a measurement expert in the metaevaluation of the Follow Through evaluation (House et al., 1987).

Two of the metaevaluations used semistructured reviews, in which criteria are prespecified, but the process of synthesizing material is left to the expertise of the metaevaluator (Lynch, 2003; Migotsky & Stake, 2001). Four used the audit approach described by Guba and Lincoln (1981) and Schwandt and Halpern (1988) (Greene et al., 1988; Greene et al., 1992; Ray, 1988; Whitmore & Ray, 1989). We have distinguished an audit, which Schwandt (1989) defines as "a systematic examination of the procedures and reports of an evaluation…[requiring] a set of audit procedures, audit standards, and some agreed-upon criteria for judging evaluation quality" (p. 34), from a semistructured review in two ways. First, they describe themselves as evaluation audits. Second, being guided by Guba and Lincoln and/or Schwandt and Halpern, they are more structured in their procedures than the semistructured reviews. The remaining four metaevaluations used a checklist approach to synthesizing the data on evaluation quality (Finn et al., 1997; Hartmann & Loizides, 2001; McKinley, 1999; Stufflebeam & Wingate, 2002).

Metaevaluation Criteria

Mirroring the procedures for synthesizing the metaevaluation data, the criteria used in the sample of metaevaluations ranged from predetermined and structured to emergent and unstructured. As shown in Table 2, there is a strong association between the metaevaluation procedure and the metaevaluation criteria used. The table organizes the criteria used in the eighteen metaevaluations into four categories: emergent; tailored; the PgES; and the trustworthiness criteria described by Guba and Lincoln (1989).

Table 2
Metaevaluation Method by Metaevaluation Criteria (N = 18)

	Metaevaluation Criteria			
Method	**Emergent**	**PgES**	**Tailored**	**Trustworthiness**
Narrative reviews	7		1	
Checklists		4		
Semi-structured reviews		1	1	
Evaluation audits			1	3

Seven of the metaevaluations used narrative reviews of information about the evaluations to identify *emergent criteria*. These varied from Burbules' (2000) inductive identification of gaps in the information that the evaluation provided in a way that indicates that quality includes the extent to which the evaluation addresses the primary purposes of the program to the technical quality issues raised in the metaevaluations by House et al. (1978) and House (1987). This set of metaevaluations also included the two case studies conducted by Farrar and House (1987; see also House, 1988) and Stake (1986) which, while primarily emergent, included an explicit focus on the match of the evaluation as implemented to the stakeholder-based evaluation model. The final metaevaluation in this category is described by Smith (1999), the critical friend whose narrative review of evaluation reports led to comments

about contextual issues, potential areas for drawing larger lessons, the use of recommendations, and other aspects of the evaluation.

Five of the metaevaluations used the PgES as quality criteria (Finn et al., 1997; Hartmann & Loizides, 2001; Lynch et al., 2003; McKinley, 1999; Stufflebeam & Wingate, 2002).[3] As most evaluators know, the PgES are comprised of thirty prescriptive statements organized by attributes of utility, feasibility, propriety, and accuracy (see www.wmich.edu/evalctr/jc/ for a list of the statements.) The PgES were developed through an extensive process of consultation and review and are internationally recognized. (The American Evaluation Association and the Canadian Evaluation Society are among the sponsors of the Joint Committee, the Board of the Australasian Evaluation Society has endorsed the standards, and the African Evaluation Society used them as a starting point for the African Evaluation Guidelines.) The specificity of the PgES makes their application in the form of a checklist straightforward. Stufflebeam (1999a, 1999b) has developed detailed checklists based on the PgES in which each standard is broken out into six (short version) or ten (long version) elements ("checkpoints"). However, the four metaevaluations in our sample that applied the PgES using a checklist approach used simpler forms. Finn et al. rated the evaluation's compliance with each standard with a scale of not met, partially met, met, or insufficient information. In addition to the ratings, the metaevaluation described notable strengths and weaknesses for each of the four attributes. The other three used the form provided in *The Program Evaluation Standards* (1994), which rates each standard as fully addressed, partially addressed, not addressed, or not applicable, and provided a rationale for each rating. The metaevaluation by Hartmann and Loizides is a special case. It focused specifically on the Internet survey component of an evaluation, so the metaevaluators applied a checklist of Dillman's (2000) guidelines for Internet surveys in addition to the PgES. The fifth metaevaluation that used the PgES selected a subset of standards (some from each attribute) and used them as a semistructured guide to an internal reflection on the strengths and weaknesses of the evaluation (Lynch et al., 2003).

Criteria that are *tailored* to a specific evaluation were applied in a narrative review (Greene, 1999), an audit (Greene et al., 1992), and semistructured review (Migotsky & Stake, 2001). Greene (1999) described three sets of metaevaluative criteria used in the narrative review: (1) criteria intrinsic to the responsive evaluation approach (the approach being used in the evaluation), such as the richness and detail with which the quality of the evaluand is described; (2) criteria "commonly accepted in the larger evaluation community and generally consonant with responsive ideals" (p. 7), such as the usefulness of the evaluation; and (3) criteria "valued by me as an evaluator and generally consonant with responsive ideals" (p. 7), including fair inclusion and representation of as many stakeholders as possible, particularly those whose views are often not included.

In their audit of the evaluation of the East Central AIDS Education and Training Center (ECAETC), Greene et al. (1992) reviewed an evaluation progress report and other materials and interviewed the evaluator to develop five guiding questions as a focus for their metaevaluation. Two of the questions concern "methodological quality and supportability of evaluation results," two "address the program improvement role" of the evaluation, and the fifth "provides a summary assessment of evaluation quality" (pp. 84-85). For their semistructured metaevaluation, Migotsky and Stake (2001) developed fourteen guiding

[3] Finn et al. (1997) used the first edition of the PgES, *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981). The others used the standards described in the second edition, *The Program Evaluation Standards* (Joint Committee, 1994).

questions, which addressed such issues as the accordance of the fieldwork with the contract, the competence of team members individually and as a team, efforts to make site visits and subsequent reports useful to site projects, orientation of the site visit protocol to actual activities, and so on.

Three metaevaluations (Greene et al., 1988; Ray, 1989; Whitmore & Ray, 1989) used the trustworthiness criteria of confirmability and dependability and applied them using auditing procedures. According to Greene et al. (1988; citing Guba and Lincoln, 1981, and Lincoln & Guba, 1985), "an external evaluation audit is a metaevaluative activity intended to judge the quality of an interpretive evaluation study. The dimensions of quality targeted by an audit are dependability or methodological soundness and confirmability or the substantive grounding of conclusions in the data" (p. 81). Schwandt (1997) defines dependability as "focused on the process of the inquiry and the inquirer's responsibility for ensuring that the process was logical, traceable, and documented" (p. 164). A dependability audit involves a review of the evaluation data and documentation of design, methodological, and analytic decisions to "assess underlying logic and defensibility and for adherence to professional standards, and to assess the degree of inquirer bias" (Whitmore & Ray, 1989, p. 79). Schwandt defines confirmability as focused on "linking assertions, findings, interpretations, and so on to the data themselves in readily discernable ways" (p. 164). A confirmability audit reviews the analytic categories that have emerged from the data and determines whether findings and conclusions are clearly grounded in the data and follow from the analysis (Guba & Lincoln, 1989; Whitmore & Ray, 1989).

In two cases (Ray, 1988; Whitmore & Ray, 1989), credibility was also used as a sign of quality. Guba and Lincoln (1989) describe credibility as a trustworthiness criterion that is focused on "establishing the match between the constructed realities of respondents (or stakeholders) and those realities as represented by the evaluator and attributed to various stakeholders" (p. 237). Credibility has a specific meaning within the constructivist paradigm, but is commonly referred to as a metaevaluative criterion outside of the paradigm as well, with varied definitions. For example, Stufflebeam wrote, credibility "concerns whether the audience trusts the evaluator and supposes him to be free of bias in his conduct of the evaluation" (1974, p. 8-9). In a somewhat different take on the construct, Patton (1997) defines credibility as "a complex notion that includes the perceived accuracy, fairness, and believability of the evaluation" (p. 250). In a metaevaluation project conducted for synthesis purposes involving multiple evaluations, we used credibility (operationalized as the extent to which inferences were supported by evidence, similar to the confirmability criterion of trustworthiness) because it is an important criterion across various design options, and it is translatable across different representations of evaluation practice (Cooksy & Caracelli, 2005).

For the most part, the studies do not explain why one set of criteria or one method of applying the criteria was selected over another set or method. Only the documents about the evaluation audits discuss the rationale underlying the choices. For example, in the metaevaluation conducted by Greene and her colleagues (1992), trustworthiness criteria were identified as relevant in their process of developing tailored criteria, but they did not start out with these criteria because the evaluation was not conducted within an interpretive paradigm. On the other hand, while evaluation paradigm was a consideration in Greene et al.'s selection, a lack of an interpretive framework for the evaluation metaevaluated by Ray (1988) was not considered an obstacle to using credibility, confirmability, and dependability criteria.

Discussion

Defining Metaevaluation

The metaevaluations in our sample vary in criteria and method, from the use of the PgES to reflect on the evaluation's strengths and weaknesses, to detailed case studies of the context, experiences, and other aspects of an evaluation, to short reports summarizing the findings on emergent or tailored criteria. Despite casting a wide net, we identified only eighteen metaevaluations that met our primary definition of evaluating a single evaluation and self-identifying as a metaevaluation. Moreover, with one exception (Lynch et al., 2003), the metaevaluations were associated with only three sources—the University of Illinois' Center for Instructional Research and Curriculum Evaluation, Western Michigan University's Evaluation Center, and Cornell University's graduate program in evaluation.

What we did not include, however, is the informal practice of metaevaluation. For example, in response to our posting on EVALTALK, we received a personal e-mail from an evaluator saying that she uses the PgES to guide her evaluation and her reflections on the evaluation's strengths and weaknesses once it is complete. In other personal communications, Stake wrote to us, "I cannot think of anything else we called a metaevaluation, but of course there has been informal metaevaluation in everything we have done;" and House said, "Actually, there are many so-called audits. Of course, most are not well documented." Our sample also did not include internal quality controls, such as Peshkin's (1988) self-monitoring methods or managerial reviews of evaluations or registries that rate the quality of studies, such as the Campbell Collaboration. The dearth of documented cases of metaevaluation that met our criteria is a problem for research on metaevaluation, but does not necessarily mean that the evaluation field is not paying sufficient attention to evaluation quality.

Another definitional issue relates to the degree to which the metaevaluation is truly evaluative. Greene et al. (1988) ask, "How many questions must be answered in the negative or how serious must an evaluator's errors be for the evaluation to 'fail' the audit?" (p. 371). Our sample was limited to metaevaluations that were identified by their authors as metaevaluations, but Smith (1987) questioned the identification of *Quieting Reform* (Stake, 1986) as a metaevaluation on the grounds that "it is difficult to arrive at an *overall* sense of the evaluation's worth" (p. 364). In contrast, Stufflebeam (1999a, 1999b) identifies specific standards that are essential for an evaluation to meet if it is to "pass" the metaevaluation. However, if metaevaluations require an explicit, overall or integrated statement of an evaluation's merit or worth, then we would lose more than *Quieting Reform* from our already small sample.

Ways of Thinking about Criteria

As discussed in the introduction and demonstrated by the metaevaluations that have been described, conceptions of quality can differ. These different ways of thinking about quality reflect the diverse paradigmatic stances and sometimes conflicting views about what goals evaluation should pursue and what approaches should be used (Benson, Hinn, & Lloyd, 2001; Schwandt, 1989). In our sample, a sort of goal-free metaevaluation can be found in the form of the emergent criteria. While rigor, in Schwandt and Halpern's (1988) sense of "the right methods properly applied" (p. 54), was the most consistent focus of these emergent criteria, questions of responsiveness to local context, fairness, efficiency, and effects on program efforts also surfaced. For those metaevaluations that used a more structured approach, the limited use of quality indicators that have been developed for specific kinds of

evaluation is notable. For example, Western Michigan University maintains a Web site with several checklists tailored to different evaluation models (www.wmich.edu/evalctr/checklists), and Datta (1997) has developed checklists for reviewing mixed-method evaluations and case study reports. From a practical perspective, an evaluator contracting with an external metaevaluator may want to be sure that the criteria used in a metaevaluation are made explicit or at least that there is an agreement on the goal-free nature of the review (Cooksy & Caracelli, 2005; Schwandt, 1989; Whitmore & Ray, 1989).

Forms of Single-Evaluation Metaevaluation

Although our sample does not provide much information about the use of the metaevaluations, it is likely that different forms of metaevaluation, that is, different clusters of characteristics (timing, evaluator location, method, criteria, data sources), would be more or less appropriate for different uses and users. For example, a self-administered checklist (with verifiable justifications for the ratings) seems well-suited to an accountability purpose. In contrast, a goal-free metaevaluation, using emergent criteria applied by an external evaluator in a narrative review, can provide a full range of strengths and weaknesses that may be useful in improving a final report. If the metaevaluation is intended to establish the evaluation's credibility to an external audience as well as strengthen the final report, however, the use of some tailored or prespecified quality criteria could provide a stronger basis for conclusions about the evaluation's quality. While one hopes that the intended purpose of the metaevaluation is a primary determinant of its form, other potential influences on the choice of form include funding, metaevaluator competency, and other resources. Most obviously, it is cheaper to complete a checklist as an internal evaluator than to hire an external evaluator to do a tailored metaevaluation.

Conclusion

In 1987, House said, "I believe we are at the point in the development of the evaluation profession where we need some quality control measures. The various written standards are fine—well done I think—yet they do not go far enough in assuring a high quality product" (p. 55). Based on our research thus far, it is not clear that metaevaluation has become standard practice for individual evaluations (except perhaps in the form of the kind of management quality control reviews that occur in government agencies such as the U.S. Government Accountability Office or large consulting firms). The variety in our sample also suggests that a lack of clarity about what constitutes a metaevaluation: Are peer reviews metaevaluations? Does a semiformal reflection on the PgES qualify? If specific standards have been used to guide the evaluation, when is it sufficient to simply say so and when does the statement need external verification? As our research program expands to examine the wider pool of metaevaluations and metaevaluative practices, we hope to learn more about how the evaluation field views and practices metaevaluation. Metaevaluation is, as Stufflebeam (2001) stated, a professional imperative. As such, it is imperative that we develop some common understandings about its defining characteristics.

References

Benson, A. P., Hinn, D. M., & Lloyd, C. (Eds.). (2001). Visions of quality: How evaluators define, understand and represent program quality. *Advances in Program Evaluation* (Vol. 7). New York: Elsevier.

Bickman, L. (1997). Evaluating evaluation: Where do we go from here? *Evaluation Practice*, *18*(1), 1-16.

Brandon, P. (1998). A meta-evaluation of schools' methods for selecting site-managed

projects. *Studies in Educational Evaluation*, *24*(3), 213-28.

*Burbules, N. C. (2000). Meta-evaluation of the Milwaukee Teacher Education Center evaluation. In M. Chandler, R. Stake, M. Montavon, G. Hoke, R. Davis, J-H. Lee, & S. Rierson, *Final report: Evaluation of the MTEC Alternative Teacher Education program (*Appendix A). Chicago: University of Illinois, Center for Instructional Research & Curriculum Evaluation.

Bustelo, M. (2002, October). *Metaevaluation as a tool for the improvement and development of the evaluation function in public administrations*. Paper presented at the meeting of the European Evaluation Society Conference, Seville. Retrieved October 24, 2007, from http://evaluationcanada.ca/distribution/20021010_bustelo_maria.pdf

Cleave-Hogg, D., & Byrne, P. N. (1988). Evaluation of an innovation in a traditional medical school: A metaevaluation. *Evaluation & the Health Professions, 11,* 249-271.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmannn, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation research. *Evaluation Quarterly*, *2*(1), 5-51.

Cooksy, L. J., & Caracelli, V. J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, *26*(1), 31-42.

Cooksy, L. J., & Caracelli, V. J. (2007, November). *The practice of metaevaluation: Does evaluation practice measure up?* Panel presentation at the meeting of the American Evaluation Association, Baltimore, MD.

Curran, V. R. (2000). An eclectic model for evaluating Web-based continuing medical education courseware systems. *Evaluation & the Health Professions, 23*, 318-347.

Datta, L. (1997). Multimethod evaluations: Using case studies together with other methods. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 344-359). Thousand Oaks, CA: Sage.

Datta, L. (1999). CIRCE's demonstration of a close-to-ideal evaluation in a less-than-ideal world. *American Journal of Evaluation, 20,* 345-354.

Dickersin, K. & Berlin, J. A. (1992). Meta-analysis: State-of-the-science. *Epidemiologic Reviews*, *14*, 154-176.

Dillman, D. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). San Francisco: Wiley.

*Farrar, E., & House, E. R. (1983). The evaluation of Push/Excel: A case study. In A. S. Byrk (Ed.), Stakeholder-based education (pp. 31-57). *New Directions for Program Evaluation, 17*.

*Finn Jr., C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. In H. L. Miller, Jr. (Guest Ed.), The New York City Public Schools Integrated Learning Systems Project: Evaluation and meta-evaluation. *International Journal of Educational Research*, *27*(2), 159-174.

Grasso, P. G. (1999). Meta-evaluation of an evaluation of reader focused writing for the Veterans Benefits Administration. *American Journal of Evaluation, 20*, 355-371.

Greene, J. C. (1992). A case study of evaluation auditing as metaevaluation. *Evaluation and Program Planning*, *15*(1), 71-74.

*Greene, J. C. (1999). *Meta-evaluation: Evaluation of the VBA Appeals Training Module.* Unpublished paper, University of Illinois at Urbana-Champaign.

*Greene, J. C., Doughty, J., Marquart, J. M., Ray, M. L, & Roberts, L. (1988). Qualitative evaluation audits in practice. *Evaluation Review, 12*, 352-375.

*Greene, J. C., Dumont, J., & Doughty, J. (1992). A formative audit of the ECAETC year 1 evaluation: Audit procedures, findings, and issues. *Evaluation & Program Planning, 15,* 81-90.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. San Francisco: Jossey-Bass.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage

Hanssen, C. E., Lawrenz, F., & Dunet, D. O. (2008). Concurrent metaevaluation: A critique. *American Journal of Evaluation, 29*, 572-582.

*Hartmann, D., & Loizides, G. (2001). Metaevaluation of the Web-based ATE survey evaluation system. Western Michigan University: Kercher Center for Social Research. Retrieved July 24, 2008, from the Western Michigan University Evaluation Center Web site: http://www.wmich.edu/evalctr/ate/webbasedmetafinal.pdf

Henry, G. T. (2001). How modern democracies are shaping evaluation and the emerging challenges for evaluation. *American Journal of Evaluation*, *22*(3), 419-429.

*House, E. R. (1987). The evaluation audit. *Evaluation Practice*, *8*(2), 52-56.

House, E. R. (1988). *Jesse Jackson and the politics of charisma: The rise and fall of the PUSH/Excel Program*. Boulder, CO: Westview.

*House, E. R., Glass, G. V., McLean, L. D., & Walker, D. F. (1978). No simple answer: A critique of the Follow Through evaluation. *Harvard Educational Review, 48*(2), 128-160.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for the evaluation of educational programs, projects, and materials*. New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage.

*Kemmis, S. (1997). Metaevaluation executive summary. In R. Stake, R. Davis, & S. Guynn, *Evaluation of Reader-Focused Writing for the Veterans Benefits Administration* (pp. 143-148). Retrieved June 25, 2008, from http://www.ed.uiuc.edu/CIRCE/RFW/15metaeval.pdf

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. *American Journal of Evaluation*, *16*, 1-12.

Leeuw, F. L. & Cooksy, L. J. (2005). Evaluating the performance of development agencies: The role of metaevaluations. In G. K. Pitman, O. N. Feinstein, & G. K. Ingram (Eds.), Evaluating development effectiveness (pp. 95-108). *World Bank series on evaluation and development (Vol. 7)*. New Brunswick: Transaction.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. In D. S. Cordray (Ed.), Utilizing prior research in evaluation planning (pp.7-28). *New Directions for Program Evaluation, 27*.

*Lynch, D. C., Greer, A. G., Larson, L. C., Cummings, D. M., Harriett, B. S., Dreyfus, K. S., & Clay, M. C. (2003). Descriptive metaevaluation: Case study of an interdisciplinary curriculum. *Evaluation & the Health Professions, 26,* 447-461.

*McKinley, K. H. (1999). *Metaevaluation report of the Evaluation of the Michigan Public School Academy Initiative*. Retrieved July 24, 2008, from the Western Michigan University Evaluation Center Web site: http://www.wmich.edu/evalctr/charter/reports/metaeval.html

*Migotsky, C., & Stake, R. (2001). An evaluation of an evaluation: CIRCE's metaevaluation of the site visits and issue papers of the ATE program evaluation. Retrieved July 24, 2008, from the Western Michigan University Evaluation Center Web site: http://www.wmich.edu/evalctr/ate/sitevisitmetafinal.pdf

Mixed-method Collaboration. (1994). Mixed-method evaluation: Developing quality criteria through concept mapping. *Evaluation Practice*, *15*, 139-152.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Peshkin, A. (1988). In search of subjectivity—one's own. *Educational Researcher*, *17*, 17-21.

*Ray, M. L. (1988). Evaluation audit. In L. J. Cooksy, *The nature of participation in a mandatory employment assistance program* (Appendix). Unpublished doctoral dissertation, Cornell University, New York.

Rebolloso, E., Fernández-Ramírez, B., Cantón, P., & Pozo, C. (2002). Metaevaluation of a total quality management evaluation system. *Psychology in Spain*, *6(1)*, 12-25.

Sanders, J. R. (1999). Metaevaluation of "The Effectiveness of Comprehensive, Case Management Interventions: Evidence from the National Evaluation of the Comprehensive Child Development Program." *American Journal of Evaluation*, *20*(3), 577-582.

Schwandt, T. A. (1989). The politics of verifying trustworthiness in evaluation auditing. *American Journal of Evaluation, 10*(4) 33-40.

Schwandt, T. A. (1997). *Qualitative inquiry: A dictionary of terms*. Thousand Oaks, CA: Sage.

Schwandt, T. A. & Halpern, E. S. (1988). *Linking auditing and metaevaluation: Enhancing quality in applied research. Applied Social Research Methods Series, 11*. Thousand Oaks, CA: Sage.

Scriven, M. (1969.). Introduction to metaevaluation. *Educational Product Report*, *2*, 36-38.

Scriven, M. (1975). Evaluation bias and its control. *Occasional Paper #4*. Retrieved August 23, 2008, from the Western Michigan University Evaluation Center Web site: http://www.wmich.edu/evalctr/pubs/ops/ops04.pdf

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Scriven, M. (2001). Evaluation: Future tense. *American Journal of Evaluation*, *22*(3), 301-307.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (Eds.). (1995). *Guiding principles for evaluators. New Directions for Program Evaluation, 66*.

*Smith, L. M. (1999). *Meta-evaluation of the Veterans Appeals Training program evaluation*. Unpublished manuscript, Washington University, St. Louis, MO.

Smith, N. L. (1987). Book Review: Quieting reform: Social science and social action in an urban youth program by Robert E. Stake. *Educational Evaluation & Policy Analysis, 9*(4).

*Stake, R. E. (1986). *Quieting reform: Social science and social action in an urban youth program*. Chicago: University of Illinois Press.

Stake, R., & Davis, R. (1999). Summary evaluation of reader focused writing for the Veterans Benefits Administration. *American Journal of Evaluation, 20*, 323-344.

Stufflebeam, D. (1974). Metaevaluation. *Occasional Paper* #3. Retrieved January 22, 2008, from the Western Michigan University Evaluation Center Web site: http://www.wmich.edu/evalctr/pubs/ops/ops03.pdf

Stufflebeam, D. L. (1978). Metaevaluation: An overview. *Evaluation & the Health Professions*, *2*(1), 17-43.

Stufflebeam, D. L. (1999a). Program evaluation metaevaluation checklist (based on the Program Evaluation Standards) [short version]. Retrieved November 1, 2008, from the Western Michigan University Web site: http://www.wmich.edu/evalctr/checklists/program_metaeval.pdf.

Stufflebeam, D. L. (1999b). Program evaluation metaevaluation checklist (based on the Program Evaluation Standards) [long version]. Retrieved November 1, 2008, from the Western Michigan University Web site: http://www.wmich.edu/evalctr/checklists/program_metaeval_10point.pdf

Stufflebeam, D. L. (2001a). The meta-evaluation imperative. *American Journal of Evaluation, 22*(2), 183-209.

Stufflebeam, D. L. (2001b). Evaluation models. *New Directions for Evaluation, 89*.

*Stufflebeam, D. L. & Wingate, L. (2002). Metaevaluation: Attestation of the Evaluation's adherence to professional standards for program evaluation. In D. L. Stufflebeam, A. Gullickson, & L. Wingate, *The Spirit of Consuelo: An Evaluation of Ke Aka Ho'ona* (Appendix D). Retrieved November 11, 2007, from the Western Michigan University Web site: http://www.wmich.edu/evalctr/pubs/consuelo/

U.S. General Accounting Office. (1992). *The evaluation synthesis* (GAO/PEMD-10-1.2). Washington, DC: Author.

U.S. Office of Management and Budget. (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. *Federal Register*, February, 22, 2002. Retrieved March 20, 2005, from http://www.whitehouse.gov/omb/fedreg/reproducible2.pdf

Weiss, C. H. *Evaluation* (2nd ed.). (1998). Upper Saddle River, NJ: Prentice-Hall.

*Whitmore, E., & Ray, M. L. (1989). Qualitative evaluation audits. *Evaluation Review*, *13*(1), 78-90.

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation*, *22*(3), 409-418.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97-110). New York: Russell Sage Foundation.

Appendix A: Metaevaluations Described by Criteria, Application, Metaevaluator Location, and Purpose

Metaevaluation	Evaluand	Criteria	Application	Location	Purpose
A. Burbules (2000)	CIRCE Evaluation of the Milwaukee Teacher Education Center Alternative Teacher Education Program	Emergent	Narrative review	External	S
B. Farrar & House (1987)	AIR evaluation of PUSH/Excel	Emergent	Narrative review	External	F/S
C. Finn Jr., et al. (1997)	New York City Public Schools Integrated Learning Systems Project Evaluation	PgES*	Checklist	External	F/S
D. Greene (1999)	CIRCE VBA Appeals Training Module (Phase II)	Tailored	Narrative review	External	S
E. Greene et al. (1988)	Summary of two audits "on the qualitative evaluation data collected in two small-scale local human service program evaluations"	Trustworthiness	Audit	External	F
F. Greene et al. (1992)	East Central AIDS Education & Training Center Evaluation	Tailored	Audit	External	F/S
G. Hartmann & Loizides (2001)	WMU's evaluation of the NSF Advanced Technological Education project [Web-based survey only]	PgES Dillman's (2000) guidelines for Internet surveys	Checklist	External	F/S
H. House (1987)	Evaluation of the Promotional Gates program	Emergent	Narrative review	External	S
I. House et al. (1978)	Abt evaluation of Follow Through	Emergent	Narrative review	External	S
J. Kemmis (1997)	CIRCE evaluation of VBA Reader-Focused Writing program	Emergent	Narrative review	External	F/S
K. Lynch et al. (2003)	Evaluation of the Interdisciplinary Rural Health Training Program	PgES	Semi-structured review	Internal	F/S
L. McKinley (1999)	WMU's Michigan Public School Academy Initiative Evaluation	PgES	Checklist	External	S
M. Migotsky & Stake (2001)	WMU's evaluation of the NSF Advanced Technological Education project [site visits & issue papers]	Tailored	Semi-structured review	External	S
N. Ray (1988)	Cooksy's evaluation of a welfare-to-work program [qualitative component only]	Trustworthiness	Audit	External	S
O. Smith (1999)	CIRCE VBA Appeals Training Module (Phase I)	Emergent	Narrative review	External	S
P. Stake (1986)	AIR Evaluation of Cities-in-Schools	Emergent	Narrative review	External	S
Q. Stufflebeam & Wingate (2002)	WMU evaluation of the Spirit of Consuelo	PgES	Checklist	Internal	S
R. Whitmore & Ray (1989)	Whitmore's youth empowerment evaluation of youth employment needs	Trustworthiness	Audit	External	S

*1981 version of the Joint Committee on Standards for Educational Evaluation.