The Metaevaluation Imperative

Daniel L. Stufflebeam

American Journal of Evaluation 2001; 22; 183
DOI: 10.1177/109821400102200204

The Metaevaluation Imperative

DANIEL L. STUFFLEBEAM

ABSTRACT

The evaluation field has advanced sufficiently in its methodology and public service that evaluators can and should subject their evaluations to systematic metaevaluation. Metaevaluation is the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation's utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses. Formative metaevaluations—employed in undertaking and conducting evaluations—assist evaluators to plan, conduct, improve, interpret, and report their evaluation studies. Summative metaevaluations—conducted following an evaluation—help audiences see an evaluation's strengths and weaknesses, and judge its merit and worth. Metaevaluations are in public, professional, and institutional interests to assure that evaluations provide sound findings and conclusions; that evaluation practices continue to improve; and that institutions administer efficient, effective evaluation systems. Professional evaluators are increasingly taking their metaevaluation responsibilities seriously but need additional tools and procedures to apply their standards and principles of good evaluation practice.

INTRODUCTION

Metaevaluation—the evaluation of evaluation—is a professional obligation of evaluators. Evaluation has emerged as an important profession as societal groups increasingly commission evaluators to examine and pass judgment on many and varied consumer programs, products, and services. Attaining and sustaining the status of professionalism requires one to subject her or his work to evaluation and use the findings to strengthen services. This dictum pertains as much to evaluators as to physicians, lawyers, engineers, and accountants. It means evaluators should assure that their evaluations are themselves evaluated. Moreover, metaevaluations are needed in all types of evaluation, including evaluations of programs, projects, products, systems, institutions, theories, models, students, and personnel.

Daniel L. Stufflebeam • Harold and Beulah McKee Professor of Education and Director, The Evaluation Center, Western Michigan University, Kalamazoo, MI 49008-5237; Tel: (616) 387-5895; Fax: (616) 387-5923; E-mail: daniel.stufflebeam@wmich.edu.

American Journal of Evaluation, Vol. 22, No. 2, 2001, pp. 183–209.
ISSN: 1098-2140

This article's purposes are to emphasize the importance of metaevaluation, stimulate and contribute to American Evaluation Association (AEA) members' dialogue on and development of metaevaluation, and offer both conceptual and practical guidance for conducting metaevaluations. Particularly, the article (1) presents a case for metaevaluation, including both formative and summative metaevaluation; (2) suggests an overall conceptualization of metaevaluation, including reference to the available standards and guiding principles for judging evaluations; (3) references a range of metaevaluations to outline a methodology for planning and conducting metaevaluations; (4) cites a repository of checklists designed for use in planning, monitoring, and judging evaluations; and (5) discusses the role of context and resource constraints in deciding whether and, if so, how to do a metaevaluation.

The ensuing analysis is based on my longstanding commitment to help professionalize evaluation, substantial experience in leading the development of professional standards for evaluations, considerable experience in designing and conducting metaevaluations, review of pertinent literature, and personal efforts to generate useful metaevaluation tools. While I can contribute only a limited perspective and set of experiences, I hope this article will stimulate and help inform the dialogue and development work needed to advance the theory and practice of metaevaluation and also provide some tangible assistance.

A CASE FOR METAEVALUATION

In general, assuring that evaluations are rigorously evaluated is in professional, public, institutional, and personal interests. As professionals, evaluators need metaevaluations to assure the quality of their evaluations, provide direction for improving individual studies as well as their developing evaluation approaches, and earn and maintain credibility for their services among both clients and other evaluators. Consumers need metaevaluations to help avoid accepting invalid evaluative conclusions and, instead, to use sound evaluation information with confidence. Those who are not necessarily professional evaluators but who house and oversee evaluation systems need metaevaluations to help assure that their institution's evaluation services are defensible, functional, and worth the investment. Also, the subjects of evaluations—particularly personnel evaluations—have a right to expect that the systems used to evaluate their competence and performance have measured up to appropriate standards for sound personnel evaluations.

As with other societal endeavors, evaluations can be good, bad, or somewhere in between. Many things can and do go wrong in evaluations. Evaluations might be flawed by inadequate focus, inappropriate criteria, technical errors, excessive costs, abuse of authority, shoddy implementation, tardy reports, biased findings, ambiguous findings, unjustified conclusions, inadequate or wrong interpretation to users, unwarranted recommendations, and counterproductive interference in the programs being evaluated. These and other problems are seen in evaluations of programs, products, personnel, students, institutions, and other evaluands and across a wide array of disciplines and service areas. If such problems are not detected and addressed in the evaluation process, evaluators will present erroneous findings and may deliver services that are overly expensive, ineffective, and/or unfair. If flawed reports are issued without being exposed by sound metaevaluations, evaluation audiences may make bad decisions based on the erroneous findings. In the evaluation vernacular, formative metaevaluations are needed to plan and carry out sound evaluations, and summa-

tive metaevaluations are needed to judge completed evaluations. As Scriven (1994) pointed out, even the highly respected and widely used *Consumer Reports* magazine should be independently evaluated to help readers see the limitations as well as the strengths of the many product evaluations published in this magazine. Invalid personnel evaluations can have bad consequences for employees or conversely for an institution and its clients. Similarly, evaluations that accredit unworthy programs or institutions are a disservice to potential students or other constituents. Additionally, professional standards and principles for evaluations will be little more than rhetoric if they are not applied in the process of judging and improving evaluation services.

A CONCEPTUALIZATION OF METAEVALUATION

Michael Scriven introduced the term *metaevaluation* in 1969 in the *Educational Products Report.* He used this label to refer to his evaluation of a plan for evaluating educational products. Essentially, Dr. Scriven defined a metaevaluation as any evaluation of an evaluation, evaluation system, or evaluation device. He argued that issuance of inaccurate and/or biased reports could seriously mislead consumers to purchase unworthy or inferior educational products and then use them to the detriment of children and youth. Thus, he stressed that the evaluations of such products must themselves be evaluated and that such metaevaluations are critically important to the welfare of consumers.

An Operational Definition

Operationally, metaevaluation is defined in this article as *the process of delineating, obtaining, and applying descriptive information and judgmental information—about the utility, feasibility, propriety, and accuracy of an evaluation and its systematic nature, competent conduct, integrity/honesty, respectfulness, and social responsibility—to guide the evaluation and/or report its strengths and weaknesses*. This definition is designed particularly to serve the needs of AEA members and other evaluators who may want to meet AEA's Guiding Principles for Evaluators and also adhere to the standards issued by the Joint Committee on Standards for Educational Evaluation. It is instructive to consider this definition's main elements.

The *process* elements note that metaevaluation includes both group process and more isolated technical tasks. The group process tasks of delineating and applying denote the metaevaluator's interactions with stakeholders of the evaluation being assessed. In planning the metaevaluation, the evaluator identifies and communicates with the client and other audiences for the metaevaluation to reach mutual understandings on the important metaevaluation questions, how they are to be addressed, and how and when the findings are to be reported. In the metaevaluation's concluding stages, the metaevaluator meets or otherwise communicates with the client and other stakeholders to help them understand, correctly interpret, and apply the metaevaluation findings. In discussing the use of findings, the metaevaluator should help the audience avoid overgeneralizing as well as draw justified conclusions.

The *obtaining* elements, in the definition of metaevaluation, are the technical tasks required to acquire and assess the information needed to judge the target evaluation. Especially involved are the collection and assessment of evaluation contracts, plans, instru-

ments, data, and reports and evaluator credentials. Additionally, the metaevaluator(s) may interview, survey, and otherwise collect information and perspectives from persons involved in or affected by the evaluation process. Furthermore, as the definition notes, both *descriptive and judgmental information* is needed.

The definition's basis for judging program evaluations are *The Program Evaluation Standards* (Joint Committee, 1994) and *The AEA Guiding Principles* (American Evaluation Association, 1995). *The Personnel Evaluation Standards* (Joint Committee, 1988) are advocated for use in judging personnel evaluations and personnel evaluation systems. Such other standards and principles that may be appropriate in particular metaevaluations are also advocated. To its credit, *The American Journal of Evaluation* (AJE) recently began including a section for reporting metaevaluations and suggested that contributors reference the AEA *Guiding Principles* and the Joint Committee's standards in their submitted metaevaluation reports. Sanders (1995) has shown that the Joint Committee standards and the AEA *Guiding Principles* are compatible and complementary. Grasso (1999) merged and jointly applied both sets of requirements. In a later section of this article, I will reference and discuss checklists grounded in both the Joint Committee's standards and the *Guiding Principles*.

The Joint Committee *Standards* are focused on evaluations of education and training programs and educational personnel; require that evaluations be useful, feasible, proper, and accurate; and are considerably more detailed than the *AEA Guiding Principles*. However, the *Guiding Principles* apply to the full range of program evaluations, not just those associated with education and training. The *Guiding Principles* require that evaluations be systematic and data based, conducted by evaluators with the requisite competence, and honest; embody respect for all participating and affected persons; and take into account the diversity of interests and values that may be related to the general and public welfare. Both the Joint Committee's standards and the *Guiding Principles* merit consideration as appropriate bases for judging evaluations conducted by AEA members and/or others who desire to employ the Joint Committee standards and the *Guiding Principles*.

A caveat is that the Joint Committee developed its program and personnel evaluation standards expressly for application in the United States and Canada and cautioned against uncritical use of these standards outside this context. In this regard, it should be remembered that the Joint Committee defined a standard as a principle commonly agreed to by the parties whose work will be evaluated against the standard. The Joint Committee argued that evaluators in other countries should carefully consider what standards are acceptable and functional within their cultures and should not uncritically adopt and apply standards developed in a different country. There are definite problems in transferring North American standards on human rights, freedom of information, and other matters covered by the Joint Committee *Standards* to cultures outside the U.S. and Canada (Beywl, 2000; Jang, 2000; Smith, Chircop, & Mukherjee, 2000; Taut, 2000; Widmer, Landert, & Bacmann, 2000). AEA seems to espouse the reasonable position that AEA members and other evaluators—wherever they may be conducting evaluations—should adhere at least to the *Guiding Principles* if they intend to claim consistency between their evaluations and what AEA recommends for conducting sound evaluations.

The final part of this article's operational definition of metaevaluation notes its basic purposes as *guiding the evaluation and/or reporting its strengths and weaknesses*. Like any other kind of evaluation, metaevaluations may have a formative role in helping an evaluation succeed and a summative role in helping interested parties judge the evaluation's merit and worth.

Metaevaluation vis-a-vis Meta-Analysis

Before leaving this conceptualization of metaevaluation, it may be helpful to contrast metaevaluation and meta-analysis. Whereas these terms refer to quite different concepts, they are often inappropriately equated. As stated above, a metaevaluation assesses the merit and worth of a given evaluation. On the other hand, a meta-analysis is a form of quantitative synthesis of studies that address a common research question. In program evaluation research contexts, this usually involves a treatment-control (or treatment A-treatment B) comparison. Across a selected set of similar studies, the investigator calculates and examines the magnitude and significance of effect sizes.

While metaevaluation and meta-analysis are different activities, metaevaluations have applications in meta-analysis studies. Metaevaluations are used first to evaluate and determine which candidate comparative studies qualify for inclusion in a defensible meta-analysis database. Also, a metaevaluation can and should be conducted to assess the merit and worth of the completed meta-analysis study. The meta-analysis technique is rarely applicable in a metaevaluation, since most evaluations do not involve multiple comparative studies in a particular program area.

A GENERAL METHODOLOGY FOR PLANNING AND CONDUCTING METAEVALUATIONS

Given the proceeding conceptualization of metaevaluation, this article now pursues the development of a general methodology for conducting metaevaluations. I have referenced a particular metaevaluation to identify main tasks in any metaevaluation. Subsequently, a range of metaevaluations are referenced to identify a variety of methods of use in conducting the metaevaluation tasks in different evaluation domains.

This approach is only a beginning. It helped identify key tasks in a metaevaluation and a few procedures and notions that merit consideration in carrying through each task. However, the presented procedures are far from a comprehensive set.

A Metaevaluation of the TFA Teacher Evaluation System

The metaevaluation selected for reference in identifying tasks in the metaevaluation process is a metaevaluation of the teacher evaluation system employed by TEACH FOR AMERICA (TFA). This organization's purpose is to recruit, train, and certify graduates of various baccalaureate programs for service as teachers in inner city schools. These teacher trainees had four-year degrees grounded in an Arts and Sciences discipline, but most had no university-based teacher education. TFA's role was to recruit able students desiring to serve inner city students; provide them a year of on-the-job, supervised teacher training in inner city schools; rigorously evaluate their performance and potential during and immediately following this probationary period; and subsequently recommend only satisfactory performers for certification as effective teachers.

William Mehrens, Jason Millman, and I served as the metaevaluation team that, in 1995, evaluated TFA's system for evaluating the probationary teachers. It was called the Performance Assessment System (PAS). Our metaevaluation role was to determine whether the PAS—in design and execution—fairly, reliably, and accurately evaluated beginning teachers.

Our metaevaluation function was important to TFA's sponsors, administrators, constituent school districts, and TFA trainees. Many inner city schools lacked, desperately needed, and sought to employ additional competent teachers. The certifying bodies—be they state education departments or city school districts—needed assurances that they would be making sound certification decisions. The teaching candidates themselves needed assurances that they would be functioning in a profession for which they were well suited and appropriately prepared. They also deserved to be credentialed or screened out based only on fair, valid, impartial assessments. Politically, TFA needed to demonstrate the program's quality and integrity, since it was an innovative program and posed a quite radical alternative to the traditional approach to training and certifying teachers. TFA's teaching candidates had received no university-based teacher education, and some members of the teacher education establishment charged that TFA's program was inferior to traditional college and university teacher education programs. If TFA could expect state government to approve, trainees to enroll, and schools to employ the graduates, it needed to achieve and maintain credibility for the soundness of its bold alternative to traditional teacher preparation and certification programs.

TFA's evaluation of the teacher trainees was a crucial task in awarding certification and getting only qualified teachers into the schools. TFA commissioned the metaevaluation to help assure that TFA would be basing its judgments and recommendations on sound evaluations of the beginning teachers and to show that the evaluation process was subjected to independent assessments.

The five main components of TFA's teacher evaluation system were a teacher-compiled portfolio, portfolio assessors, a system of training and calibration for the assessors, actual assessments of portfolios, and certification recommendations derived from the assessments. The evidence in each teacher's portfolio included students' work, videotaped teaching, teaching plans, the teacher's assessment devices, and an analysis of the students' academic growth. In addition, the portfolio included survey results from the teacher's principal, other supervisors, teacher colleagues, parents, and students.

Two assessors independently evaluated each portfolio according to pre-established rubrics and produced subscores for the specified certification criteria and a total score. A third assessor resolved any unacceptable discrepancies between the first two sets of ratings.

TFA's main metaevaluation questions asked whether each of the following was adequate: the performance assessment design, assessments of teachers' impacts on students' learning, assessors' selection and training, implementation of the portfolio review process, quantitative analysis of the assessors' ratings of probationary teachers, legal defensibility of the PAS, and implications for PAS's wider use. The metaevaluation also assessed the PAS against the requirements of the 21 Joint Committee (1988) *Personnel Evaluation Standards* to determine PAS's utility, feasibility, propriety, and accuracy.

We first obtained from TFA and studied documents related to the targeted metaevaluation questions and the 21 Joint Committee standards. Among others, these documents included the teacher trainees' academic records; the credentials of the assessors who would evaluate the evidence on each probationary teacher; and the TFA plan and associated recruitment, training, and assessment materials. We also observed and prepared field notes on the training of the assessors and examined the beginning teachers' portfolios. Subsequently, we observed the assessors' actual assessments of the teachers' materials and analyzed the ratings and resulting certification recommendations, especially for reliability of subscores and agreements on final recommendations. Throughout the process, we conducted both

telephone and face-to-face interviews with a range of participants. Examination of the obtained evidence was used to judge whether TFA's performance assessment system met, partially met, or failed to meet each of the 21 Joint Committee standards. We also referenced pertinent policies, statutes, and laws to assess the legal viability of TFA's assessment structure and process. Finally, we produced an executive summary, a full-length report, and a technical appendix for the completed metaevaluation. In accordance with the metaevaluation contract, these reports were delivered to TFA for its discretionary use.

The basic findings were that TFA's evaluation team did a creditable and legally viable job in conducting and reporting summative evaluations of the TFA probationary teachers. Also, TFA was judged to have performed professionally in informing its state department and school district clients about the evaluation findings and engaging them in making appropriate decisions based on the findings. On the other hand, the metaevaluation identified areas where TFA needed to improve the PAS, especially in providing less hurried training for the assessors, strengthening the assessments of the teacher trainees' effects on student learning, and better matching assessors and trainees on areas and grade levels taught.

The Tasks in the Metaevaluation Process

While the preceding example is abbreviated, it nevertheless points up main tasks that should at least be considered in planning a metaevaluation process. Basically, these are generic tasks. Because metaevaluation is only a special type of evaluation, it should not be surprising that the identified tasks apply to evaluations in general and not only to metaevaluations. Also, it should be noted that these tasks were derived from a particular metaevaluation case; not all the tasks would necessarily apply in all formative and summative metaevaluations, and it is likely that additional tasks will sometimes be needed. The following tasks, then, are suggested mainly as a heuristic for use in planning metaevaluations and also are used in this article to analyze and draw out methods and lessons from a range of different metaevaluation cases.

Up front, the metaevaluator should **clarify the client and appropriate audiences** for the metaevaluation reports. In this case, the client group included TFA's leaders and staff. The metaevaluation had many other stakeholders, including the participating state education departments, school districts, teachers, and the students in the involved school districts.

The client **commissioned qualified metaevaluators**. Beyond having previous similar experiences in evaluating teacher evaluation systems, Dr. Millman, Dr. Mehrens, and I were selected to address three specialized areas. Dr. Millman, a past president of the National Council on Measurement in Education, focused especially on technical measurement questions, a topic for which he was eminently qualified. Dr. Mehrens examined the PAS's legal viability, reflecting his extensive experience in assessing the legal viability of state teacher evaluation systems. I assessed the system against the 21 Joint Committee *Personnel Evaluation Standards*, whose development I had led. These considerations aside, clearly TFA did not select us for our gender and racial diversity, although these are often relevant considerations. Presumably, our team was selected both to provide expertise that the client group lacked and to provide an independent, credible perspective on TFA's evaluation process. However, evaluators will not always be able to or need to engage an independent metaevaluator. In some resource-poor evaluations and especially in formative evaluations, the evaluators might appropriately do much or all of the formative metaevaluation themselves. Such self-metaevaluation practice is better than conducting no metaevaluation at all.

Another early task in the metaevaluation process is to **clarify in writing the agreements needed to guide the metaevaluation and often to negotiate a formal metaevaluation contract**. Among the important agreements reached in the TFA case were clarification of the metaevaluation issues and questions, the standards to be used in judging the evaluation system, guaranteed access to the needed information, substance and timing of reports, designated authority to edit and release the metaevaluation reports, and provision of the required resources. Formal contracts aren't always required, especially in small, formative, internally conducted metaevaluations. In general, though, it is wise and can prove prudent and functional to clarify and record as much as feasible the basic agreements that will guide and govern the metaevaluation.

The next metaevaluation task was to **compile and analyze the available, relevant information**. The initial information collection process typically culminates in a desk review of relevant documents. Following this "stay-at-home" work, the metaevaluator often must **collect additionally needed information**. For example, this metaevaluation included telephone and on-site interviews, observations, and study of portfolios. To reach valid conclusions, metaevaluators need access to all the relevant available information and to be able to collect any further needed information. Basically, the metaevaluator should obtain the full range of information required to apply all the applicable standards and principles and address the metaevaluation questions.

Next the metaevaluator should **analyze the obtained information, determine conclusions, and (usually) write up the metaevaluation findings**. This metaevaluation presented tables showing both quantitative and qualitative analyses and provided judgments on TFA's adherence to each of the Joint Committee's 21 personnel evaluation standards. We then divided the writing assignments and produced draft sections for the report. After discussing these, one member compiled the entire report and submitted the semifinal draft to the client for review. After considering critiques from the client and other stakeholders, the **metaevaluation report was finalized** and given to the client. As with other tasks, the need to fully implement this one depends on the metaevaluation context. For example, if the metaevaluation is sensitive, large-scale, and summative, formal written reports of findings will be required along with supporting technical appendixes; however, in more formatively-oriented metaevaluations, findings may appropriately be conveyed via oral communications, e-mail messages, letters, discussion sessions, and so forth

Subsequently, the metaevaluation team stood ready to **help the client and other stakeholders interpret the findings**. This task can be crucially important for improving an evaluation system, helping the interested stakeholders use the metaevaluation findings appropriately and productively, and also helping to assure that findings are not misinterpreted and/or misused. Metaevaluation follow-up procedures won't always be desired by the client or feasible. Clearly, metaevaluators cannot be expected to wait upon the client following completion of the metaevaluation if there hasn't been a prior agreement and associated funding to make this feasible. It is recommended—especially in large, summatively oriented metaevaluations—that the metaevaluator and client carefully consider in the initial contracting process the possibility of engaging the metaevaluator to provide service following delivery of the final report. Often such follow-up service can contribute importantly to the metaevaluation's impact, but this is unlikely to occur if not planned and budgeted in advance.

The above analysis reveals that a metaevaluation (or for that matter evaluations of other stripes) may be divided into eleven main tasks. These are presented in Table 1 as a structure for identifying alternative metaevaluation procedures.

TABLE 1.
Structure for Identifying Alternative Metaevaluation Procedures

1.	Determine and arrange to interact with the metaevaluation's stakeholders.
2.	Staff the metaevaluation team with one or more qualified evaluators.
3.	Define the metaevaluation questions.
4.	Agree on standards, principles, and/or criteria to judge the evaluation system or particular evaluation.
5.	Develop the memorandum of agreement or contract to govern the metaevaluation.
6.	Collect and review pertinent available information.
7.	Collect new information as needed, including, for example, on-site interviews, observations, and surveys.
8.	Analyze the qualitative and quantitative information.
9.	Judge the evaluation's adherence to appropriate standards, principles, and/or criteria.
10.	Convey the metaevaluation findings through reports, correspondence, oral presentations, etc.
11.	As needed and feasible, help the client and other stakeholders interpret and apply the findings.

SELECTED PROCEDURES FOR IMPLEMENTING METAEVALUATION TASKS

Using the preceding eleven metaevaluation tasks, we next look at some of the specific procedures that have proved useful in ten particular metaevaluations—two of personnel evaluation systems, five of program evaluations, one of a needs assessment system, one of alternative theoretical approaches to evaluation, and one of a large-scale student assessment system. One of the evaluations of personnel evaluation involved the system previously used by the U.S. Marine Corps to evaluate the performance of officers and enlisted personnel, while the other addressed the system the Hawaii Department of Education uses to evaluate Hawaii's public school teachers. The examples of evaluations of program evaluations involved the New York City School District's testing of the Waterford Integrated Learning System—a computer-based skills program for elementary school students (Finn, Stevens, Stufflebeam, & Walberg, 1997); an evaluations of programs of the Appalachia Regional Educational Laboratory; an evaluation of the Reader Focused Writing program for the Veterans Benefits Administration (Datta, 1999; Grasso, 1999; Stake & Davis, 1999); an evaluation of Australia's national distance baccalaureate program called Open Learning Australia; and a small-scale, modest formative metaevaluation of Michael Scriven's first goal-free evaluation (of an early childhood program in southern California). The needs assessment example stemmed from a metaevaluation conducted by the U.S. Army Command in Europe to assess needs assessments it was using in the mid-1980s to plan and offer courses to soldiers based in Europe. The metaevaluation of theoretical approaches was the present author's assessment of alternative program evaluation models (Stufflebeam, 2001). The metaevaluation of a large-scale assessment system focused on an attempt, by the National Assessment Governing Board, to set achievement levels on the National Assessment of Educational Progress (Stufflebeam, Jaeger, & Scriven, 1992; Vinovskis, 1999).

Space does not permit in-depth discussion of any of these cases. Instead, they are cited for particular arrangements and procedures that proved useful in terms of the eleven tasks identified above and in the settings where they were conducted. The intent is to help the reader consider arrangements and procedures that might aid the conduct of particular

metaevaluation tasks and, where feasible, to show optional ways of approaching different tasks. Clearly, context is important in determining when a particular procedure is or is not applicable and likely to be effective. Caveats and commentary are provided to help readers maintain circumspection as they consider whether, when, and how to use any of the cited arrangements and procedures.

Task 1: Determine and Arrange to Interact with the Evaluation's Stakeholders

A metaevaluation for the U.S. Marine Corps was instructive regarding the identification and involvement of stakeholders. From the beginning, this client group established two stakeholder panels and arranged for systematic interaction between them and the metaevaluators. The executive-level panel included 11 generals, 4 colonels, and a sergeant major. The advisory panel included about 20 representatives from different ranks of officers and enlisted personnel.

A Marine Corps management office scheduled monthly meetings between the metaevaluators and each panel. Each meeting was scheduled for at least two hours. The metaevaluators were contractually required to deliver printed reports at least ten working days in advance of the meeting, and the panelists were expected to read and prepare to discuss the reports. Collectively, these reports spanned all major tasks in the metaevaluation, including selection of standards for judging the Corps' personnel evaluation system; plans and instruments for obtaining information; diagnoses of strengths and weaknesses in the current personnel evaluation system; assessments of alternative personnel evaluation systems used in business, industry, and six other military organizations; generation and evaluation of three alternative new evaluation plans; and a plan for operationalizing and testing the selected new personnel evaluation system. A general officer chaired each meeting for both groups.

Each meeting began with a briefing by the metaevaluators using an overhead projector, with copies of the transparencies distributed to all persons present. A period of questions, answers, and discussion followed. At the meeting's end the presiding general officer asked each panelist to address a bottom line question. The lead general then summarized the meeting's main outcomes. Subsequently, an assigned officer prepared and distributed a report of all conclusions reached at the meeting. These meetings were highly substantive and productive, with one lasting more than five hours without a break. I can say without hesitation that these Marine Corps panels were the most professionally responsible, substantively involved, and ultimately better-informed evaluation clients I have ever served. They read, understood, critiqued, and discussed the metaevaluation reports and used them in decision making.

A down side was that the stakeholder panels were top heavy with high-ranking officers. This was an especially serious limitation, considering that they had all been promoted by the personnel evaluation system under review. Also, all members of the panels worked in the D.C. area, not, for example, in California, Hawaii, or Saipan. There was a risk that voices and concerns of rank and file members throughout the Corps would not be sufficiently represented and heard. We had to strive mightily to convince the Washington-area generals that we needed to work beyond the headquarters area to learn, report, and address the full range of personnel evaluation issues. With this accomplished through surveys and site visits, the stakeholder involvement aspect of this metaevaluation was good. The structure involved in this project for the Marine Corps could be beneficially applied in metaevaluations set in school districts, foundations, businesses, and other nonmilitary settings.

However, the stakeholder involvement procedures used in this example were largely dictated by the culture of the Marine Corps and the fact that the Commandant had mandated the metaevaluation and reform of the Corps' performance review system. In other institutions, where there is a less top-down organization and a style of informal exchange, a different, less formal process of identifying and interacting with stakeholders would be needed. Also, metaevaluators should keep in mind that some important metaevaluation stakeholders may be identifiable only as the metaevaluation unfolds. In such cases, the metaevaluator and client should consider keeping open the question of who should be involved and informed throughout the study process. Clearly, to pursue the most effective process of interaction, the metaevaluator should carefully study and take into account the metaevaluation's context and the client organization's culture and preferred style of communication and involvement.

As another caveat, not all metaevaluations need heavy involvement of stakeholders. In the formative metaevaluation of Michael Scriven's first goal-free evaluation, the purpose was to provide him a modest level of assessment he could use to detect and correct deficiencies in his evaluation plan. My reading of his plan and reaction by telephone and letters seemingly sufficed to quickly provide the needed formative metaevaluation feedback at a low cost. Also, there was minimal involvement of stakeholders in my metaevaluation of alternative theoretical approaches to evaluation. I engaged the authors of a number of the approaches evaluated to react critically to my characterizations and assessments of their approaches, also obtained critiques of the draft manuscript from colleagues, and had extensive exchange with Dr. Gary Henry who co-edits the *New Directions for Program Evaluation* where the metaevaluation report, titled *Evaluation Models,* appeared. While some metaevaluations will require extensive, more or less formal interactions with stakeholders, others will require little if any interaction with either a narrow or wide range of stakeholders. The metaevaluator needs to carefully consider the study's setting and should exercise judgment in deciding how best to involve stakeholders.

Task 2: Staff the Metaevaluation with One or More Qualified Metaevaluators

Turning to the second metaevaluation task, the metaevaluation team should be credible. The members should be competent and trusted by the stakeholders. In setting up the metaevaluation team for the Marine Corps, it was important to include persons with military personnel evaluation experience, as well as expertise in the different aspects of a metaevaluation. The metaevaluation for the New York City School District's basic education program included the perspectives of educational research, program evaluation, educational policy, and school district operations, as well as the perspectives of women and minorities. This team also could have used at least another perspective representing school and classroom-level operations and possibly others. Generally, the metaevaluation team's leader should clarify the metaevaluation work to be done and involve the client and stakeholders in appointing a qualified metaevaluation team.

There will be cases in which the client can afford to employ only a single metaevaluator. Then, one should engage the most credible, capable metaevaluator one can find. For example, we engage Dr. William Wiersma to conduct The Evaluation Center's metaevaluations for the Appalachia Regional Educational Laboratory. He meets this need exceptionally well. He is a highly accomplished research methodologist, with extensive successful experience in schools. He is thoroughly familiar with professional standards for evaluation and measure-

ment. He writes easily and well, as reflected in his widely circulated, readable research methodology textbooks and well-received evaluation reports. He understands education at all levels and relates effectively to educators, students, parents, and policymakers. Dr. Wiersma's qualifications give an indication of the characteristics one should seek for a "lone ranger" metaevaluator assignment. The published metaevaluations by Datta (1999) and Grasso (1999) of the Stake and Davis (1999) evaluation of the Reader Focused Writing program also illustrate the engagement of single, credible metaevaluators.

Even in the face of restricted resources and a formative metaevaluation orientation, the evaluator can sometimes obtain metaevaluation service from a colleague at little or no cost. For example, my involvement with Michael Scriven's goal-free evaluation involved a fee of only $100. Upon reviewing his initial plan, I judged that Dr. Scriven would miss program effects occurring on the school's playground and elsewhere outside the school. Following a revised plan, the main effects he found happened not in classrooms but on the playground.

Sometimes the evaluator cannot or need not engage even a single independent metaevaluator—especially when the target evaluation is internal, small scale, and informal. Even then, the evaluator can usefully self-assess evaluation plans, operations, and reports against pertinent professional principles and standards.

Task 3: Define the Metaevaluation Questions

The fundamental considerations in selecting questions for a metaevaluation are to assess the evaluation for (1) how well it meets the requirements of a sound evaluation (merit) and (2) the extent to which it meets the audience's needs for evaluative information (worth). Fundamentally, the metaevaluator should assess the extent to which the evaluation conforms to the requirements of a sound evaluation, for example, the AEA *Guiding Principles for Evaluators* and the Joint Committee *Program Evaluation Standards.* Also, the metaevaluator should address the client group's particular questions.

Illustrating the latter point, the National Assessment Governing Board (NAGB), in concert with its contracted metaevaluators, defined more than 20 questions concerning NAGB's attempt to set achievement levels of "basic, proficient, and advanced" on the National Assessment of Educational Progress.

Two examples illustrating clients' specific questions are as follows:

> 1.1 Is the membership of NAGB duly constituted, sufficiently representative of the National Assessment's constituencies, and effectively in touch with stakeholders so that it enjoys sufficient authority and credibility to set and secure use of achievement levels on the National Assessment?
> 2.1 Are NAGB's policy framework and specifications for setting achievement levels sufficiently clear and consistent with the state of the relevant measurement technology to assure that an appropriately representative group of standards setters can consistently and effectively set sound achievement levels on the National Assessment?

In general, the metaevaluator should carefully assure that the metaevaluation will determine the quality and overall value of the target evaluation and also address the audience's most important questions. Because some important metaevaluation questions may not be immediately clear at the metaevaluation's outset, the metaevaluator and client should consider the desirability of keeping open the possibility of identifying additional questions as the metaevaluation unfolds. Balance between the evaluation's initial structure and openness

to consider emergent questions is desirable. This helps assure that the metaevaluation begins with and maintains integrity keyed to the evaluation's initial purpose and questions and yet maintains flexibility to entertain and address new, important questions.

Task 4: As Appropriate, Agree on Standards, Principles, and/or Criteria to Judge the Evaluation System or Particular Evaluation

Evaluation is a professional activity. As such, it is often appropriate and helpful to judge evaluations against the professional standards and principles of the evaluation field. Harmonious conduct and potential impact of the evaluation are enhanced when metaevaluators and their clients reach a clear, up-front understanding of the criteria, principles, and/or standards to be applied in evaluating the target evaluation. Depending on particular situations, evaluators and clients may usefully choose from among a range of published standards and principles pertaining to evaluation. Some examples follow.

The APA (1999) *Standards for Educational and Psychological Testing* are especially useful for assessing testing programs (e.g., NAGB's attempt to set achievement levels on the National Assessment of Educational Progress) and particular assessment devices. Such applications of the "APA Test Standards" to metaevaluate measurement devices are seen in the various volumes of the Buros *Mental Measurements Yearbooks*. Other potentially useful standards include the NCES (1992) *Statistical Standards,* the NCES (1991) *SEDCAR Standards* for conducting large-scale surveys, the GAO (1994) *Government Auditing Standards,* and the American Evaluation Association's *Guiding Principles for Evaluators* (American Evaluation Association, 1995).

The standards most used in my metaevaluations are the program and personnel evaluation standards issued by the North American Joint Committee on Standards for Educational Evaluation (Joint Committee, 1988, 1994). They have been widely applied in American educational evaluations. For example, the Hawaii State Board of Education adopted the Joint Committee's program and personnel evaluation standards as state policy, stipulating that these standards be used to assess and strengthen Hawaii's system of educational accountability. While the Joint Committee's standards were developed for use in evaluating North American educational evaluations, certain groups have found them appropriate and useful in other areas. For example, with minor modifications, the U.S. Marine Corps adopted the Joint Committee's *Personnel Evaluation Standards* for use in assessing and reforming its system for evaluating officers and enlisted personnel. Similarly, General Motors (Orris, 1989) used the Joint Committee's *Personnel Evaluation Standards* to assess GM's system for evaluating executives. Also, the U.S. Army applied the Joint Committee's (1981) *Standards for Evaluations of Educational Programs, Projects, and Materials* to evaluate needs assessments conducted to help determine what courses the Army should provide to U.S. soldiers stationed in Europe.

In metaevaluations of the Stake and Davis (1999) evaluation, Datta (1999) employed the AEA *Guiding Principles* and Grasso (1999) mainly applied the *Guiding Principles* but also referenced the Joint Committee's (1994) *Program Evaluation Standards.*

It is acknowledged that metaevaluators and client evaluators need not always reach an advanced agreement on an explicit set of criteria, standards, or principles for judging an evaluation. Such formal negotiation of the bases for judging an evaluation tend to be unnecessary to the extent that only the evaluator needs the feedback, the orientation is formative rather than summative, the target is a draft evaluation plan or a particular issue in

the evaluation, and the need for feedback is immediate. An example of this occurred when Michael Scriven asked me to evaluate his plan for employing goal-free evaluation to evaluate the early childhood education program referenced previously in this article. Presumably, he was seeking my assessment based on my experience, general view of sound evaluation, and independent perspective. Metaevaluators should formally invoke pertinent evaluation criteria, standards, and principles when the evaluation would be strengthened thereby and when doing so is feasible. This will usually be the case in summative metaevaluations and often so in formative metaevaluations of fairly broad scope and large size.

Experience with and research on metaevaluation is too limited to yield definitive advice on weighting different standards for judging evaluations. In general, metaevaluators are advised to begin by assuming that all the involved standards should be accorded equal importance. Following deliberation with stakeholders and careful thinking about a particular metaevaluation, one should subsequently, if appropriate, differentially weight the standards. Sometimes it will be clear that some standards are not applicable in the particular metaevaluation. For example, the U.S. Army command in Europe decided that all the Joint Committee's accuracy, feasibility, and utility standards were highly applicable for judging the target needs assessment system but that there was no need to invoke the propriety standards. If more were known about this case, one might justifiably disagree with the Army command's decision to exclude the propriety standards. However, it is nevertheless true that metaevaluators have to make choices about assigning relative importance to different standards. In doing so, a metaevaluator should exercise careful judgments and document the basis for the judgments. In general, fewer standards will be applicable and important to the extent that the target evaluation is small and formative and directed only to the evaluator or a small audience. Large scale, summative evaluations employing the Joint Committee's standards often require that all the standards be applied. In such situations, one might justifiably decide that certain standards are so important that a failing grade on any of them will cause the evaluation to fail, even though high marks were attained on the other standards. In the case involving evaluation of alternative evaluation models, I decided that no model would be given a passing grade if it failed any of the following standards: Service Orientation, Valid Information, Justified Conclusions, and Impartial Reporting. My rationale was that a model would be an unacceptable guide to evaluations if it did not focus on beneficiaries' needs, answer the questions to which it was addressed, present defensible conclusions, and issue findings that were independent of the biases of the evaluation participants.

Task 5: Issue a Memo of Understanding or Negotiate a Formal Metaevaluation Contract

As with any evaluation, a metaevaluation should be firmly grounded in a sound memorandum of agreement or formal contract. According to the Joint Committee on Standards for Educational Evaluation (Joint Committee, 1994), evaluators and their clients should negotiate and document evaluation agreements that contain ". . .mutual understandings of the specified expectations and responsibilities of both the client and the evaluator." Such an agreement clarifies understandings and helps prevent misunderstandings between the client and metaevaluators and provides a basis for resolving any future disputes about the evaluation. Without such agreements the metaevaluation process is constantly subject to misunderstandings, disputes, efforts to compromise the findings, attacks, and/or the client's withdrawal of cooperation and funds. As the Committee further states, "Having entered into

such an agreement, both parties have an obligation to carry it out in a forthright manner or to renegotiate it. Neither party is obligated to honor decisions made unilaterally by the other." Written agreements for metaevaluations should be explicit but should also allow for appropriate, mutually agreeable adjustments during the metaevaluation. A checklist designed to help metaevaluators and clients identify key contractual issues and make and record their agreements for conducting a metaevaluation may be referenced and downloaded at the Web site to be referenced near the end of this article. The referenced checklist is designed to help metaevaluators and their clients launch, stand by and, as appropriate, modify the agreements required to guide and govern a metaevaluation.

In the metaevaluation of an evaluation of the Waterford Integrated Learning Systems project in the New York City School District (Miller, 1997), the metaevaluation team (Finn et al., 1997) was contracted not by the primary evaluators or the program directors, but by an independent foundation. This helped the metaevaluators maintain their independence and issue sometimes unwelcome judgments without fear of having their contract canceled. Many metaevaluations have the potential for conflict; when feasible, obtaining a contract and funds from a third party strengthens the metaevaluation's contractual grounding and viability.

An example of the hazards of proceeding without clear, advance written agreements is seen in the metaevaluation for NAGB (Stufflebeam, 2000). Unfortunately, I had acquiesced to the client organization's urgent request for a summative evaluation before it could formally process a contract. When NAGB subsequently was offended by the draft report and refused to pay for the summative evaluation work, there was no written agreement with which to press the issue. My university never received payment for the summative metaevaluation work that had been agreed to verbally but not in a formal written contract.

Task 6: Collect and Review Pertinent Available Information

After agreeing on the terms to govern the evaluation, the metaevaluator needs to examine the target evaluation against pertinent evidence. Initially, this involves collecting and assessing existing information. In some metaevaluations, this is the only information used to reach the metaevaluation conclusions. Legitimate reasons for collecting additional information are that the existing information is technically inadequate, insufficient, and/or not sufficiently credible to answer the metaevaluation questions. When the existing information is fully acceptable for producing a sound metaevaluation report, further data collection can be wasteful.

Datta (1999) and Grasso (1999) referenced both the Stake and Davis (1999) published summary of their evaluation of the Reader Focused Writing program and their full-length report, which was available on a University of Illinois web site. In conducting a defensible metaevaluation, it is important to look beyond an executive summary and even the full-length report. A key lesson for evaluators seen in the Datta and Grasso metaevaluations is that evaluators and their clients can facilitate the conduct of metaevaluations by placing evaluation reports and supporting materials on a Web site. Evaluators and their clients are advised to consider this and make provisions for such access to reports when negotiating the evaluation contract.

My metaevaluation of the evaluation of the Open Learning Australia distance education program is instructive about the kinds of extant information from which to begin a metaevaluation and how to handle that information. It was in the interest of Open Learning Australia to control the metaevaluation's costs, since travel from the U.S. to Australia entails

a sizable expense. Thus, it was agreed that Open Learning Australia would send pertinent information to Kalamazoo, where I could reference it in reaching at least tentative judgments about the adequacy of the evaluation. A wide array of documents was involved. These included letters, plans, budgets, contracts, data collection forms, journal and newspaper articles, field notes, reports, and responses to reports. Substantive foci were the nature of Open Learning Australia, the background of the evaluation, the evaluation plans and procedures, the evaluation process, the findings, publicity for the program, and guidelines for the metaevaluation.

The client for the metaevaluation emphasized that all judgments of the evaluation should be grounded in references to the pertinent evidence. This, they thought, would distill any stakeholder notions that the metaevaluation from afar was only a set of ill-informed opinions. Accordingly, I catalogued every piece of information used in the metaevaluation, giving its year of origination and a unique number within that year. In reporting a judgment for each of the 30 Joint Committee standards, I referenced each catalogued information item used in reaching the judgment. Thus, the client group and its constituents could review essentially all the evidence I used to reach the metaevaluation conclusions. In other metaevaluations, I have found this documentation procedure to be useful, not only for bolstering the evaluation report's credibility, but also for maintaining a quite definitive history of the metaevaluation that facilitates revisiting and studying the metaevaluation in later years.

Task 7: Collect New Information as Needed, Including, for Example, On-Site Interviews, Observations, and Surveys

While the extant information for evaluating the evaluation of Open Learning Australia (OLA) was substantial, it was insufficient to produce the needed metaevaluation report. Thus, I went to Australia to fill in some important information gaps. In addition to talking with OLA's leaders and participating faculty, I also met with OLA's students and with leaders and faculty in the more traditional higher education programs. This additional input led to the determinations that the assumed need for Open Learning Australia was questionable, the quality of OLA offerings was highly variable and, based on my analysis, these findings were at variance with the evaluation of OLA. In retrospect, the additional information obtained by making a site visit to Australia was vital to the validity of the metaevaluation report.

Another example of supplementing extant information with new information to reach metaevaluation conclusions occurred in an assessment of Hawaii's teacher evaluation system. Jerry Horn of our Center first used extant information to judge the Hawaii system against each of the 21 *Personnel Evaluation Standards* (Joint Committee, 1988). He then supplemented this information with surveys of stratified random samples of Hawaii's public school teachers and administrators. The survey items were keyed to the 21 personnel evaluation standards. The additional information not only corroborated the initial judgments but also provided an even stronger case that the existing teacher evaluation system was badly in need of reform.

Task 8: Analyze the Findings

The wide array of information used in various metaevaluations requires a variety of qualitative and quantitative analysis procedures. In The Evaluation Center's metaevaluations, we have used line and bar graphs, pie charts, reanalysis of data from the target evaluation,

and computer-assisted content analysis, among other techniques. In her metaevaluation of Stake and Davis (1999) evaluation of the Reader Focused Writing program, Datta (1999, p. 350) employed a cross-break table to contrast the topics addressed in each of five case study reports. Based on this analysis, she observed, "Because there seemed to be only a few common elements reported on in each site. . . the reliability in areas such as productivity seem uncertain. Sorting out idiosyncratic findings from incomplete inquiry is a bit difficult."

In a reanalysis of cost and effectiveness data for several alternative reading improvement programs, I arrived at conclusions that strongly contradicted the findings of the primary evaluation. The key issue was what number of students should be included in the denominator used to determine per-pupil costs. One program that concentrates reading improvement resources on only those most needy students and does so until their reading proficiency is satisfactory and sustainable can actually be less expensive and more cost-effective for a school as a whole than a program that year after year spends a modest amount for reading improvement on every student in the school, whether or not all of them need the remedial service. Computing the per-pupil expenditures based only on those students served can produce misleading results. Such erroneous analysis can lead a school to choose a program that appears to incur low per-pupil costs but actually is overly expensive for the school as a whole and not cost-effective. It seems much more plausible that a school that concentrates its resources for reading improvement on poor readers and does so until the needed improvements occur can actually spend less on the reading problem and do so more successfully than a school that year after year spends a relatively puny amount, but on each and every student.

Task 9: Judge the Evaluation's Adherence to the Selected Evaluation Standards, Principles, and/or Other Criteria

Following analysis and display of the obtained information, the evaluator must judge the target evaluation. Particularly important is the approach to analyzing judgments keyed to the employed standards. Datta (1999) and Grasso (1999) basically keyed their narrative assessments of the Stake and Davis (1999) evaluation to an outline of the main standards in the AEA *Guiding Principles* and subparts for each one.

My metaevaluations typically have been keyed to each of the 21 standards in the Joint Committee's (1988) *Personnel Evaluation Standards* or the 30 standards in the Joint Committee's (1994) *Program Evaluation Standards* and, more specifically, to between 6 and 10 specific points associated with each standard. To support narrative judgments I usually score the target evaluation on all points for each standard and then assign a scaled value meaning (e.g., excellent, very good, good, poor) to the evaluation's adherence to each standard. Sometimes I have subsequently followed a set procedure to aggregate the scores across standards and produce judgments of the evaluation on each of the main requirements of utility, feasibility, propriety, and accuracy.

The following charts illustrate how colleagues and I developed and presented judgments of the Marine Corps' personnel evaluation system. The rubrics in Table 2 were used to determine the degree to which the personnel evaluation system had satisfied standards in the four categories—utility, feasibility, propriety, and accuracy. All available relevant evidence was then used to identify the personnel evaluation system's strengths and weaknesses related to each standard. Using these lists, judgments were formed about whether the system met, partially met, or failed to meet each standard. Then, to summarize the results, the rubrics from Table 2 were used to prepare the summary matrix seen in Table 3. Based heavily on this

TABLE 2.
Rubrics Used to Determine Whether a Military Branch's Personnel Evaluation System Satisfies the Conditions of Utility, Feasibility, Propriety, and Accuracy[1]

Categories of standards	*Degree of Fulfillment of Requirements*[2]		
	Not met	*Partially met*	*Met*
Utility	1. Three or more standards are not met.	2. At least 3 of the 5 standards are met or partially met, and at least 1 standard is not met. or 3. Fewer than 4 standards are met, all 5 are either met or partially met, and no standard is unmet.	4. At least 4 or 5 standards are met, and none are unmet.
Feasibility	5. Three or 4 standards are not met.	6. At least 2 of the 4 standards are met or partially met, and at least 1 standard is not met. or 7. Fewer than 2 standards are met, and no standard is unmet.	8. At least 2 of the 4 standards are met, and none are unmet.
Propriety	9. Three or more standards are not met.	10. At least 3 of the 5 standards are met or partially met, and 1 or 2 standards are not met. or 11. Fewer than 4 standards are either met or partially met, and no standard is unmet.	12. At least 4 of the 5 standards are met, and none are unmet.
Accuracy	13. Four or more standards are not met.	14. At least 5 of the 8 standards are met or partially met, and at least 1 standard is not met. or 15. Fewer than 5 standards are met, at least 4 are either met or partially met, and no standard is unmet.	16. At least 5 of the 8 standards are met, and none are unmet.

[1]This form is designed for use in judging the overall utility, propriety, feasibility, and accuracy of an evaluation. Use of the decision rules on this form requires that the user first judge whether the evaluation meets, partially meets, or fails to meet the detailed rquirements of each of the 21 standards as they appear in Joint Committee on Standards for Educational Evaluation, *The Personnel Evaluation Standards* (Sage, 1988) and an additional standard (Transition to the New Evaluation system) developed for this project.

[2]In some cases, a standard appropriately may be judged as not applicable, and such standards would have no impact on determining which of the above rubrics fits the pattern of judgments.

TABLE 3.
Conclusions on the Degree to Which a Military Branch's Personnel Evaluation System Satisfies Standards of Utility, Propriety, Feasibility, and Accuracy

Category of standards	*Conclusion*	*Rubric (from Table 2, above)*
Utility	Not met	1
Propriety	Partially met	10
Feasibility	Partially met	6
Accuracy	Not met	13

analysis, the U.S. Marine Corps decided to replace its personnel evaluation system with one that would better meet the standards.

Task 10: Prepare and Submit the Needed Reports

Throughout the metaevaluation, there are important occasions for preparing and submitting evaluation reports. Typical reports include an initial metaevaluation plan, interim reports keyed to the evaluation's important aspects, and the final report. For each report, it is usually advisable to prepare and submit a draft, follow this up with a feedback workshop designed to orally communicate and discuss the draft reports, and subsequently complete and submit the finalized version of the particular report. The core contents of the reports can be keyed to the guiding standards and principles for evaluations. The charts presented above illustrate ways to display standards-based judgments in the reports. For many reports, it is often appropriate to prepare an executive summary, a set of supportive transparencies and/or a Power Point® presentation, a full-length report, and a technical appendix. Depending on advance agreements, it may also be appropriate to post the metaevaluation reports on a web site, submit an executive summary for publication in a professional journal, or both.

Task 11: As Appropriate, Help the Client and Other Stakeholders Interpret and Apply the Findings

Throughout the metaevaluation process, it is desirable to have regular, periodic exchanges with representatives of the key audiences. Our evaluation for the Marine Corps is illustrative of an extensive, functional reporting process. As noted previously, two client panels were established at the metaevaluation's beginning, along with specifications and a schedule for the reports. The reporting schedule was closely linked to the client organization's schedule for making decisions about the retention or replacement of the subject personnel evaluation system.

Our experience with the Hawaii Department of Education was similar to the experience with the Marine Corps. At the metaevaluation's outset the department appointed a review panel that represented the various interests in the state's public education system. Included were the president of the state board of education, the majority leaders of the two houses of the state legislature, a representative of the military establishment, the president of the state teachers union, the state superintendent of public instruction, the chief executive officer of one of the state's largest companies, two school principals, other schoolteachers and staff members, the head of the Pacific Regional Educational Laboratory, and representatives of parents and the general public. The metaevaluator and members of Hawaii's Department of

Education regularly met with this group to discuss and obtain input for the ongoing metaevaluation of Hawaii's systems for evaluating students, teachers, administrators, and schools. The panel was charged to receive, critique, discuss, and help the department use the metaevaluation findings. The review panel helped clarify the metaevaluation questions, provided valuable critiques of draft reports, and used the findings to generate recommendations for improving the state education systems of public accountability. By being involved in the metaevaluation process, the review panel developed ownership of the findings and became a powerful, informed resource for helping to chart and obtain support for the needed reforms.

I think it was important that this group was termed a review and not an advisory panel. The orientation was that they were qualified to critique draft plans and reports from their perspectives but not necessarily to provide technical advice for improving the metaevaluation work. In my experience, groups sometimes become dysfunctional and counterproductive when they are accorded an aura of unmerited expertise by virtue of being labeled an advisory panel.

Parallel to the involvement of the review panel, the metaevaluators engaged panels of Hawaii educators in carrying out the metaevaluation. These stakeholders also made valuable inputs to the metaevaluation process and developed interest and confidence in the findings.

As seen in the preceding examples, metaevaluation can and often should be a collaborative effort. This is especially so when the aim is to help an organization assess and reform its evaluation systems. When the aim is to protect the public from being misinformed by evaluations of specific entities, the evaluator must maintain proper distance to assure an independent perspective. Even then, however, metaevaluators should communicate appropriately with audiences for the metaevaluation reports to secure their confidence, interest, assistance, understanding, and informed uses of findings.

Comparative Metaevaluations

Sometimes a metaevaluation involves a comparative assessment of a number of evaluations. For example, professional societies, such as AEA and the American Educational Research Association, do so when they rate evaluations as a basis for making awards to outstanding evaluations. Table 4 provides an example of how nine hypothetical candidate evaluations might be subjected to a comparative metaevaluation. The hypothetical evaluations are listed in order of merit. The ratings are in relationship to the Joint Committee (1994) *Program Evaluation Standards* and were derived by the author using a special checklist keyed to the *Standards*.

Assume that each evaluation was rated on each of the 30 Joint Committee program evaluation standards by judging whether the study meets each of 10 key features of the standard. Each approach was then judged as follows: 9–10 Excellent, 7–8 Very Good, 5–6 Good, 3–4 Fair, 0–2 Poor. The score for each evaluation on each of the 4 categories of standards (utility, feasibility, propriety, accuracy) was then determined by summing the following products: 4 × number of Excellent ratings, 3 × number of Very Good ratings, 2 × number of Good ratings, 1 × number of Fair ratings. Judgments of each evaluation's strength in satisfying each category of standards would then be determined according to percentages of possible quality points for the category of standards as follows: 93 to 100% Excellent, 68 to 92% Very Good, 50 to 67% Good, 25 to 49% Fair, 0 to 24% Poor. This is conducted by converting each category score to the percentage of the maximum score for the category and

TABLE 4.
Ratings of Candidate Program Evaluations

Evaluation approach	*Graph of overall merit* 0 P F G VG 100 E	*Overall score & rating*	*Utility rating*	*Feasibility rating*	*Propriety rating*	*Accuracy rating*
Evaluation 1		92 (VG)	90 (VG)	92 (VG)	88 (VG)	98 (E)
Evaluation 2		87 (VG)	96 (E)	92 (VG)	81 (VG)	79 (VG)
Evaluation 3		87 (VG)	93 (E)	92 (VG)	75 (VG)	88 (VG)
Evaluation 4		83 (VG)	96 (E)	92 (VG)	75 (VG)	69 (VG)
Evaluation 5		81 (VG)	81 (VG)	75 (VG)	91 (VG)	81 (VG)
Evaluation 6		80 (VG)	82 (VG)	67 (G)	88 (VG)	83 (VG)
Evaluation 7		80 (VG)	68 (VG)	83 (VG)	78 (VG)	92 (VG)
Evaluation 8		72 (VG)	71 (VG)	92 (VG)	69 (VG)	56 (VG)
Evaluation 9		60 (G)	71 (VG)	58 (G)	58 (G)	50 (G)

Within types, listed in order of compliance with The Program Evaluation standards.

multiplying by 100. The 4 equalized scores are next summed, divided by 4, and compared to the total maximum value of 100. The evaluation's overall merit is then judged as follows: 93 to 100 Excellent, 68 to 92 Very Good, 50 to 67 Good, 25 to 49 Fair, 0 to 24 Poor. Regardless of each evaluation's total score and overall rating, I would judge any evaluation as failed if it received a Poor rating on the vital standards of P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, A11 Impartial Reporting. The scale ranges in Table 4 are **P** = Poor, **F** = Fair, **G** = Good, **VG** = Very Good, **E** = Excellent.

It should be noted that the above procedure unequally weights different standards in the process of computing a total score and overall rating. This is because there are unequal numbers of standards in the four categories. An alternate means of determining a total score and overall rating is to sum and average the 30 individual standard scores. My practice is to compute, assess, and discuss the extent of agreement between the two total score/overall rating approaches.

CHECKLISTS FOR USE IN METAEVALUATION STUDIES

As illustrated above, checklists can be useful in metaevaluations. The Evaluation Contracting Checklist and the Joint Committee Program Evaluation Standards Checklist plus additional checklists may be accessed at the web site to be referenced near the end of this article. Included in that repository are a range of checklists designed for use in evaluating personnel, program, and materials evaluations. Among others they include Scriven's Key Evaluation Checklist, a checklist by Ernest House and Kenneth Howe for guiding and assessing Deliberative Democratic evaluations, Lorrie Shepard's Checklist for Assessing State Educational Assessment Systems, my personnel and program evaluation checklists keyed to the Joint Committee's evaluation standards, and another of mine focused on the tasks in the evaluation process. By the time this article is published, I expect to put up on the Web site and invite critical feedback on a draft AEA Guiding Principles Checklist. Interested readers may also consult my article on evaluation checklists that appeared in the last issue of *AJE* (Stufflebeam, 2001).

THE ROLE OF CONTEXT AND RESOURCE CONSTRAINTS IN DECIDING WHETHER AND, IF SO, HOW TO DO A METAEVALUATION

The preceding analysis and discussion of metaevaluation needs to be tempered by considerations of the reality constraints in evaluation work. It will not always be important and/or feasible to do a formal metaevaluation. The cases referenced in this article are all examples wherein a client requested and funded a metaevaluation. Even then, the cases varied considerably in the need for extensive, formal feedback and the amount of employed resources. The amounts of money invested generally were in the range of $10,000 to $30,000, but the smallest metaevaluation done on Michael Scriven's goal-free evaluation cost only $100, while the metaevaluation for the Marine Corps cost more than $450,000. Generally, the small, formative target evaluations required much less money and metaevaluation effort than did the large-scale, summative evaluations. Usually, cost should not be a deterrent to obtaining some level of metaevaluation. Typically, the size of budgets for metaevaluations are minuscule compared with the cost of the target evaluation, often less than 1% of the target evaluation's annual budget. Moreover, in large-scale, high stakes evaluations, metaevaluations can often be judged cost-free when their costs are compared with the value of the benefits they produce.

Nevertheless, sometimes evaluators will not request and/or need a formal metaevaluation. Examples are evaluation systems that have been subjected to a metaevaluation relatively recently and that subsequently have operated relatively free of complaints and observed problems. Also, individual personnel evaluations typically require no metaevaluations, except when one is triggered by an appeal of the findings. Many government agencies, accrediting organizations, and charitable foundations seek no metaevaluations of the evaluations they sponsor, presumably because they trust their system of monitoring and oversight (although such trust is not always justified). Also, very small-scale, formative evaluations—as when one evaluates a small project or a course for purposes of improvement—might not need or be amenable to any kind of formal metaevaluation.

Nevertheless, despite evaluations that require little or no metaevaluation, it is always appropriate for an evaluator to plan and carry through even small-scale formative evaluations with a metaevaluation mind-set. One of the best ways to do this is to thoroughly study and internalize the key messages of the Joint Committee's standards, the AEA *Guiding Principles*, and other standards such as the APA *Standards for Educational and Psychological Testing*. Having the underlying metaevaluation principles in mind is invaluable in planning evaluations, dealing with issues and problems as they arise, advising evaluation participants regarding the dilemmas they face, and—after the fact—taking stock of what the evaluation accomplished.

SUMMARY AND RECOMMENDATIONS

Metaevaluations serve all segments of society. They help assure the integrity and credibility of evaluations and are thus important to both the users and producers of evaluations. Metaevaluations are often needed to scrutinize evaluations of charitable services; research and development projects; equipment and technology; state assessment systems; new, expensive curricula; automobiles and refrigerators; hospitals and other organizations; and

engineering plans and projects. Metaevaluations are also needed to assess and help improve the systems used to evaluate physicians, military officers, researchers, evaluators, public administrators, teachers, school principals, students, and others. In cases of appeal, metaevaluations are needed to assess the soundness and fairness of evaluations of individual persons. As seen in these examples, metaevaluations are in public, professional, institutional, and personal interests.

Examples of the societal value of metaevaluations abound. Purchasers of consumer products need assurance that evaluations of alternatives are sound. Taxpayers, needing to intelligently decide on levels of support, require assurance that they are getting dependable information on the needs and performance of their town's service organizations and schools. Parents need to know whether they are getting reliable and valid evaluations of the schools and colleges where they might send their children and of their children's academic progress. In cases of high stakes evaluations, students and their parents are entitled to know whether the students' efforts and achievements have been fairly and validly graded. School board members and school administrators need assurances that they are getting relevant and technically defensible evaluations of schools, programs, and personnel. Politicians and their constituents need to know whether international, comparative evaluations of educational achievements in different countries are valid and defensible for use in reviewing and revising national education policy. In all of these examples, users of evaluations need sound metaevaluation information to help them assess the relevance, dependability, and fairness of evaluative information they are receiving.

As professionals, evaluators themselves need to regularly subject their evaluation services to internal and independent review. Sound metaevaluations provide evaluators with a quality assurance mechanism they can use to examine and strengthen evaluation plans, operations, draft reports, and means of communicating the findings. Also, the prospect and fact of metaevaluations should help keep evaluators on their toes, push them to produce defensible evaluation services, and guide them over time to improve their services.

Metaevaluation is as important to the evaluation field as auditing is to the accounting field. Society would be seriously at risk if it depended only on accountants for its financial information, without acquiring the scrutiny of independent auditors. Likewise, parents, students, educators, government leaders, business persons, and consumers, in general, are at risk to the extent they cannot trust evaluation findings.

Despite the strong case that can be made for metaevaluation, not all evaluations will require or merit a metaevaluation. Small-scale, locally focused, and improvement-oriented evaluations may not require any special metaevaluation. Making such determinations is a matter for careful judgment by the evaluator and client; they should take into account the local setting and especially the audience for the evaluation. The many evaluations that do merit a special metaevaluation will vary considerably concerning the type and level of needed effort. In general, evaluators should bring a strong metaevaluation orientation to their work, whatever the type and level of evaluation to be conducted. Small, improvement-oriented evaluations often should employ a modicum of formative metaevaluation. Large-scale, summative evaluations usually should secure at least a formal, summative metaevaluation and often a formative metaevaluation as well. In deciding whether or not to commission or conduct a metaevaluation, evaluators and their clients should keep in mind that a metaevaluation's cost is typically small compared with the cost of the target evaluation and that the value of the metaevaluation's benefits can far outweigh the metaevaluation's costs.

Metaevaluation is defined operationally in this article as *the process of delineating,*

obtaining, and applying descriptive information and judgmental information—about the utility, feasibility, propriety, and accuracy of an evaluation and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility—to guide the evaluation and publicly report its strengths and weaknesses. This definition is designed particularly to serve the needs of AEA members plus other evaluators who subscribe to AEA's position on the meaning of sound evaluation.

Based on this definition, a general, 11-task methodology for metaevaluation was suggested: (1) determine and arrange to interact with the metaevaluation's stakeholders; (2) staff the metaevaluation team with one or more qualified evaluators; (3) define the metaevaluation questions; (4) agree on standards, principles, and/or criteria to judge the evaluation system or particular evaluation; (5) develop the memorandum of agreement or contract to guide and govern the metaevaluation; (6) collect and review pertinent available information; (7) collect new information as needed, including, for example, on-site interviews, observations, and surveys; (8) analyze the qualitative and quantitative information; (9) judge the evaluation's adherence to appropriate standards, principles, and/or criteria; (10) convey the metaevaluation findings through reports, correspondence, oral presentations, and so forth; and (11) as needed and feasible, help the client and other stakeholders interpret and apply the metaevaluation findings.

Example metaevaluations were referenced to identify a sample of metaevaluation techniques of use in carrying out each metaevaluation task. The examples included evaluations of program and personnel evaluations, a needs assessment system, alternative evaluation models, assessment devices, and a state assessment system. Among the procedures presented for use in metaevaluations were review panels, a contracting checklist, a suggestion—when feasible—to contract a metaevaluation with a third party, a range of checklists for judging evaluations, rubrics and analysis protocols for judging evaluations, and feedback workshops. A range of checklists designed for use in metaevaluations can be accessed and downloaded at the following web site: <*www.wmich.edu/evalctr/checklists*>. Also included in this web site are a paper by Michael Scriven (2000) titled "The Logic and Methodology of Checklists" and one by me (Stufflebeam, 2000) titled "Guidelines for Developing Evaluation Checklists."

Undergirding this article is the strong recommendation that evaluators should ground their metaevaluations in professional standards and principles for evaluations. Those evaluators and clients wanting to meet AEA's expectations for sound evaluations are advised to employ both the AEA *Guiding Principles* and the professional standards for program and personnel evaluations promulgated by the Joint Committee on Standards for Educational Evaluation. Applying parallel, systematic methodologies and checklists for both the *Guiding Principles* and the Joint Committee's standards would be consistent with the virtue of applying multiple methods in evaluation work to foster rigor and reproducibility of judgments and to identify discrepancies that should be taken into account.

As seen in this article, evaluators are making progress in conducting metaevaluations. Sustaining and increasing efforts to systematize and increase the rigor, relevance, and contributions of metaevaluations are in the interest of professionalizing the evaluation field. AEA has strongly supported the metaevaluation imperative. Perhaps the association could do even more through such actions as stimulating and sponsoring research and development in the metaevaluation area and establishing a special interest group on metaevaluation.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

American Evaluation Association, Task Force on Guiding Principles for Evaluators. (1995). Guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *New directions for program evaluation* (pp. 19–26). San Francisco: Jossey-Bass.

American Psychological Association. (1973). *Ethical principles in the conduct of research with human participants*. Washington, D.C.: Author.

Beywl, W. (2000). Standards for evaluation: On the way to guiding principles in German evaluation. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 60–67). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5*, 437–474.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. J. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

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.

Evaluation Center. (1995). *The USMC personnel evaluation standards*. Kalamazoo, MI: Western Michigan University Evaluation Center.

Finn, C. E., Stevens, F. I., Stufflebeam, D. L., & Walberg, H. J. (1997). A meta-evaluation. In H. Miller (Ed.), The New York City public schools integrated learning systems project. *International Journal of Educational Research, 27*, 159–174.

General Accounting Office. (1994). *Government auditing standards*. Washington, D.C.: Author.

Glass, G. V. (1974). *Excellence: A paradox*. Speech presented at the second annual meeting of the Pacific Northwest Research and Evaluation Conference sponsored by the Washington Educational Research Association.

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.

Guba, E. G. (1969). The failure of educational evaluation. *Educational Technology, 9*, 29–38.

House, E. R. (1977). *Fair evaluation agreement*. Urbana-Champaign, IL: University of Illinois Center for Instructional Research and Curriculum Evaluation.

House, E. R., Glass, G. V, McLean, L. D., & Walker, D. F. (1977). *No simple answer: Critique of the "Follow Through Evaluation."* Urbana-Champaign, IL: Center for Instructional Research and Curriculum Evaluation.

Jang, S. (2000). The appropriateness of Joint Committee standards in non-western settings: A case study of South Korea. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 41–59). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.

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

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Newbury Park, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*, 2nd ed. Thousand Oaks, CA: Sage.

Krathwohl, D. R. (1972). Functions for experimental schools evaluation and their organization. In G.

V. Glass, M. L. Byers, & B. R. Worthen (Eds.), *Recommendations for the evaluation of experimental schools projects of the U.S. Office of Education* (pp. 174–194). Report of the Experimental Schools Evaluation Working Conference, Estes Park, CO, December 1971. Boulder, CO: University of Colorado Laboratory of Educational Research.

Lessinger, L. (1970). *Every kid a winner: Accountability in education*. Palo Alto, CA: SRA.

Miller, H. L. (Ed.) (1997). The New York City public schools integrated learning systems project: Evaluation and meta-evaluation. *International Journal of Educational Research, 27*, 159–174.

Millman, J. (Ed.) (1981). *Handbook of teacher evaluation*. Beverly Hills, CA: Sage.

Millman, J., & Darling-Hammond, L. (Eds.) (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: Sage.

National Center for Education Statistics. (1991). *SEDCAR (Standards for Education Data Collection and Reporting)*. Rockville, MD: Westat, Inc.

National Center for Education Statistics. (1992). *NCES statistical standards*. Washington, D.C.: Author.

Orris, M. J. (1989). *Industrial applicability of the Joint Committee's personnel evaluation standards*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo, MI.

Sanders, J. R. (1995). Standards and principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *New directions for program evaluation* (pp. 47–52). San Francisco: Jossey-Bass.

Sanders, J. R., & Nafziger, D. H. (1977). *A basis for determining the adequacy of evaluation designs.* Kalamazoo, MI: Western Michigan University Evaluation Center. Occasional Paper Series #6.

Schnee, R. (1977). Ethical standards for evaluators: The problem. *CEDR Quarterly, 10*, 3.

Scriven, M. S. (1969). An introduction to meta-evaluation. *Educational Products Report, 2*, 36–38.

Scriven, M. S. (1973). *Maximizing the power of causal investigations—The modus operandi method.* (Mimeo).

Scriven, M. S. (1975). *Bias and bias control in evaluation.* Kalamazoo, MI: Western Michigan University Evaluation Center.

Scriven, M. (1994). Product evaluation: The state of the art. *Evaluation Practice, 15*, 45–62.

Scriven, M. (2000). The logic and methodology of checklists. [On-line]. Available: <www.wmich.edu/evalctr/checklists/>.

Shepard, L. A. (1977). *A checklist for evaluating large-scale assessment programs.* Kalamazoo, MI: Western Michigan University Evaluation Center. Occasional Paper Series Paper #9.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels.* Stanford, CA: National Academy of Education.

Smith, N. L., Chircop, S., & Mukherjee, P. (2000). Considerations on the development of culturally relevant evaluation standards. In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 29–40). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.

Sroufe, G. E. (1977). Evaluation and politics. *NSSE Yearbook, Part 2, 76*, 287.

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. L. (1974). *Meta-evaluation*. Kalamazoo, MI: Western Michigan University Evaluation Center. Occasional Paper Series #3.

Stufflebeam, D. L. (2000). Guidelines for developing evaluation checklists. [On-line]. Available: <www.wmich.edu/evalctr/checklists/>.

Stufflebeam, D. L. (2000). Lessons in contracting for evaluations. *American Journal of Evaluation, 21*, 293–314.

Stufflebeam, D. L. (2001). *Evaluation models. New directions for evaluation, 89.* San Francisco: Jossey-Bass.

Stufflebeam, D. L. (2001). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation, 22*, 71–79.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. (1971). *Educational evaluation and decision making*. Itasca, IL: F. E. Peacock.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1992). *A retrospective analysis of a summative evaluation of NAGB's pilot project to set achievement levels on the national assessment of educational progress.* Annual meeting of the American Educational Research Association, San Francisco.

Taut, S. (2000). Cross-cultural transferability of *the program evaluation standards.* In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 5–28). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.

Vinovskis, M. (1999). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Washington, D.C.: National Assessment Governing Board.

Widmer, T. (2000). Evaluating evaluations: Does the Swiss practice live up to the program evaluation standards? In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 67–80). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.

Widmer, T., Landert, C. & Bacmann, N. (2000). Evaluation standards recommended by the Swiss Evaluation Society (SEVAL). In C. Russon (Ed.), *The program evaluation standards in international settings* (pp. 81–102). Kalamazoo, MI: The Evaluation Center Occasional Papers Series #17.