Meta-Evaluation

Originally published as Paper #3, Occasional Paper Series, December, 1974

Daniel L. Stufflebeam
Western Michigan University

Good evaluation requires that evaluation efforts themselves be evaluated. Many things can and often do go wrong in evaluation work. Accordingly, it is necessary to check evaluations for problems such as bias, technical error, administrative difficulties, and misuse. Such checks are needed both to improve ongoing evaluation activities and to assess the merits of completed evaluation efforts. The aim of this paper is to present both a logical structure and methodological suggestions for evaluating evaluation.

For ease of communication, Michael Scriven's label "Meta-Evaluation" will be used to refer to evaluations of evaluation (Scriven, 1969). The term "primary evaluation" will refer to the evaluations that are the subject of meta-evaluations.

This paper is based on work performed in the Ohio State University Evaluation Center between 1963 and 1973. That work provided many occasions for addressing meta-evaluation issues, including developing evaluation systems and assessing the work of such systems, designing and conducting evaluation studies, training graduate students and practitioners to conduct evaluation work, and critiquing many evaluation designs and reports. These experiences and the attendant problems are the basis for this paper.

Part I of this paper analyzes background factors and problems associated with meta-evaluation, the need for meta-evaluation, and summarizes some of the pertinent literature. Suggestions are made concerning what criteria should guide the development of a meta-evaluation methodology. The final and major portion of Part I is an enumeration of six classes of problems that jeopardize evaluation, and need to be addressed by a meta-evaluation methodology.

The second part of the paper is a conceptual response to the first. Part II defines and sets forth premises for meta-evaluation and presents a logical structure for designing meta-evaluation studies.

The third part of the paper is an application of the logical structure presented in Part II. Basically, Part III contains five meta-evaluation designs, four for use in guiding evaluation work, and the fifth for judging completed evaluation work. Taken together, the three parts of the paper are intended to provide

a partial response to the needs for conceptual and practical developments of meta-evaluation.

I. Background and Problems

The Importance of Meta-Evaluation

The topic of meta-evaluation is timely because evaluators increasingly are being required to evaluate their work. During the past ten years there has been a great increase in evaluation activity at all levels of education. Thousands of federal projects have been evaluated, over half of the states have started work on accountability systems, and several school districts have instituted departments of evaluation. Such activity has cost millions of dollars. It has been of variable quality, and there has been great controversy over its worth. See, for example, Egon Guba's article on the "Failure of Educational Evaluation" (Guba, 1969). Overall, evaluators have come under much pressure to insure and demonstrate they are doing quality work.

Available Meta-Evaluation Concepts and Approaches

The literature of evaluation provides some guidance for evaluating evaluation work. Michael Scriven (1969) introduced the term metaevaluation in the Educational Products Report, and applied the underlying concept to the assessment of a design for evaluating educational products. Leon Lessinger (1970), Malcolm Provus (1973), Richard Seligman (1973), and others have discussed the concept under the label of educational auditing. The APA technical standards for test development (1954) and the-Burg's Mental Measurement Yearbooks (1965) are useful meta-evaluation devices, since they assist in evaluating evaluation instruments. Likewise the Campbell-Stanley (1963) piece on quasi-experimental design and true experimental design is a useful tool for evaluating alternative experimental designs. Campbell and Stanley (1963), Bracht and Glass (1968), The Phi Delta Kappa Study Committee on Evaluation (Stufflebeam et al, 1971 [a]), Krathwohl (1972), and Stufflebeam (1972) have prepared statements of what criteria should be applied in meta-evaluation work.

As a part of an NIE effort to plan evaluation of R and D Centers and regional laboratories, teams chaired by Michael Scriven and myself prepared alternative plans for evaluating lab and center evaluation systems (Scriven et al, 1971), (Stufflebeam et al, 1971 [d]). Richard Turner (1972) has presented a plan for evaluating evaluation systems in NIE's Experimental Schools Program. Thomas Cook (1974) has written an extensive paper on secondary evaluation, and Michael Scriven (1974) recently has developed a paper on how to assess bias in evaluation work. Also, AERA, APA and NCME are in the initial stages of developing a joint statement on technical standards for evaluation. There is, then, an emergent literature in meta-evaluation, and there are some devices for carrying out meta-evaluation work.

The Need for New Meta-Evaluation Concepts and Tools

However, the state of the art of meta-evaluation is limited. Discussions of the logical structure of meta-evaluation have been cryptic and have appeared in only a

few fugitive papers. These conceptualizations lack reference to research on evaluation, and they do not include extensive analyses of problems actually encountered in practical evaluation work. The writings have lacked detail concerning the mechanics of meta-evaluation. While some devices, such as technical standards for tests, exist, the available tools for conducting meta-evaluation are neither extensive nor well organized. Finally, there are virtually no publicized meta-evaluation designs. Overall, the state of the art of meta-evaluation is primitive, and there is a need for both conceptual and technical development of the area.

Meta-Evaluation Criteria

In developing a methodology for meta-evaluation, it is important to have in mind an appropriate set of criteria. These are needed to prescribe necessary and sufficient attributes of evaluation reports and designs. A good place to start is with accepted criteria for research, because both research and evaluation reports must contain sound information.

Criteria for judging research are suggested in the writings of Campbell and Stanley (1963); Gephart, Ingle and items tad (1967); and Bracht and Glass (1968). Basically, these authors have agreed that research must produce findings that are internally and externally valid; i.e., the findings must be true and they must be generalizable. While evaluations must also meet these technical standards, they are not sufficient for judging evaluation findings.

In addition to producing good information; i.e., technically sound information, evaluation must produce findings that are useful to some audience; and the findings must be worth more to the audiences than the cost of obtaining the information—the concern of cost/effectiveness.

These three standards of technical adequacy, utility and cost/effectiveness have been spelled out by Guba and Stufflebeam (1970), and the Phi Delta Kappa Study Committee (Stufflebeam et al, 1971 [a]) in the form of eleven specific criteria.

The criteria of technical adequacy are:

1. Internal Validity. This criterion concerns the extent to which the findings are true. Does the evaluation design answer the questions it is intended to answer? Are the results accurate and unequivocal? Clearly any study, whether research or evaluation, must at a minimum produce accurate answers to the questions under consideration.
2. External Validity. This criterion refers to the generalizability of the information. To what persons and program conditions can the findings be applied? Does the information hold for only the sample from which it was collected or for other groups? Is it time bound, or are the findings predictive of what would occur in future applications? Basically, meeting the criterion of external validity means that one can safely generalize to some population of interest, some set of program conditions, and some milieu of environmental circumstances. Thus, in evaluation (as in research) it is important to define the extrapolations one wants to make from the study results and to demonstrate whether the findings warrant such extrapolations.

3. Reliability. This criterion concerns the accuracy of the data. How internally consistent are the findings. How consistent would they be under test-retest conditions? If the findings lack precision and reproducibility, one should be concerned whether they are simply random and therefore meaningless.
4. Objectivity. This criterion concerns the publicness of the data. Would competent judges agree on the results? Or, are the results highly dependent on the unique experiences, perceptions, and biases of the evaluators. It is possible that findings provided by a set of judges could be reproducible and therefore reliable but heavily biased by the judges' predilections. Unless the findings would be interpreted similarly by different but equally competent experts, the true meaning of the results is subject to question. To meet standards of utility, evaluation reports must be informative to practitioners and must make a desirable impact on their work. The six criteria that are relevant here involve an explicit or implicit interaction between the evaluative findings and some audience.
5. Relevance. This criterion concerns whether the findings respond to the purposes of the evaluation. What are the audiences? What information do they need? To what extent is the evaluation design responsive to the stated purposes of the study and the information requests made by the audiences? The concern for relevance is crucial if the findings are to have more than academic appeal and if they are to be used by the intended audiences. Application of the criterion of relevance requires that the evaluation audiences and purposes be specified. Such specifications essentially result in the questions to be answered. Relevance is determined by comparing each datum to be gathered with the questions to be answered.
6. Importance. This involves determining which particular data should be gathered. In any evaluation study a wide range of data are potentially relevant to the purposes of the study. Since practical considerations dictate that only a part of the potentially relevant data can be gathered, the evaluator should choose those data that will most usefully serve the purposes of the study. To do this, the evaluator needs to rate the importance of each potentially relevant datum and he needs to know the priorities the audiences assign to the various data. Then, based on his own judgments and those of his audiences, the evaluator needs to choose the most significant data.
7. Scope. A further condition of utility is that evaluative information have adequate scope. Information that is relevant and important may yet fail to address all of the audience's important questions. The Michigan Assessment Program (House et al, 1974) is a case in point. This program's purpose is to assess the educational needs of students in Michigan. In practice it mainly provides data about the reading and mathematics performance of 4th and 7th graders. While these

data are relevant and important for assessing certain educational needs of many students, the data are very limited in scope. They pertain to students in only two grades and provide little information about interests, motivation, self-concept, or emotional stability. Nor do they assess needs in science, art, music, or a lot of other areas.

8. Credibility. This criterion concerns whether the audience trusts the evaluator and supposes him to be free of bias in his conduct of the evaluation. Audiences often are not in a position to assess the technical adequacy of a study. The next best thing they can do is decide whether they have confidence in the group that conducted the study. This factor is often correlated with the matter of independence. In some cases the audience for a study wouldn't trust the results if they were self-assessments, but would accept perhaps identical results if they had been obtained by some impartial, external evaluator. In other cases a self-assessment conducted by an internal team might be completely acceptable to the audience. It is crucial that the criterion of credibility be met by the study. However technically adequate the findings may be, they will be useless if the audience puts no stock in their credibility. The meta-evaluator needs to assess how much trust the audience places in the evaluation. Whether an insider or an external agent, the evaluator can do much to insure credibility for his study by carrying out his study openly and by consistently demonstrating his professional integrity.
9. Timeliness. This is perhaps the most critical of the utility criteria. This is because the best of information is; useless if it is provided too late to serve its purposes. In research we are not concerned about timeliness, for the sole aim is to produce new knowledge that is internally and externally valid. It is thus appropriate that researchers take whatever time they need to produce information that is scientifically adequate. In evaluation, however, the purpose is not to produce new knowledge but to influence practice. Therefore, the practitioner must be given the information he needs when he needs it. In evaluation work this almost always creates a conflict. If the evaluator optimizes the technical adequacy of the information he obtains, he almost certainly will not have his report ready when it is needed. If he meets the time constraints of his audiences, he probably will have to sacrifice some of its technical adequacy. The position taken here is that the evaluator should strive to provide reasonably good information to his audience at the time they need it.
10. Pervasiveness. This final utility criterion concerns the dissemination of the evaluation findings. Clearly the utility of an evaluation can be partially gauged by determining whether all of the intended audiences receive and use the findings from the evaluation. If an evaluation report that was intended for use by teachers and administrators were provided to a chief administrator who in turn did

not distribute the findings to his teachers, we would say the findings were not pervasive. This criterion is met when all persons who have a need for the evaluation findings do in fact receive and use them. Overall the four technical and six utility criteria listed above underscore the difficulty of the evaluator's assignment. The evaluator's work must be judged on similar grounds to those that are used to judge the technical adequacy of research reports. But the evaluator's report will also be judged for its relevance, importance, scope, credibility, timeliness, and pervasiveness. To make matters worse for the evaluator, there is yet a third standard to be applied to his work. This is the prudential concern of cost/effectiveness.

11. Cost/effectiveness. This one refers to the need to keep evaluation costs as low as possible without sacrificing quality. Proper application of the utility criteria of relevance, scope, importance, and timeliness should eliminate the grossest of inefficiencies. However, there are always alternative ways of gathering and reporting data, and these vary in their financial and time requirements. Thus, care must be taken to choose the most effective ways of implementing the evaluation design. It is also important that evaluators maintain cost and impact records of their evaluation activities. In this way they will be able to address questions about the cost/effectiveness of their work. In the long run, evaluators must demonstrate that the results of their efforts are worth more than they cost. In some cases, evaluators should be able to show that their studies actually saved more money than they cost, e.g., through influencing the elimination of wasteful activities.

In summary, evaluations should be technically adequate, useful, and efficient. The eleven criteria presented above are suggested to meta-evaluators for their use in assessing evaluation designs and reports. It is apparent that the evaluator cannot insist on optimizing any one criterion if he is to optimize his overall effort. Rather he must make many compromises and strike the best balance he can in satisfying standards of technical adequacy, utility, and cost/effectiveness.

Problems that Jeopardize Evaluation

It is one thing to determine whether evaluation results meet the criteria described above. It is quite another thing to insure that these criteria will be met.

For the latter purpose, one must predict the problems that may jeopardize an evaluation study and introduce appropriate preventive measures.

In my past evaluation experience I have encountered a great many problems, and for some time have thought it would be helpful if evaluators could have available a list of such problems. Such a list would help evaluators predict and counter problems before they happen.

The following pages introduce and delineate six classes of problems that I believe are commonly encountered in evaluation work. These classes are conceptual, sociopolitical, contractual/ legal, technical, administrative, and

moral/ethical. Basically, these problems are suggested for use in improving ongoing evaluation work (the matter of formative meta-evaluation), but they should also prove useful in assessing and diagnosing completed evaluation studies (a concern of summative metaevaluation). Each of the six problem areas is defined and then explicated through the identification of specific subproblems.

Conceptual Problems

This problem area concerns how evaluators conceive evaluation. Evaluation is typically a team activity. As a basis for effective communication and collaboration among the team members, it; is necessary that they share a common and well defined view of the nature of evaluation. Otherwise their activities won't complement each other toward achieving some shared objectives of the evaluation.

Also alternative conceptualizations of evaluation might be adopted. Depending on which one is chosen, the evaluators will produce evaluation outcomes that differ both in kind and quality. For example, a conceptualization that insists on the evaluation of goals will produce different results from one that insists on a goal-free approach, and an approach that emphasizes impressionistic analyses likely will yield less valid and reliable results than an approach that requires close conformance to technical standards. Since adherence to different conceptualizations of evaluation may lead to different results, it is important that teams of evaluators carefully consider and document the approach used to guide their activities.

In addressing this problem area, I believe evaluators should answer eight general questions: What is evaluation? What is it for? What questions does it address? What information does it require? Who should it serve? Who should do it? How should they do it? By what standards should their work be judged? Each is presented and described below.

1. What is evaluation? One can respond to this question in a variety of ways.

One way is to define evaluation as "determining whether objectives have been achieved." This is the most common and classical way of defining evaluation. It focuses attention on outcomes and suggests that stated objectives be used to determine the worth of the outcomes. It doesn't call for the assessment of objectives, project plans, and process, nor does it emphasize ongoing feedback designed to help design and develop projects. "Relating outcomes to objectives," then, is one way of responding to the question "What is evaluation;" and this response has certain characteristics and limitations.

Another possible response is that evaluation is "the process of providing information for decision making." This definition explicitly offers ongoing evaluative feedback for planning and conducting programs. Also this definition is broader than the previously described definition, because providing information for decision making implies that the evaluation would not only focus on outcomes but would also provide information for choosing goals and designs and for carrying out the designs. A chief limitation of this definition is that it does not reference the need for retroactive evaluation to serve accountability.

A third possible response is the one found in standard dictionaries. This one amounts to saying that evaluation is the ascertainment of merit. This definition is broad enough to encompass all questions

about value, quality, or amount that one might imagine, and is not, therefore, as limited as the first two definitions. Also, its generality admits the possibility of providing information for both decision making and accountability. Of course, it is communicable since it is consistent with common dictionary definitions. Its weakness is its lack of specificity.

These three definitions illustrate that there are alternative ways of defining evaluation. Other possibilities include equating evaluation to testing, to professional judgment, and to experimental research (Stufflebeam et al, 1971 [a]). The way that; a group chooses to define evaluation has an important influence on what they produce and is therefore an important consideration in the evaluation of their work.

2. What is it for?

The second question concerns the purposes that evaluation results are to serve. Again, one can respond in alternative ways, and the purposes that an evaluation team chooses to serve can drastically affect what data they collect, how they collect it, how they report it, and how others will judge it.

One possible purpose is implied by one of the mentioned definitions. It is to provide information for decision marling. Invoking this purpose requires that the evaluators place great emphasis on the utility of the information that they gather and report. In effect, they must conduct their evaluation work proactively so as to continually provide timely information for decision making. This is much like Scriven's (1967) notion of formative evaluation.

Another purpose that we hear a lot about these days is accountability. This means maintaining a file of data that persons can use to report and defend their past actions. Serving this purpose calls for a retrospective approach to evaluation which is similar to Scriven's (1967) concept of summative evaluation.

Still a third purpose involves developing new knowledge that is internally and externally valid. In my view, this type of activity is research and not evaluation, and troubles arise when persons equate the two concepts. If the inquirer optimizes the criteria of technical adequacy, his findings will probably lack utility. But if he claims to be doing research and doesn't insist on meeting the criteria of technical adequacy to the exclusion of utility criteria, the outcomes may likely be judged as bad on research grounds—whether or not the findings are useful.

As illustrated above, evaluations may serve different purposes, which suggest different criteria, or at least different emphases, for judging the results of evaluative efforts. Also the evaluators can get into trouble if they set out to serve a different set of purposes than those that their sponsors and audiences have in mind. Thus, evaluators should be explicit about the purposes they are serving, and meta-evaluators should assess the clarity, consensus, and implications of those purposes.

3. What questions do evaluations address?

This third question concerns the foci of the evaluation. What questions might be addressed? Which ones will be addressed?

Classically, evaluations have addressed questions about outcomes. This is certainly one important focus for evaluation work, but it is only one. Evaluations may also assess the merit of

goals, of designs for achieving the goals, and of efforts to implement the designs.

Many different questions might be asked, depending on the substance of what is being evaluated, and the purpose(s) being served. For example, if the purpose is to serve decision making and if the focus is implementation, the evaluator might concentrate on identifying potential barriers; but if the purpose is to serve accountability in relation to the implementation of a design, the evaluator would need to document the total implementation process.

One way of identifying and analyzing potential evaluation questions is through an appropriate matrix. Its vertical dimension includes the purposes of the evaluation, i.e., decision making and accountability. Its horizontal dimension includes the categories of goals, designs, implementation, and results. Figure 1 illustrates the use of such a matrix; its cells have been filled out to illustrate the evaluative questions that might be addressed in an evaluation study.

This matrix illustrates that up to eight categories of questions might be addressed in any evaluative effort, and that many specific questions might be addressed in each of the categories. The metaevaluator can assess what questions are being addressed, whether they are the right ones, and whether they are all the questions that should be asked.

4. What information does evaluation require?

In addition to assessing whether the right questions are being asked, the evaluator also needs to assess whether the right information is being collected. Some evaluations are judged harshly because they contain only summaries of questionnaire data indicating, for example, that "the participants liked the experience." Others are judged to be incomplete because they provide only hard data, such as test scores, and do not reflect the insights of persons who were closely associated with an experience. Still others are criticized because they are devoid of recommendations. Hence, meta-evaluators need to assess whether evaluations provide an appropriate mix of descriptions, judgments, and recommendations.

5. Whom will be served?

The fifth question concerns the audience for the evaluation. What persons and groups will be served? What do they need? How will they be served? These are key issues regarding both the questions to be addressed and the means of reporting back to the audiences.

Invariably, multiple audiences might be served. For example, teachers, researchers, administrators, parents, students, sponsors, politicians, publishers, and taxpayers are potentially interested in the results of evaluations of educational innovations. However, these audiences are interested in different questions, and require different amounts and kinds of information. Hence, evaluation designs need to reflect the different audiences, their different information requirements, and the different reports required to service them. If these matters are left to chance, as is often the case, the evaluation may be expected to fail to meet the needs of some of the audiences. This is because the reports from an evaluation designed to serve one audience likely will not meet the needs of other audiences.

Purpose of Evaluation Studies	Categories of Evaluation Questions			
	Goals	Designs	Implementation	Results
Pro-active Evaluation to serve Decision Making	Who is to be served? What are their needs? What problems have to be solved if needs are to be met? What funds are available for work in this area? What research findings have a bearing on problem solving in this area? What relevant technology is available? What alternative goals might be chosen?	Are the given objectives stated operationally? Is their accomplishment feasible? What relevant strategies exist? What alternative strategies can be developed? What are the potential costs and benefits of the competing strategies? What are the operating characteristics of the competing strategies? How compatible are the competing strategies with the system? How feasible are the competing strategies?	What is the schedule of activities? What are the personnel assignments? What's the program budget? What potential problems attend the design? What are the discrepancies between the design and the operations? What design changes are needed? What changes in implementation are needed?"	What results are being achieved? Are they congruent with the objectives? Are there any negative side effects? Are there any positive side effects? Do the results suggest that the goals, designs, and process should be modified? Do the results suggest that the project will be a success?
Retroactive Evaluation to serve Accountability	What goals were chosen? What goals were considered, then rejected? What alternative goals might have been considered? What evidence exists to justify the goals that were chosen? How defensible is this evidence? How well have the goals been translated into objectives? Overall, what is the merit of the goals that were chosen?	What strategy was chosen? What alternative strategies were considered? What other strategies might have been considered? What evidence exists to justify the strategy that was chosen? How defensible is this evidence? How well was the chosen strategy translated into an operational design? Overall, what is the merit of the chosen strategy?	What was the operational design? To what extent was it implemented? What were the strengths and weakness of the design under operating conditions? What was the quality of the effort to implement it? What was the actual design that was implemented? Overall, what is the merit of the process that was actually carried out?	What results were achieved? Were the stated objectives achieved? What were the positive and negative side effects? What impact was made on the target audience? What long-term effects may be predicted? What is the relation of costs to benefits? Overall, how valuable were the results and impacts of this effort?

Figure 1
A Matrix for Identifying and Analyzing Evaluative Questions

This was dramatically illustrated in the U. S. Office of Education evaluation of the first year of the Title I Program of the Elementary and Secondary Education Act (Department of Health, Education and Welfare, 1967). This multi-billion dollar program, designed to upgrade educational opportunities for disadvantaged children, was of interest to many different audiences.

Two such audiences were local educators and Congressmen. Their interests were quite different, however. Local level educators were especially concerned about how to make the individual projects succeed. The Congressmen wanted to know what the total program was accomplishing. Clearly, no single evaluation study or report could serve the different needs of these two audiences.

The U. S. Office—being responsible to the Congress for evaluating the Title I Program—had to decide on the audiences, questions, design, and reports for a national evaluation of Title I. USOE officials did not distinguish between different audiences to be served, nor did they plan how different information requirements at national and local levels would be met. USOE officials allowed each school district to design its evaluation exclusively to serve local information requirements. Due to potential political problems, no requirements were placed on the schools to use common instruments by which information could be gathered on a uniform basis for submission to the Congress.

Incredible as it may seem, USOE officials assumed that the thousands of local school Title I evaluation reports could be aggregated into a single report that would respond to interests of the Congress. USOE did develop a report that attempted to integrate and aggregate the local school reports, but the result was a disaster and an embarrassment to all concerned.

This illustrates that it is important early in the process of designing an evaluation to carefully identify and analyze the information needs of the different audiences for the evaluation. This type of audience analysis must be used in designing data collecting and reporting activities. Careful attention to this area can assist greatly in satisfying criteria of relevance, importance, scope, and timeliness for the evaluation. The meta-evaluator, then, will do well to assess evaluation designs for their attention to the audiences to be served.

6. Who should do the evaluation?

This sixth question concerns the agent for the evaluation. Should educators do their own evaluations? Should they employ evaluation specialists and have them do it? Should they subcontract to some external evaluation company? Should the educators do their own evaluation but engage an external auditor to check their work? Or what?

These questions are complicated but they become even more complicated when the dimension of purpose of the evaluation is added. Who should do evaluation for decision making? Who should do evaluation that is intended to serve accountability?

Answers to these questions are important, because there are different costs, benefits, and problems associated with the use of different evaluation agents. It may be cheapest to do one's own evaluation, but to do so invariably sacrifices the important criteria of objectivity and credibility. Conversely, the employment of external evaluation agents enhances objectivity and credibility, but it can increase disruption, costs, and threat. The meta-evaluator should check on how the question of evaluation agents has been handled, or might be handled, and he should assess alternate consequences of the different possible arrangements. For further information on this problem, see Scriven's recent and important paper on bias control. (Scriven, in press).

7. How should the evaluation be conducted?

The seventh question regarding the conceptualization of evaluation concerns the methodology of evaluation. What is the process of evaluation? What steps have to be implemented in the course of doing an evaluation? To what extent have sound procedures been worked out for each step in the evaluation process?

There are alternative conceptualizations of evaluation and each one has its different steps. While most authors do not view the evaluation process as linear, they have recommended varying lists of steps that evaluators should carry out. Stake (1967) has suggested an approach that involves describing a program, reporting the description to the audience, obtaining and analyzing their judgments, and reporting the analyzed judgments back to the audience.

Michael Scriven (1972 [a]) has suggested nine steps in his Pathway Comparison Model. They are: (1) characterizing the program to be evaluated, (2) clarifying the conclusions wanted, (3) checking for cause and effect relationships, (4) making a comprehensive check for consequences, (5) assessing costs, (6) identifying and assessing program goals, (7) comparing the program to critical competitors, (8) performing a needs assessment as a basis for judging the importance of the program, and (9) formulating an overall judgment of the program.

Newton S. Metfessel and William B. Michael (1967), in writing about Ralph Tyler's rationale for evaluation, have suggested an eight step evaluation process. Their steps are: (1) involvement of all interested groups, (2) development of broad goals, (3) construction of behavioral objectives, (4) development of instruments, (5) collection of data, (6) analysis of data, (7) interpretation of the meaning of the findings, and (8) formulation of recommendations.

Malcolm Provus (1971) has proposed a five step process. It is (1) clarifying the design of a program, (2) assessing its implementation, (3) assessing its interim results, (4) assessing its long-term results, and (5) assessing its costs and benefits.

As a final example, the Phi Delta Kappa Study Committee on Evaluation (Stufflebeam et al, 1971 [a]) presented a three-step process. It included (1) interacting with the audiences to delineate information requirements, (2) collecting, organizing, and analyzing the needed information, and (3) interpreting and reporting the findings back to the audience.

These different conceptions of the evaluation process illustrate that evaluators will do different things depending on which conceptualization of evaluation they use. If Scriven's approach is followed, great attention will be given to steps that insure the technical adequacy and inclusion of judgments in the evaluation; but little concern will be given to interactions (with audiences) that are designed to insure the utility of the evaluation reports. Conversely, the other approaches place heavy emphasis on interactions with audiences to insure that the obtained information will be used by the intended audiences.

The meta-evaluator should identify what evaluation process is being followed, examine the implications of the selected process in relation to criteria of technical adequacy, utility, and cost/effectiveness; and check on the provisions for carrying out the evaluation process. Feedback of such information to evaluators should help them decide whether their design is in need of modification or explication.

8. By what standards should the evaluation be judged?

This paper has already asserted that evaluations should meet standards of technical adequacy, utility, and cost/effectiveness. These standards were further defined in the form of the eleven criteria of external validity, internal validity, reliability, objectivity, relevance, scope, importance, credibility, timeliness, pervasiveness, and cost/efficiency. In accordance with this position, meta-evaluators should assess the extent to which evaluations have been designed to meet these criteria.

There are a number of considerations in making such checks. What priorities do the evaluators assign to each of the eleven criteria? What priorities do the audiences apply to the different criteria? Are the priorities for different criteria likely to be in conflict? To what extent is the overall conceptualization of evaluation consistent with the standards of adequacy for the evaluation that evaluators and their audiences have in mind?

This concludes the discussion of conceptual problems in evaluation. In summary, evaluators and their audiences need to hold in common some defensible conceptualization of evaluation that can guide their collection and use of evaluation data. There are alternative conceptualizations that might be adopted. Meta-evaluators are encouraged to check the clarity, common acceptance, and adequacy of the conceptualization being used by a particular group.

This section has presented some of the questions and alternative responses to be considered. Given this analysis of conceptual problems, we next turn to the sociopolitical problems.

Sociopolitical Problems

This problem area reflects that evaluations are carried out in social and political milieus. Thus, the evaluator must face many problems in dealing with groups and organizations.

Unless the evaluation design includes provision for dealing effectively with the people who will be involved in and affected by the evaluation, these people may well cause the evaluation to be subverted or even terminated. As any evaluator can testify, sociopolitical problems and threats are real; and evaluation training programs and textbooks do not prepare evaluators to deal with these problems. In evaluation it is of utmost importance to check for the existence of potential sociopolitical problems and to plan how they can be overcome. My list includes seven such problems.

1. Involvement

This first sociopolitical problem concerns interaction with the persons on whose cooperation the success of the evaluation depends. A principle of educational change is that unless persons who will need to support the change are involved early, meaningfully and continuously in the development of an innovation, they likely will not support the operation and use of the innovation.

This principle applies in evaluation as well. Bettinghaus and Miller (1973) have pointed this out in their analysis of resistance throughout Michigan to the newly developed state accountability system. Their explanation is that much of the resistance would not have developed if more people throughout Michigan had been involved earlier in the design of the Michigan Accountability System.

Evaluation and accountability at best are threatening concepts. If persons whose work is to be evaluated are not involved in discussions of criteria by which their work will be judged, methods by which data will be supplied, and audiences who will receive the reports, these persons can hardly be expected to be supportive of the evaluation. More likely they will resist, boycott, or even attempt to subvert the evaluative effort.

What can the evaluator do to involve persons whose support is required if the evaluation is to succeed? One thing he could do is to design the evaluation work to the last detail then present the design at a meeting comprised of persons representing all interested parties. While he could do this, and while evaluators often do it this way, this is just about the worst thing they can do.

Presenting a "canned" design to previously uninitiated but interested persons at a large meeting is pregnant with involvement problems. The attendees likely will include administrators, sponsors, evaluators, and teachers. Certainly some of the persons will be reluctant participants, and none of them, outside the designer of the evaluation, will have any commitment to the prepared design. Any one person who wants to delay or cancel the evaluation task will find it easy at such a meeting to rally support for his questions and reservations. The evaluator, on the other hand, may find no one in his "corner." So the first checkpoint in regard to the involvement problem is that evaluators not plan to orient participants in the evaluation through presenting them with a finished evaluation design at a large group meeting.

Instead, evaluators must involve groups in the development of the design before it is ever presented in anything like final form. Small advisory panels can be established and convened for the purpose of hearing their recommendations. Small groups can be engaged in working sessions to provide recommendations regarding such matters as logistics. Much individual contact with interested persons should be arranged, both face-to-face and via telephone and mail, especially to get their views of what questions should be addressed by the evaluation. Unless there is some compelling reason for it, the evaluator should probably avoid holding a large group meeting to review the evaluation design; it is preferable to hold several small group meetings.

The point here is that many interested persons should be involved in developing evaluation designs to win their cooperation. The metaevaluator should examine the evaluation for evidence that persons whose support is needed are provided opportunities for real input into the evaluation planning. The meta-evaluator should also check for the existence of unnecessary situations in which adversaries of the evaluation might be given opportunities to cause the evaluation to fail. Accordingly, evaluation designs and activities should be checked for their provisions against problems that may result from a lack of involvement of interested persons in the evaluation or from bad plans for involving people.

2. Internal Communication

The second sociopolitical problem is that of internal communication. Evaluations involve many activities that are not routine for persons in the system where the evaluation is being conducted. At best these activities are disruptive, but they can become intolerable to system personnel if they occur as a succession of surprises. Conversely, if system personnel

do not understand their roles in the evaluation, they can't perform them. If they don't perform them, the evaluation can hardly be successful. Also, audiences for the evaluation can't use evaluative data if they do not know it exists. The point of this discussion is that evaluation activities should be supported by some system of ongoing internal communication.

The internal communication should focus particularly on data collection and reporting. Periodically all persons who are involved in data collection should be informed about what groups will be involved, in what ways, in providing what data, at what times. Figure 2 presents one way of communicating such information to interested persons. This figure is a chart that shows who is scheduled to respond to designated data collection instruments for each day of some explicit period. Likewise, Figure 3 is a similar chart that shows what audiences are scheduled to receive what reports on what days. The preparation of such charts can be used to inform system participants about their future involvement in the evaluation. Of equal importance, the projection of such calendars can aid evaluators to identify conflicts and feasibility problems in their data collection and reporting schedules.

These charts are suggested as devices that meta-evaluators can use to check for communication problems. The completed charts can be used to check whether the system personnel and evaluators have common understandings of their evaluation responsibilities. The charts can also be used to help the evaluators discover feasibility problems in their plans.

To insure that internal communication is systematically maintained, evaluators can use a number of techniques. They can report at staff meetings. They can issue weekly evaluation project newsletters and they can maintain advisory boards that represent the system personnel.

It is important that evaluators use appropriate means to maintain good communication with system personnel. This is necessary both to insure their cooperation—which is necessary for the technical adequacy of evaluation efforts—and to insure that the evaluation findings will be used. Meta-evaluators can provide valuable service to evaluators through checking their evaluation plans for the adequacy of provisions for internal communication.

3. Internal Credibility

A third sociopolitical problem concerns the internal credibility of the evaluation. Particularly this involves the extent that system personnel trust the evaluator to be objective and fair in his collection and reporting of data.

A common characteristic of evaluations is that evaluators must often collect data from persons at one level of a system and then collate the data and report them to persons at the next higher level of the system. For example, it is common that data must be collected from teachers for development of a school-level report to serve the school principal. This characteristic of evaluation causes a natural threat: persons who respond to evaluative queries wonder whether they are being evaluated and whether the data they provide will be used against them. It is little wonder that evaluators' requests that educators respond to interviews and questionnaires are often met with anxiety. To secure the cooperation of potential respondents to evaluation instruments, evaluators must clarify how the collected data will be reported and used; and the

evaluators must establish a climate of mutual trust and cooperation. Particularly evaluators must clarify who will receive the evaluation reports and whether the reports will be used to evaluate the persons who supplied the basic data. The evaluators must say whether or not they can guarantee anonymity; if they make such guarantees, they must demonstrate how they can live up to their commitments. Above all, the evaluators must constantly demonstrate the highest level of professional integrity.

If there are problems of internal credibility, the technical adequacy and utility of the evaluation will be threatened. To check for such problems, a meta-evaluator can pose questions to the evaluator that he might be asked by potential suppliers of data for the study. Cover letters for questionnaires can also be reviewed, and potential evaluation respondents can be interviewed. Feedback to the evaluator should identify areas that need clarification, contradictions in various communications to data suppliers, concerns of the data suppliers, and suggestions for strengthening the internal credibility of the study.

4. External Credibility

This fourth sociopolitical problem involves whether persons outside the system being evaluated have confidence in the objectivity of the evaluators.

Groups	11/1	11/22	11/25	12/15	1/20	1/23	2/20
All school principals in the district		1.	1.	2.			
A 10% random sample of 3rd grade teachers	1.			2.		1.	
All 3rd grade students in the district	3.	4.	5.	3.	5.	4.	3.
All members of the central ad-ministration					2.		
All school counselors							
A 10% random sample of parents of 3rd graders		2.					
All board members						2.	

Data Gathering Devices

I. System-level Interview Form
II. General Survey Form
III. State's objective-referenced test in reading
IV. Special scale to assess attitude toward learning
V. Physical examination

Figure 2
A Hypothetical Projection of Various Groups' Involvement in Responding to Evaluation Instruments

Groups		3/20	3/30	4/10	4/11	4/15	5/2
1.	Sponsor	1,3	2,4				
2.	News Media		2,4	5	6		
3.	System Administrator	1,3	2,4			7	7
4.	Teachers		2,4			7	7
5.	Parents				2,4	7	7
6.	Citizens at large				2,4	7	7

Evaluation Reports

1. Preliminary Technical Report
2. Final Technical Report
3. Preliminary Summary Report
4. Final Summary Report
5. Press Conference
6. TV Presentation
7. Neighborhood Hearings

Figure 3
A Hypothetical Projection of the Reporting of Evaluation Findings

To the extent that evaluators have done a good job with internal credibility, they are likely to encounter external credibility problems. If people inside the system are comfortable with and confident in the evaluator, people outside the system may think the evaluator has been co-opted. This is because outsiders commonly expect the evaluator to do an independent, objective, hard-hitting assessment of merit and they take it for granted that insiders will resist and be anything but confident in the evaluator.

The internal credibility/external credibility dilemma is a common and difficult one for evaluators. However, the technical adequacy and utility of the evaluation depends on the evaluator being credible to both insiders and outsiders. The evaluator must, therefore, be alert to problems in both areas and he must strive to overcome them.

5. Security of data

One of the ways to enhance the internal credibility of the evaluation is through attending to the fifth sociopolitical problem. This is the problem of security of data which, of course, concerns whether the obtained data are under the complete control of the evaluator.

To be kept, guarantees of anonymity must be backed by strong security measures. Some of these are common sense, such as storing data in locked files and strictly controlling the keys to the files. Another effective method is to insist that respondents not place their names on the forms they fill out. Also, matrix sampling can be used-as in the case of "National Assessment"—to prevent any person, school, or school district from being identified by a particular score. Of course there are limits to the guarantees

of security that can be upheld as became evident in the infamous Watergate case. The evaluator should provide reasonable assurances, he should make provisions for upholding these, but he must not make promises that he cannot keep.

Problems of security can influence the evaluator's ability to collect data and thus affect the technical adequacy of the results. In the long run, if security of data is not maintained, the evaluator will likely encounter great resistance in his attempts to collect data.

6. Protocol

One commonly hears that school districts and schools maintain standard protocol procedures that outsiders are expected to use. Problems in this area may develop when evaluators don't find out and use the system's protocol procedures.

Essentially, protocol involves interactions with the chain of command. In some schools outsiders must always get clearance from a teacher's or principal's immediate superior before visiting or communicating with the teacher or principal about school affairs. Also it is common for school superintendents to clear contacts for outsiders with school board members. In some cases evaluators are asked not to contact school personnel until a school official has formally announced and sanctioned the evaluation plans. In extreme cases, administrators have been known to require that evaluators be accompanied in their visits to school personnel by the administrator or his representative. Clearly, there are many alternative protocol arrangements that evaluators may be expected to honor.

Such requirements present a dilemma to the evaluator. While it is inexcusable for evaluators not to find out what protocol expectations exist, it is not at all clear on a priori grounds how they should respond to them. If the evaluator doesn't first clear his questionnaires with the school principal, the teachers may not respond to the questionnaires. If the evaluator goes along with an administrator's request that questionnaires be administered and collected by the administrator, this may bias how teachers respond to the instruments. If the evaluator allows the administrator to be present in interviews, a serious question will exist concerning the validity of the interview results. Thus the evaluator must deal carefully with the deceptively simple-appearing problems of protocol.

7. Public relations

The seventh and final sociopolitical problem is that of public relations. This problem concerns the public's and the news medians interest in evaluation work and how evaluators should treat such interests.

Evaluations are often of interest to many groups—sometimes for the evaluations' informative aspects and other times for their sensational qualities. Thus, reporters frequently seek to learn about the nature and findings of evaluation studies. Newspaper articles, press conferences, television releases, etc., are common occurrences in evaluation work. As a consequence, evaluators, whether they like it or not, must deal in public relations.

This situation, like so many others, presents the evaluator both with opportunities and problems. Cooperation with the news media is a desirable means of keeping the public informed about the evaluation activities and results. However, reporters are not always respectful of the evaluator's concern for controlling what

information is publicly disseminated; hence, if they can get it, reporters may publicly report information that the evaluator had agreed to report privately to some restricted audience. Also reporters may edit and slant an evaluator's report. If the utility of their findings are not to be jeopardized, evaluators must work very carefully with representatives of the news media.

The posture of this paper is that evaluators should take the initiative in the public relations area. They should make contact with reporters. They should project a schedule of news releases, and they should reach agreements about what information is out-of-bounds for public release. Protocol should be established for the release and editing of the evaluative information.

The main problem to be avoided in the public relations area is in avoiding it. Meta evaluators should probe to find out what arrangements have been made in this area, and they should critique these arrangements for their appropriateness.

With the public relations problem the discussion of sociopolitical problems has been completed. The seven problems in this area remind one that evaluations occur in the real world and that evaluators must be mindful of this if their work is to be technically adequate and useful. Meta-evaluators can help by checking for the existence of sociopolitical problems and developing appropriate recommendations.

Contractual/Legal Problems

The third problem area, pertaining to contractual and legal matters, indicated that evaluations need to be covered by working agreements among a number of parties both to insure efficient collaboration and to protect the rights of each party. Successful evaluation requires that evaluators, sponsors, and program personnel collaborate. If this collaboration is to be effective, it needs to be guided by working agreements. If these are to hold, they often need to be in the form of some legal instrument such as letters of agreement and contracts. Such legal instruments should be reflective of possible disputes that might emerge during the evaluation and of the assurances that each party requires in regard to these possible disputes.

One way of conceptualizing contractual and legal problems in evaluation is to project items that might be standard in most contracts between an external evaluation agent and some system or sponsor. I have in mind eight such contractual items. They are: (1) definition of the client/evaluator roles, (2) specification of evaluation products, (3) projection of a delivery schedule for evaluation products, (4) authority for editing evaluation reports, (5) definition of the limits of access to data that must be observed, (6) the release of evaluation reports, (7) differentiation of responsibility and authority for conducting evaluation activities, and (8) the source and schedule of payments for the evaluation. Satisfactory performance in these task areas is essential if the evaluation is to be conducted efficiently and if it is to succeed in meeting standards of technical adequacy and utility. Each of these contractual and legal problems is defined in more detail below.

1. Definition of the Client and Evaluator Roles

Clarifying the roles involved in the evaluation work and identifying the agencies and persons who are responsible for those roles is the first contractual/legal

problem. Problems of role clarification are common in programs, whether they occur within agencies or involve relationships among several agencies. If the evaluation is to be conducted smoothly and if it is to serve its audiences well, the roles required for conducting and using the evaluation must be defined and the agencies and persons who will be responsible for these roles must understand and accept their roles. Hence, the legal agreements that govern evaluation work must clearly define the roles to be implemented.

Basically, evaluation functions can be grouped according to the main roles of sponsor, evaluee, and evaluator. These roles may be implemented independently by separate agents, or they may be combined and assigned to agents in a variety of ways. The extreme, but not unusual, case is when all three roles are assigned to one person. This, of course, is the instance of self-evaluation. A number of questions can be asked to determine whether the evaluation roles have been adequately defined.

Concerning the role of sponsor, who commissioned the evaluation, and who will pay for it? Why do they want it conducted? What support will they provide for it? To what extent do they intend to participate in gathering information? To what extent will the sponsor's work be a subject of the evaluation? What information do they want? How will they use it? By what authority have they commissioned the evaluation? These and similar questions are appropriate for determining whether the role of sponsor has been clarified to the satisfaction of all parties who must enter into an agreement for the conduct of an evaluation.

There are also a number of questions to be considered in clarifying the role of the evaluee. Whose work will be evaluated? What is the nature of their work? Are they bound to cooperate? Have they agreed to cooperate? What do they expect to receive from the evaluation? What do they require as conditions for conducting the evaluation? How will they participate in the evaluation? What is their relationship to the sponsor and the evaluator? Clearly it is important that such questions be answered by the main parties to the evaluation if the evaluee is to play a constructive role in the evaluation.

A third role is that of evaluator. What group will do the evaluation? What is their relationship to the sponsor? To the evaluee? What are their qualifications to perform the evaluation? Why have they agreed to conduct the evaluation? What services do they expect to perform? What persons will they assign to perform this work? What support do they require? General responses to these questions provide a basic definition of the evaluation role to be served. Of course, the evaluator's role is explicated in the detail of the technical evaluation design.

There are then a number of roles that need to be defined and included in the written agreements that govern evaluation. By including these definitions in the legal instruments that govern evaluation there is a basis for allocating specific areas of responsibility and authority in the evaluation work. Placing agreements about roles in the evaluation contract gives assurances and safeguards concerning collaboration among the various groups that must support the evaluation.

2. Evaluation Products

The second contractual/legal item concerns the products to be produced by the evaluation. Just as program personnel

should clarify their objectives, so should evaluators specify the evaluation outcomes to be produced.

Basically, evaluation outcomes refer to the reports to be prepared. How many reports are to be produced? What are their content specifications? How will they be disseminated? Who will use the reports? How will the quality of the reports be assessed? Generally, it is desirable that the different parties to an evaluation reach agreements early concerning the evaluation products to be produced.

3. The Delivery Schedule

Related to the evaluation products is the delivery schedule for the specified evaluation products. If they are to be useful, they must be timely. Hence, it is important to determine in advance when the evaluation reports will be needed and to reach agreements about whether the reports can be produced on such a schedule.

Attempts to reach such agreements often reveal potential timing problems. To meet the sponsor's timetable, the evaluator often would have to sacrifice the quality of his work. But meeting his own qualitative specifications would often prevent the evaluator from producing timely reports. Frequently evaluators and sponsors must compromise concerning technical and time requirements in order to insure that the evaluation will achieve a reasonable balance of technical adequacy and timeliness. It is best that such compromises be affected early in the evaluation work. For this reason, the timing of evaluation reports should be worked out and included as a specific item in the statement of agreement that will govern the evaluation work.

4. Editing Evaluation Reports

Basically, this concerns who has authority for final editing of evaluation reports, but it also concerns the need for checks and balances to insure that reports contain accurate information. Evaluators, sponsors, and evaluees have legitimate concerns regarding editing.

The evaluator needs assurances that he has the ultimate authority in determining the contents of the reports that will carry his name. There are all too many instances of evaluation reports being edited and released by sponsors, without first getting the approval of the evaluator. It is not proper for sponsors to revise evaluation reports so they convey a different (usually more positive or negative) meaning than that presented by the evaluator. It is proper and often a necessary protection that evaluators require an advance written agreement that they will retain final authority regarding the editing of their reports.

But, sponsors and evaluees also deserve certain assurances regarding editing. All evaluation procedures are subject to error. Therefore, all evaluation reports potentially contain misinformation. Moreover, the reporting of false results can be unjustly damaging. Hence, it is reasonable that sponsors and evaluees require that evaluation designs contain reasonable checks and balances to guarantee the accuracy of evaluation reports before they are released.

A suggested contractual provision covering editing of evaluation reports is as follows:

a. The evaluator will have final authority for editing evaluation reports.
b. The evaluator will provide a preliminary draft of his report to

designated representatives of the evaluee and sponsor for their review and reactions.

c. These representatives will be given a specified number of days within which to file a written reaction to the report.
d. If received prior to the deadline, the evaluator will consider the written reactions in the preparation of the final report.

These points are intended to guarantee final editing authority to the evaluator, but to provide the evaluee and sponsor with a means of raising questions about the accuracy of preliminary reports. The point is that evaluations involve potential disputes over editing and accuracy that can be minimized through the reaching of advance written agreements. At the same time, the evaluator must realize that when the evaluee is given an opportunity to review the draft report he is also being supplied with an opportunity to discredit the evaluation if he so desires.

5. Access to Data

Generally, evaluators must gather existing data from files and new data from system personnel. This situation presents potential problems to evaluees and sponsors as well as evaluators.

The evaluees and sponsors have a special concern for protecting the rights of system personnel and for maintaining good relationships with them. Certain data in system files are confidential. The system administrators need to guard the confidentiality of this information or reach special agreements about its use in the evaluation. Also, system personnel are not automatically willing to submit to interviews or to respond to lengthy questionnaires. Nor, based on their contracts, are they bound to do so. If their cooperation is to be obtained, it must be requested in advance, and agreements with the system personnel need to be worked out. Hence, the evaluees and sponsors have an interest in writing advance agreements about access to data.

Of course this is a crucial item as far as the evaluator is concerned. He can't conduct his evaluation unless he can get the data he needs. Hence, he also needs to have advance agreements concerning what information he can expect to get from system files, and concerning what new data he can obtain. If the evaluator can't get such assurances in advance, his work is in jeopardy, and he might just as well cancel the evaluation before it starts.

6. Release of Reports

Basically, this is a matter of who will release the reports and what audiences may receive them.

A potential problem exists in the possibility that the evaluations may be released by the sponsor only if they match his predilections and serve his ulterior motives. This, of course, is a biased use of evaluation and is to be avoided by professional evaluators. Instead they should insist that their reports be provided to the prespecified audiences regardless of the nature of the findings. If there is some doubt about whether the sponsor will release the report to the prespecified audience, the evaluator, in writing, should reserve the right to do so himself.

A related problem is in determining what audiences should receive what reports. In some cases, for example, in evaluating the early developmental work of a new program, it is entirely appropriate that the developers engage an evaluator to provide them with private

feedback for their own use in improving their work. If the evaluator and developers agree to this condition in advance, it would be inappropriate for the evaluator to release his report to the public. In other cases the evaluator and the sponsor might appropriately agree that a report on the overall merit of a program be developed and released to the public. In such a case, if the sponsor didn't like the results and decided not to make them public, the evaluator should release the results. Otherwise, his integrity and the credibility of his work will be justifiably threatened.

It is patently evident that evaluators and their sponsors should agree in advance regarding what reports should be released to what audiences. Not all reports should be released to all audiences. But reports should not be selectively released based on the nature of the findings. Both evaluators and their sponsors need assurances in this matter. It is, therefore, urged that their advance written agreements contain an item pertaining to the release of reports.

7. Responsibility and Authority

A prior contractual item concerned the definition of roles for the evaluator, the evaluee, and the sponsor. This seventh item, concerning responsibility and authority, emphasizes that specific work needs to be performed by each group in the conduct of the evaluation and that specific agreements about work assignments should be worked out in advance. Including this item in the contract is intended to insure that the rights of all parties will be protected and that the evaluation design will be implemented.

Any evaluator knows that he can't do everything that is required to implement an evaluation. Cooperation is required from many different groups. Administrators must secure the cooperation of their staffs; and teachers, students, administrators, community personnel, and others often are asked to provide information. Often, teachers are engaged to administer tests to their students. In short, the evaluator is dependent on receiving help from many groups in carrying out the evaluation design.

But, evaluators don't have automatic authority to assign responsibilities to the various groups on whose cooperation the success of the evaluation depends. They either need to define and work from explicit agreements concerning who will do what, or they need to depend on their wits and the good will of the people with whom they intend to work. By far the best practice is to work out advance written agreements that delineate areas of authority and responsibility for all parties who will be involved in the evaluation.

8. Finances

The eighth and final contractual item concerns finances for the evaluation. Who will pay for the evaluation? How much money has been budgeted for it? How may this money be spent? What is the schedule of payments? What are the conditions for payment? How is the schedule of payments correlated with the delivery schedule for evaluation reports? The matter of finances is, of course, the most common one in evaluation contracts. Advance agreements regarding finances should be written to protect both the sponsor and the evaluator. The sponsor should insure that payments will be made only if the evaluation objectives are achieved. The evaluator should be assured that funds will not be cut off midway in the evaluation due to the nature (as

opposed to quality) of the results that are produced. Hence, the evaluator and the sponsor should agree in advance to a schedule of payments that is dependent only on the evaluator meeting the mutually agreed upon product specifications.

This concludes the discussion of contractual/legal problems. Basically, all parties involved in an evaluation require protection and assurances. The suggestion here is that the evaluation be governed by an advance contract developed and signed by the appropriate parties. The meta-evaluator's concern here should be to ascertain whether the evaluation is covered by a set of written agreements that would adequately forestall potential contractual and legal problems in the evaluation.

So far, this discussion of evaluation problems has considered conceptual, sociopolitical, and contractual/legal problems. But, little has been said about technical problems, which are the ones that have received the most attention in the formal literature of evaluation. By considering technical problems fourth in the discussion of six classes of evaluation problems, the point is hopefully being made that technical problems are one important type of problem the evaluator must face, but by no means the only type.

Technical Problems

Evaluators must be prepared to cope with a wide range of difficult technical problems, including nine that are discussed in this section. Attention to these items should assist evaluators to convert an abstract evaluation plan to a detailed technical plan for carrying out the evaluation work.

1. Objectives and Variables

The first technical problem concerns the identification of the variables to be assessed. The problem here is twofold. First, there are potentially many variables that might be included in any study, and the evaluator has the difficult task of identifying and choosing among them. Second, it is usually not possible to choose and operationally define all the variables before the study starts; hence, the evaluator often must continually add new variables to his evaluation design. Meta-evaluators should check evaluation designs for their inclusion of variables that meet conditions of relevance, scope, and importance; and the meta-evaluators should check designs for their flexibility and provisions for adding new variables through the course of the study.

There are a number of ways of dealing with the problem of identifying evaluative variables. The classic way is to get program personnel to define their objectives in behavioral terms. This focuses the evaluation on what the program personnel perceive to be desirable outcomes. Devices that are of assistance in defining objectives include the Bloom (1956) and Krathwohl (1964) taxonomies of cognitive and affective objectives. Also, an enormously useful article by Metfessel and Michael (1967) presents a long list of behavioral indicators for use in evaluation studies.

This focus on objectives has served well in countless studies, but it yields variables that are limited in scope. For example, if evaluators focus exclusively on those variables that relate to the developers' objectives, other important outcomes and side effects may be missed. Also, variables such as cost, readability of materials, staff time in a program, and socioeconomic background of students

may be ignored. Hence, there is a need for a broader framework of variables than that afforded in the concept of behavioral objectives.

A number of broader perspectives have been suggested in the literature. Clark and Cuba (1967) have suggested a range of variables that they believe should be considered in assessing various change process activities. Hammond (1969) has proposed his EPIC cube as a means of choosing variables that reflect student behavior, institutional involvement, and curricular elements. Hammond (1975) presented an algorithm based on facet analysis wherein evaluators and program personnel may systematically assign priorities to the potential variables in the EPIC cube. Stake in his Countenance Model (1967) has suggested a framework that interrelates antecedent conditions, transactions and outcomes with program persons' intents and evaluators' observation. These perspectives illustrate that the views of what variables should be incorporated in evaluation have broadened greatly from the early ideas that evaluations should focus exclusively on outcomes that relate to given objectives.

2. The Investigatory Framework

The second technical problem concerns what investigatory framework should guide the evaluation. An investigatory framework specifies the conditions under which data are to be gathered and reported, and the assumptions to be made in interpreting the findings. In all evaluation studies, evaluators must choose either implicitly or explicitly among a number of alternative investigatory frameworks, e.g., experimental design, survey, case study, and site visitation.

No one investigatory framework is superior in all cases. None is always best in serving the criteria of technical adequacy, utility, and efficiency. Also, different frameworks work differentially well under different sets of feasibility constraints. Thus, evaluators may choose different investigatory frameworks depending on the evaluative purposes to be served, the priorities assigned to the different criteria for judging evaluation reports, and the unique conditions under which evaluations are to be conducted. The task is to choose the framework that will optimize the quality and use of results under realistic constraints.

Whereas true experimental design is theoretically the best way of determining cause and effect relationships (through its provisions for internal and external validity), it is often not feasible to conduct true experiments. This is because it is frequently impossible to control treatment variables and the assignment of the experimental units to the treatment conditions. For example, one would not use experimentation to assess the effects of Sputnik I on subsequent U. S. Educational policy. Neither would one say that it is not appropriate to make post hoc evaluative interpretations about such linkages. Also—regarding the matters of relevance, scope, and timeliness—experimental design often would not assess the right variables or provide timely feedback for decision making. This is especially true when the concern is to conduct needs assessments to assist in formulating goals, or to conduct Process evaluations to assist in implementing a project. Experimental design should be used when it addresses the questions of interest and when it is practicable to use it; otherwise, some alternative framework should be chosen (Stufflebeam, 1971 [b]).

The literature presents a number of valuable alternatives to experimental design. Campbell and Stanley (1963) have discussed quasi-experimental design. O'Keefe (1968) has suggested a comprehensive methodology for field-based case studies. Scriven (1972 [b]) has introduced "Goal-Free Evaluation" and more recently "Modus Operandi Analysis" (Scriven, 1973). Reinhard (1972) has explained "Advocate Team Methodology," and Wolf (1974) has explicated the "advocacy-adversary" model. Thus, evaluators are not bound to use any single investigatory framework.

The meta-evaluator can perform a valuable service in helping an evaluator identify and assess alternative investigatory frameworks. To do this, the purposes (i.e., decision making or accountability) and the foci of the evaluation study (e.g., goals, design, process and/or results) need to be known. Also it is necessary to determine any feasibility constraints. Subsequently, the meta-evaluator can suggest and assess frameworks that are potentially responsive to the given conditions, and the evaluator can choose that framework that best optimizes the given conditions.

3. Instrumentation

Considering the purposes of the study, which of the available data gathering instruments and techniques are most appropriate? Moreover, are any of them adequate? Must instruments be especially developed to serve the purpose of this study? Is it feasible to develop new instruments? If it is, what sacrifices will have to be made regarding the technical adequacy of the instruments? These questions illustrate measurement problems commonly encountered in evaluation work.

The evaluator can, of course, get help from the literature in identifying potentially useful instruments. The Buros Mental Measurements Yearbooks (1949-1965) catalog and assess many instruments, especially in the cognitive domain. A recent book by Miles, Lake, and Earle (1973) identifies and discusses a number of instruments in the affective domain. Under the leadership of Gene Glass, the Laboratory of Educational Research at the University of Colorado maintains a set of fugitive instruments that have been developed and used in federal projects. Also one can identify many different instruments by checking through completed doctoral dissertations. So the evaluator can be greatly assisted in choosing instruments by surveying the relevant literature.

Even then, however, he may not find appropriate instruments. In this case he is often faced with a dilemma. Should he choose an inappropriate instrument that has been validated? Should he develop and use new instruments that respond directly to the purposes of the study, when there is no possibility of validating the instruments before they are used? The position in this paper is that the latter course of action often is the only feasible one. In any case, problems of instrumentation are key concerns in assessing evaluation designs.

4. Sampling

The fourth technical problem in evaluation studies concerns sampling. What's the population? Is an inference to be made to this population? How large a sample is needed? Can a random sample be drawn? Should the sample be stratified according to certain classification variables? Can the experimental units be randomly assigned to program

conditions? How much testing time can be expected from each sampled element? Is examines sampling necessary or would matrix sampling be better? Is matrix sampling feasible? If random selection and assignment of experimental units are not feasible, what can be done to guard against bias in the sample?

These questions denote a number of sampling-related difficulties that often are encountered in evaluation work. Even under the best of circumstances, inference to a population based on the performance of a random sample is logically not possible. As Campbell and Stanley (1963 p. 187) have pointed out, "generalization always turns out to involve extrapolations into a realm not represented in one's sample." Evaluators, however, are rarely able to even draw a random sample, so their problems of extrapolation are even worse. The least they can do is to consider and respond as best they can to questions such as those posed above.

5. Data Gathering

It is one thing to choose instruments and samples, but it is quite another actually to gather the data. Often the evaluator must rely on a number of persons in addition to himself for the gathering of data. For example, teachers must often be relied on to administer tests to students. This fifth technical problem of data gathering presents a number of difficulties to which the evaluator must be sensitive and responsive.

Who will deliver instruments to the data gathering sites? What is to prevent teachers from teaching to the tests? How can the cooperation of test administrators and respondents be secured? What can be done to insure motivation of the respondents and prevent cheating? In what settings will the respondents work? Who will administer the instruments? Who will monitor the data gathering sessions? How will standardization of data gathering conditions be assured? Unless evaluators consider and respond to such questions, their evaluations may fail due to poor implementation of the data-gathering plan.

6. Data Storage and Retrieval

The sixth technical problem concerns the storage and retrieval of data. Once the data have been gathered it is necessary to check them for accuracy, to code them properly, and to store them for future use. Meta-evaluators should check whether provisions have been made to accomplish these tasks. While the tasks are fairly routine, failure to deal effectively with them can destroy the effectiveness and efficiency of the evaluative effort.

7. Data Analysis

Both statistical and content analysis are involved in the seventh technical problem. The meta-evaluator should ascertain what plans have been made to analyze the data that will be obtained. He should check the plans for their appropriateness in responding to the study questions. He also should check whether assumptions required for the data analysis will be met by the data. Lastly, he should assess the provisions that have been made for performing the actual data analysis.

Many texts are available to assist in the analysis of data. Those prepared by Glass and Stanley (1970), Winer (1962), Guilford (1965), and Siegel (1956) are viewed in this paper as especially useful.

8. Reporting

The eighth problem concerns the preparation of evaluation reports. What different reports will be required for the different audiences? How will they be organized? What tables will they include? How long should they be? How will they be presented and interpreted to the audience?

This problem area is a reminder that evaluations must be informative. Doing an outstanding job of data collection and analysis will fall short of meeting the purposes of an evaluation if the results are not communicated effectively to the designated audiences. Therefore, metaevaluators should ascertain whether appropriate communication techniques will be used to interpret the findings to the prespecified audiences.

A common dilemma in reporting evaluation findings is that evaluators often have more data to present than their audiences are willing or able to receive. If evaluators make their reports very brief, their audiences are likely to judge the reports as cryptic and non-responsive to important questions. If, on the other hand, the evaluators present all their findings, their audiences likely won't read them and will be critical of their length.

One solution that the Dallas, Texas Independent School District's Office of Research and Evaluation is trying is to present three versions of each report. These include an abstract, an executive report, and a technical report. It seems important, so far, that these not be combined but presented as three distinct volumes. It further seems important that the shortest reports be provided to the audiences first, and that the others be provided only if they are requested.

Additional concerns are writing style and mode presentation. Technical language and jargon provide efficient communications to certain audiences but are uninterpretable to others. Also, oral reporting is sometimes more effective than printed communications. Also, reporting can be a continuing natural occurring function only if the evaluator maintains contact with his audience.

9. Summarizing the Technical Adequacy of the Design

The ninth and final problem involves summarizing the technical adequacy of the evaluation plan. Have the evaluation variables been identified and are they the right ones? Has a relevant and feasible investigatory framework been chosen? Has this framework been fleshed out in the form of appropriate instruments, sampling techniques, and analysis procedures? Have sufficient provisions been made for collecting, storing, retrieving, and reporting the information? Overall, will the evaluation yield results that are reliable, valid, objective, and useable?

If the evaluator can summarize his evaluation design through answering affirmatively to the above questions, he can be sure that his technical plan is sound. If he cannot, he should review and revise his technical plan. While technical problems are not the only problems that evaluators must address, they certainly are crucial ones.

Management Problems

So far, it has been noted that evaluation problems are conceptual, sociopolitical, contractual/legal, and technical in nature. The fifth area to be considered emphasizes that evaluators must cope with a number of crucial management problems. Specifically, ten such problems

will be introduced and discussed. It is to be noted that evaluators should not only deal with these problems, they should do so in such a way as to enhance the ability of the parent agency to improve its long-range capabilities to manage evaluation studies.

1. The Organizational Mechanism

The first management problem concerns the organizational mechanism for the evaluation. This is a matter of determining what organizational unit will be responsible for the evaluation.

Alternative possibilities exist. An in-house office of evaluation might be assigned to do the evaluation. An external evaluation group might be commissioned. A consortium of agencies might set up an evaluation center that they jointly support and this center might be assigned to do the work. The program staff, themselves, might perform a self-evaluation; or they might do it themselves but engage an external auditor periodically to assess their work.

Each of these approaches has been applied in evaluating educational programs, and each has differing costs and benefits. The meta-evaluator should identify the chosen alternative and compare its costs and benefits with those of alternative organizational arrangements.

2. Organizational Location of the Evaluation

The second management problem concerns where the evaluation is located within the organization. Will the evaluators report directly to the executive officer of the agency in which the program is housed? Will the evaluator also be able to report directly to staff members at lower levels of the system? Will he be enabled to communicate directly with members of the agency's policy groups? In general, through what channels may the evaluator influence policy formulation and administrative decision-making?

This is a crucial issue that affects particularly the pervasiveness, credibility, and timeliness of evaluation work. If reports are submitted only through the chief executive officer, other members of the system may doubt the credibility of the reports. On the other hand, if reports are sent directly to persons at all levels of the system, the chief decision makers may feel greatly threatened by the evaluation, especially if the evaluator interacts directly with members of the agency's policy board. Moreover, if reports must pass through the chief executive's office, the reports may fail to meet criteria of pervasiveness and timeliness. An illustration of this is when individual student diagnostic records are sent by a testing company to the central administration of a school district and only weeks later reach the teacher who could make constructive use of the results. Clearly, the matters of organizational location and reporting channels are crucial concerns in any evaluation study.

3. Policies and Procedures

A third management concern is that of policies and procedures which govern evaluation activities. The evaluator needs to find out about existing policies and procedures that will affect or govern his work. Also, he should be alert to opportunities that he might use to help the agency that commissioned the evaluation to develop and adopt policies and procedures to govern its future evaluation work.

Such policies and procedures might include a number of items. Delineation of evaluation roles and assignment of responsibilities for those roles are fundamental concerns. A conceptual scheme to guide the agency's evaluation work might also be provided, as was done in Michigan through mandating a six-step accountability model. Of course, such a statement of policies and procedures should specify how the evaluation work is to be financed. Examples of formal manuals of evaluation policies and procedures are those adopted by the Saginaw, Michigan Public Schools (Adams, 1970) and the Ohio State University College of Education (Assessment Council, 1970).

4. Staffing Problems

The fourth administrative problem concerns the staffing of the evaluation work. Who will have overall responsibility for the work? Who will be assigned the operational responsibility? What other roles are to be manned? Who will be assigned to these roles? What recruitment of personnel must be done? Who will be considered? What criteria will be used to assess their qualifications? Who will choose them? Quite obviously, evaluations are often team efforts and it is crucial to choose qualified personnel to perform the evaluations.

Beyond meeting the immediate evaluation requirements, the staffing of an evaluation sometimes provides significant opportunities for upgrading the long-range evaluation capability of the agency whose program is being assessed. Evaluation projects are excellent settings within which to train evaluators. If persons are recruited partially because they want to become evaluators in the agency whose program is being evaluated, they can be trained through their immediate evaluation assignment and subsequently be kept on by the agency as evaluators. Illustrations of this are that Dr. Jerry Walker (who heads evaluation in the Ohio State University Center for Vocational and Technical Education), Dr. Howard Merriman (a prominent evaluator in the Columbus, Ohio Public Schools), and Mr. Jerry Baker (Director of Evaluation in the Saginaw, Michigan Public Schools) were recruited, trained, and later employed on a continuing basis exactly in this way.

Staffing is obviously a key problem in the management of evaluation work. The quality of the evaluation will largely depend on the competence and motivation of the staff. At the same time there is often an opportunity to upgrade an agency's long-term evaluation capability through judicious recruitment and training of persons who may want to stay on in the agency in the capacity of evaluator after the initial evaluation assignment has been completed. The meta-evaluator should carefully assess the evaluator's provisions for meeting his staffing needs and serving opportunities for longer-range staffing payoff.

5. Facilities

The fifth management problem in evaluation concerns the facilities needed to support the evaluation. What office space, equipment, and materials will be needed to support the evaluation? What will be available? Answers to these questions can affect the ease with which evaluations are carried out and even their success. Thus, evaluators should be sure that their management plans are complete in their provisions of the necessary facilities.

6. Data Gathering Schedule

The sixth management problem to be identified involves the scheduling of data collection activities. What samples of persons are to respond to what instruments? When are they to respond? Is this schedule reasonable, and is it acceptable to the respondents? When will the instruments and administration arrangements need to be finalized? Will the instruments be ready when they are needed? Will students still be in school when the administrations are to occur? Are there any potentially disastrous conflicts between the data gathering schedule and other events in the program to be evaluated? Overall, is the data-gathering schedule complete and feasible?

The above questions illustrate difficulties that do plague evaluation studies. In one case, a government-sponsored $250,000 evaluation study of programs for disadvantaged students actually was scheduled so that student data had to be gathered in July and August. The evaluators, who were from outside the field of education, had forgotten that most students do not attend school in the summer. In another situation an evaluator planned to administer ten different instruments to the same group of principals during a three-week period. While it is important in many studies to ascertain the school principals' perceptions, bombarding them with questionnaires will neither elicit good will nor cooperation. As a final example, an evaluator scheduled observations of teachers during a week when they were administering state tests. This would have been fine if the purpose of the study had been to determine teacher competence in test administration, but it certainly was a poor time to assess their use of a new curriculum. These examples argue that meta-evaluators should pay attention to the appropriateness of the data-gathering schedule.

7. Reporting Schedule

The seventh management problem also concerns scheduling, but in this case the scheduling of reports. What reports will be provided, to what audiences, according to what schedule? Meta-evaluators should check reporting schedules for their completeness in these respects and for their potential for communicating effectively to the prespecified audiences. Also it is important that such schedules be checked for their feasibility. The scheduling of reports bears directly on how useful the reports will be to the designated audiences.

8. Training

The eighth management problem in evaluation concerns training. As mentioned previously in this paper, evaluation is largely a team activity, and the evaluation team must often depend on the cooperation of system personnel in conducting the evaluation. If the various persons, including evaluators, teachers, and administrators, are to perform their roles effectively, they often need special evaluation training. Hence, evaluators should be prepared to meet such training requirements.

In most situations, the training should be both general and specific. The specific training is needed for the performance of specific evaluation tasks, e.g., the administration of a particular test or interview, or the coding of a particular set of data. However, it is also desirable to give training in the general principles of evaluation. Such training assists persons

to understand their particular roles; it provides them with general guidelines for making specific decisions in the course of implementing their role; and it improves their overall ability to perform future evaluations. Thus, training activities within evaluation studies should prepare persons to perform their particular assignments, but it should also present them with opportunities for upgrading their general understanding of evaluation.

A variety of approaches to training within evaluation studies can be applied. Blaine Worthen, Director of Evaluation in the Northwest Regional Educational Laboratory, runs periodic sack lunch seminars that focus on topics selected by his staff. The Columbus Schools Department of Evaluation at one time supported two full-time persons whose primary assignment was to continually provide consultation to existing evaluation staff and in-service training in evaluation for administrators, teachers, and new evaluation staff members. Several agencies have engaged external review panels to study their evaluation operations and provide training based on the analyses. The Western Michigan University Evaluation Center periodically invites evaluators to present their work to the Center staff, whereupon the Center's staff members critique the work. (This is especially good because both parties gain from the exchange of information and discussion and neither charges the other.) Also, NIE, USOE, and AERA have sponsored the development of a large number of evaluation training packages. Thus, different means can be found to conduct needed training within evaluation studies.

The content for such training can be highly variable. Considerations in determining what training should be provided include who will be trained, what their assignments are, what they want and need to know, how they will use evaluation in the future, and what opportunities exist for providing the training. A good source of information about the content for evaluation training programs is a doctoral dissertation by Darrell K. Root (1971) on the topic of differential evaluation training needs of administrators and evaluators.

Overall, training is a key area in evaluation work. It is potentially very cost/effective since it enhances the ability of persons to implement their specific evaluation assignments; and it uses training opportunities to prepare these same persons for future evaluation work.

9. Installation of Evaluation

The ninth management concern in evaluation is more an opportunity than a problem. This concerns the matter of using specific evaluations as a means of installing systematic evaluation in a system. The position in this paper is that evaluators should be alert to such opportunities and capitalize on them whenever possible. In this way, evaluators can aid the systems that house the programs being evaluated to increase their capacities to evaluate their own activities.

This is a crucial need in education. There never will be sufficient evaluation companies to perform all the needed evaluation. In any case much of the needed evaluation should not be done by external agents since they are sometimes too threatening and too expensive. But, as Adams (1970) discovered when he surveyed all the school districts in Michigan, few educational agencies have their own evaluation capabilities. Thus, there is a need to aid educational agencies

to develop their own systems of evaluation.

A standard practice of the Ohio State University Evaluation Center was to use evaluation service contracts as a means of assisting agencies to develop their own evaluation systems. Notable examples are evaluation projects performed for the Columbus, Ohio, and Saginaw, Michigan Public Schools. In both cases the school districts had no evaluation capability, had encountered requirements to evaluate their federally supported projects, and engaged the Ohio State University Evaluation Center to conduct the evaluations. That Center contracted both to con-duct the needed evaluations and to develop evaluation departments for the school districts.

Both purposes were served through a common approach. The evaluation effort was staffed with teachers from the two school districts who declared interests in becoming system evaluators and who gave promise of becoming good evaluators. These teachers were enrolled in graduate programs in evaluation and were provided field-based training in evaluation. Of course, that training revolved around the work assignments in the evaluation projects. At the end of the evaluation projects, the Columbus and Saginaw personnel, now with graduate training and degrees in evaluation, returned to their school system to man new departments of evaluation.

The continued operation and the achievements of both departments attest to the power of this approach. The Saginaw, Michigan Department of Evaluation has been rated by the Michigan Department of Education as a model evaluation system. The School Profile (Merriman, 1969) developed by the Columbus Schools Department of Evaluation has been adopted nationally by a number of school districts. Of course, the achievements of Dr. Howard Merriman (present Vice-President of the American Educational Research Association's evaluation division), who was one of the Columbus teachers chosen to work on the Ohio State contract with Columbus, dramatically illustrates that school districts may have potentially outstanding evaluators in their own teaching and administrative ranks.

The position in this paper is that special evaluation projects should be viewed as potential opportunities for upgrading an agency's evaluation capability. Meta-evaluators should ascertain whether evaluation staffs have sought out and responded to such opportunities.

10. Budget for the Evaluation

The tenth and final management item involves the budget. Is there one? Does it reflect the evaluation design? Is it adequate? Does it have sufficient flexibility? Will it be monitored appropriately? While these are obvious questions, it is surprising how often grandiose evaluation plans are not accompanied by supporting budgets. It has become the habit of the author, when evaluating evaluation plans, to first review the budget for evaluation. If none exists, it matters little how good the technical plan is, for it will not be possible to implement it. If a budget does exist, it clearly needs to be checked for its sufficiency.

This concludes my discussion of management problems in evaluation. Hopefully the ten management items that were discussed will prove useful to evaluators as they review their plans for managing evaluation activities. The position in this paper has been that evaluation efforts should be managed

both to achieve specific evaluation objectives as efficiently as possible, and to help the agencies involved in the evaluation to upgrade their internal evaluation capabilities.

Moral, Ethical, and Utility Considerations

The final class of evaluation problems involves moral, ethical and utility questions. Evaluations are not merely technical activities; they are performed to serve some socially valuable purpose. Determining the purpose to be served inevitably raises questions about what values should be reflected in the evaluation. Deciding on value bases also poses ethical conflicts for the evaluator. Further, as emphasized before in this paper, the evaluator must be concerned with what practical uses his reports will serve. This final set of problems includes six issues that the evaluator must face in regard to moral, ethical, and utility matters.

1. Philosophical Stance

The first issue concerns what philosophical stance will be assumed by the evaluator. Will the evaluation be value-free, value-based, or value-plural? Each of these positions has its advocates.

Some say that evaluators should merely provide data, without regard for the values of any particular group, such as consumers or producers. Persons who take this position are committed to a value-free social science. Their position is that evaluators should be objective and should not adopt any particular value position as a basis for their work. A consequence of this position is that evaluators provide data, but not recommendations. A difficulty of this approach is in determining what data to collect since there is no particular value framework from which to deduce criteria. Selection of values for interpreting the findings is left to the audiences for the reports. Overall, the value-free option emphasizes the objectivity and neutrality of the evaluation but provides no guidance for choosing variables or interpreting results.

A second option is a value-based position. Here the evaluator chooses some value position and through his work attempts to maximize the good that can be done as defined by this position. The value-based evaluator may decide that his evaluation should optimize the Protestant Ethic, equal opportunity for persons of all races, Marxism, or principles of Democracy—to name a few possibilities. Once he has chosen a value base, the appropriate variables that might be assessed and the rules for interpreting observations on those variables are theoretically determined. The value-based evaluator is neither neutral nor objective concerning what purposes his evaluation should serve. His evaluation can be viewed (and critiqued) in terms of its social mission.

A third philosophical stance might be termed a value-neutral position. According to this position, evaluators remain neutral concerning the selection of a particular value position, but they explicitly search for and use conflicting value positions in their collection and interpretation of data. Thus, they can show the consequences of a particular action in relation to the different value positions that might be served by the action.

An example of this third philosophical stance occurred when a team of evaluators was commissioned to identify and assess

alternative ways of educating migrant children. The evaluators identified value positions advocated by experts in migrant education and by the migrants themselves. The experts said the chosen alternative should be the one that gave best promise of developing reading and arithmetic skills. However, the migrants urged that the chosen alternative should be the one that would best help their children to be socialized into society. These positions represented, for the evaluators, conflicting value positions that might be used to search for and assess alternative instructional strategies.

Using either position by itself would produce a biased set of strategies, but using both would increase the range of strategies. Using criteria from both philosophical positions would produce different evaluations of each identified strategy.

As an example, two alternatives (among others) were identified. One was to operate a resident school in the desert, the other to totally integrate the migrant children into regular classrooms. The former strategy rated high in meeting criteria of improved reading and arithmetic performance, but was a disaster in relation to the socialization objective. The opposite was true for the approach involving total integration. Based on their respect for both conflicting value positions, the evaluators identified additional alternative strategies that represented a compromise position. This case illustrates that an evaluator's philosophical stance can drastically influence his evaluation outcomes.

2. Evaluator's Values

The second problem concerns the evaluator's values. Will his values and his technical standards conflict with the client system's values? Will the evaluator face any conflict-of-interest problems? What will be done about possible conflicts? Evaluators are often faced with questions like these and should deal openly and directly with them.

An example of a conflict between an evaluator's technical standards and the client system's values occurred in an evaluation of a "free school." The evaluator believed that it was essential to administer achievement tests to the school's students. The "free school" administrators said that the "free school" philosophy does not permit the testing of students. While this was an extreme case, it illustrates problems that evaluators may encounter in performing what they consider to be necessary evaluation tasks.

The evaluator can also encounter conflicts of interest. Being on the payroll of the agency whose work is being evaluated insures that potential conflicts of interest will emerge. The evaluator, being committed to the success of the agency—or at least to preserving his job—may find it difficult to report negative results. This is also the case when the evaluator has, at some previous time, served as a consultant to the agency. It is good ethical practice for evaluators to identify and report their potential conflicts of interest, to guard against their influence on their work, and, if necessary, to withdraw from the evaluation assignment.

3. Judgments

Another issue the evaluator must face is whether his reports should present judgments (or merely descriptions) of what has been observed. Will the evaluator report no judgments? Will he report his own; or will he obtain, analyze, and report the judgments of various

reference groups? The evaluator's responses to these questions will pretty well determine his role in decision making in the activity being observed.

If the evaluator decides to present no judgments, he will leave decision making completely up to his client. If the evaluator presents his own judgments, he likely will have a strong influence on decision making. If he presents judgments of various reference groups, he will not have decision making power himself, but will help the chosen reference groups to exercise such power. The point here is that the evaluator has options concerning how he should treat the matter of judgment in his evaluation and he should weigh the consequences of each option against his particular philosophical stance on evaluation.

4. Objectivity

The fourth problem is that of objectivity. As an evaluator collects and reports data during the course of a program, how can he keep his independence? If the program personnel adopt his recommendations, how can the evaluator any longer be neutral about the merit of related actions? Likewise, how can the evaluator avoid being co-opted by program personnel who win his confidence and support his ego needs?

Tom Hastings once told me that "objectivity is a matter of intelligence and integrity." I interpret this to mean that evaluators should know whether they have lost their independent perspective and that if they have they should ask that they be replaced in the evaluation job.

5. Prospects for Utility

A fifth concern in this section is whether the evaluation is merely an academic exercise or has real prospects for utility. The criteria for relevance, scope, importance, credibility, timeliness, and pervasiveness have been mentioned before. It is reiterated here that the evaluator should seriously assess and report on the prospects that his evaluation plan has for being useful.

6. Cost/Effectiveness

Finally, the evaluator should assess the cost/effectiveness of his plan. Compared to its potential payoff, will the evaluation be carried out at a reasonable cost? Is the potential payoff worth what it will cost? Might it save the system more money than the cost of the evaluation?

This completes the discussion of evaluation problems. While technical matters are a key problem area for the evaluators, he must solve many other kinds of problems. These include conceptual, sociopolitical, contractual/legal, management, and moral/ethical problems. All such problems must be anticipated and avoided if evaluations are to be technically sound, useful, and cost/effective. Consequently, evaluators need a technology by which continually to assess their evaluative plans and activities. We consider what form such a technology might have in the next part of this paper.

II. A Conceptualization of Meta-Evaluation

Included in this second part of the paper are a definition of meta-evaluation, premises for a conceptualization of meta-evaluation, and a logical structure for designing meta-evaluation activities. Taken together these are suggested as a conceptualization of meta-evaluation.

Meta-Evaluation Defined

In the introduction, meta-evaluation was defined as the evaluation of evaluation. More specifically, it is defined in this paper as a procedure for describing an evaluation activity and judging it against a set of ideas concerning what constitutes good evaluation.

This means that meta-evaluation is higher-order and includes evaluations that are secondary, tertiary, etc. This presents a practical dilemma, since meta-evaluation involves infinite regression (Scriven, in press), and since it is not practical to act on the unlimited possibilities of evaluating evaluations of evaluations of evaluations. While infinite regression is a fundamental part of the metaevaluation, the examples in this paper mainly deal with second-order evaluations, or meta-evaluations that are once-removed from the primary evaluations. It is assumed that second-order meta-evaluations are feasible, important, and sufficient in most practical situations. In any case, the principles and procedures for second-order evaluation should apply to other levels of meta-evaluation.

Premises

Since meta-evaluation is a form of evaluation, the conceptualization of meta-evaluation must be consistent with some conceptualization of evaluation. The conceptualization used in this paper has eight premises. Essentially these are the author's responses to the eight questions in conceptualizing evaluation that were discussed in the first part of this paper. These premises are listed and related to the concept of meta-evaluation below.

1. Evaluation is the assessment of merit; thus, meta-evaluation means assessing the merit of evaluation efforts.
2. Evaluation serves decision making and accountability; thus, meta-evaluation should provide information pro-actively to support the decisions that must be made in conducting evaluation work, and meta-evaluation should provide retroactive information to help evaluators be accountable for their past evaluation work. Another way of saying this is that meta-evaluation should be both formative and summative.
3. Evaluations should assess goals, designs, implementation, and results; thus, meta-evaluation should assess the importance of evaluation objectives, the appropriateness of evaluation designs, the adequacy of implementation of the designs, and the quality and importance of evaluation results.
4. Evaluation should provide descriptive and judgmental information and appropriate recommendations. Likewise, meta-evaluation should describe and judge evaluation work and should recommend how the evaluations can be improved and how the findings can appropriately be used.
5. Evaluation should serve all persons who are involved in and affected by the program being evaluated; hence, meta-evaluation should serve evaluators and all the persons who are interested in their work.
6. Evaluation should be conducted by both insiders and outsiders; generally (but not always) insiders should conduct formative

evaluation for decision-making, and outsiders should conduct summative evaluation for accountability. Hence, evaluators should conduct formative meta-evaluation and they should obtain external judgments of the overall merit of their completed evaluation activities.

7. Evaluation involves the process of delineating the questions to be addressed, obtaining the needed information, and using the information in decision-making and accountability. Hence, meta-evaluators must implement three steps. The meta-evaluators must delineate the specific meta-evaluation questions to be addressed. They must collect, organize, and analyze the needed information. Ultimately, they must apply the obtained information to the appropriate decision-making and accountability tasks.
8. Evaluation must be technically adequate, useful, and cost/ effective, and meta-evaluation must satisfy the same criteria.

A Logical Structure for Meta-Evaluation

These eight premises have been used to generate the meta-evaluation structure that appears in Figure 4. This structure portrays meta-evaluation as a methodology for assessing the merit of proposed and completed evaluation efforts (the first premise). The framework has three dimensions, they relate to the purposes, objects, and steps (the second, third, and seventh premises) of meta-evaluation studies. The contents of the cells of the structure reflect that evaluation work should meet the criteria of technical adequacy, utility and cost/effectiveness (the eighth premise). The structure reaffirms that insiders should conduct proactive meta-evaluation and that external agents should conduct retroactive meta-evaluation work (the sixth premise). It is an implicit assumption of the structure that meta-evaluation findings should provide descriptions, judgments, and recommendations (the fourth premise) to the evaluators whose work is being judged and all persons who are interested in their work (the fifth premise). Overall, this structure is presented as a guide for designing meta-evaluation activities.

Given this overview of the structure, each of its dimensions will next be considered. Then the interaction of the three dimensions will be discussed.

Purpose of the Meta-Evaluation	Steps in the Meta-Evaluation Process	Objects of the Meta-Evaluation			
		Evaluation Goals	Evaluation Design	Evaluation Processes	Evaluation Results
Pro-active Meta-Evaluation to serve Decision Making in eval. work (This is Formative Meta-Evaluation and usually is conducted by insiders)	Delineating the information requirements	Audiences Possible eval. goals Criteria for rating eval. goals	Alternative eval. Design Criteria for rating eval. designs	Work breakdown and schedule for the chosen eval. design Admin. checklist for reviewing eval. Designs	The eval. objectives Cost, quality, and impact criteria Intended users of the evaluation
	Obtaining the needed information	Logical analyses of the eval. goals Ratings of the eval. goals	Ratings of the alternative designs	Review of the eval. design Monitoring of the eval. process.	Ratings of the quality of reports Evidence of use of eval. for decision making & accountability Ratings of the value of eval. reports Monitoring of expenditures for eval.
	Applying the obtained information	Recommendations of what eval. goals should be chosen	Recommendations of what eval. design should be chosen	Periodic progress & Exception reports Recommendations for modifying the eval. design or procedure	Periodic reports of the quality, impact, & cost/effectiveness of the eval. Recommendations for improving eval. results
Retroactive Meta-Evaluation to serve Accountability in eval. work (This is Summative Meta-Evaluation and usually is conducted by outsiders)	Delineating the information requirements	Audiences Goals chosen Criteria for judging eval. goals	The chosen design The critical competitors Criteria for rating eval. designs	Work breakdown & schedule for the chosen eval. design Admin. checklist for reviewing eval. designs	The eval. objectives Cost, quality & impact criteria Intended users of the evaluation
	Obtaining the needed information	Survey of evaluation needs Audience ratings of chosen eval. goals Analysis of eval. goals related to criteria, needs, & audience ratings	Ratings of the alternative eval. designs	Case study of the eval. process Analysis of discrepancies between the eval. process & the chosen design	Ratings of the quality of reports Evidence of use of eval. for decision making & account-ability Ratings of the value of eval. reports Cost analysis for the evaluation
	Applying the obtained information	Judgment of the chosen eval. goals	Judgment of the chosen eval. design	Judgment of the implementation of the eval. design	Judgment of the quality, utility, and cost/effectiveness of the eval. activity

Figure 4
Logical Structure for Meta-Evaluation

Purposes of Meta-Evaluation

The first dimension of the matrix indicates that meta-evaluation should serve decision-making and accountability.

Supporting decision making in evaluation requires that meta-evaluation be done proactively to provide timely recommendations concerning how evaluation studies should be designed and conducted. Meta-evaluation that serves decision making may be termed formative meta-evaluation. As noted in Figure 4, formative meta-evaluation usually is conducted by insiders, i.e., those who do the evaluation that is being guided by the meta-evaluation. Conducting formative meta-evaluation is proposed as a direct way of insuring that evaluations will produce results that are technically adequate, useful, and cost/effective.

The second purpose of meta-evaluation is to serve the evaluator's need to be accountable for his work. This purpose requires that meta-evaluation be conducted retroactively to produce public judgments of the merits of the completed evaluation work. Meta-evaluation that serves accountability is synonymous with summative meta-evaluation. A careful examination of the framework reveals that much of the information required in summative meta-evaluation is potentially available from formative meta-evaluation. Thus, formative meta-evaluation potentially can provide a preliminary data base for summative metaevaluation. However, to insure the credibility of the results, metaevaluation for accountability should usually be conducted by outsiders.

Steps in the Meta-Evaluation Process

The second dimension of Figure 4 indicates there are three basic steps in conducting meta-evaluation studies, whether in the decision-making or accountability modes. These steps are delineating the information requirements, obtaining the needed information, and applying the obtained information to achieve decision-making and accountability purposes. Thus, methods for meta-evaluation should assist in determining questions, in gathering and analyzing the needed information, and in using the information to answer the meta-evaluation questions.

Objects of Meta-Evaluation

The third dimension of the structure denotes four objects of meta-evaluation. They are evaluation goals, evaluation designs, evaluation processes, and evaluation results.

Evaluation goals pertain to the ends to be achieved by the evaluation. What audiences are to be served? What are their questions? What information do they want? What information will be provided to them? How is the evaluative feedback supposed to influence the actions of the audience? These questions illustrate considerations in the formulation and assessment of alternative evaluation goals.

Basically, an evaluation goal is an intent to answer certain questions, to enlighten some audience, and to influence their actions in the direction of rationality. There are obviously alternative possible goals for any evaluative effort, hence it is important to identify and assess the competing evaluation goals.

The second object of meta-evaluation concerns evaluation designs. Obviously, there are alternatives. The choice of the appropriate design depends on what evaluation goals have been chosen and a variety of practical and sociopolitical considerations. Hence, it is important in evaluation work to identify and assess alternative evaluation designs.

The third object of meta-evaluation involves evaluation processes. It is one thing to choose a potentially strong evaluation design. It is quite another to carry it out. As discussed in Part I of this paper, a variety of practical problems can invalidate the strongest of theoretical evaluation designs. Hence, it is important to identify potential implementation problems in relation to chosen evaluation designs and to assess their impact on the evaluation results.

The fourth object of meta-evaluation concerns evaluation results. Were the study questions answered? How well? Were the findings communicated to the designated audiences? Did they understand the findings? Did they apply them? Were their applications defensible given the evaluation results? These questions illustrate the considerations in evaluating evaluation results.

Interaction of the Three Dimensions

Given these descriptions of the three dimensions of Figure 4, it is appropriate to consider their interactions. Basically, Figure 4 identifies and characterizes two major classes of meta-evaluation designs. The Proactive or Formative Meta-Evaluation Designs, and the Retroactive or Summative Meta-Evaluation Designs. Each of these classes of designs is further divided into four specific types of meta-evaluation designs. These pertain to the assessment of evaluation goals, of evaluation designs, of evaluation processes, and of evaluation results. Each type of meta-evaluation design is further defined by the delineating, obtaining, and providing tasks. Thus, Figure 4 identifies four types of proactive and four types of retroactive meta-evaluation designs. Within the figure, each design type is defined by the steps in the evaluation process.

It is to be noted that the proactive meta-evaluation designs all result in recommendations, while the retroactive meta-evaluation designs all result in judgments. Proactive meta-evaluation studies assist in choosing evaluation goals, choosing evaluation designs, carrying out chosen evaluation designs, and attaining desirable evaluation results and impacts. Retroactive evaluation results provide assessments of the merits of completed evaluation activities.

In practice, the four types of proactive meta-evaluation studies are usually conducted separately, as they relate to specific decision points in the evaluation process. However, the retroactive meta-evaluations are often combined into a single summative case study since they pertain to completed and interrelated sets of evaluation activities.

The object of the conceptualization in this second part of the paper is to assist evaluators to identify and ameliorate the problems and serve the meta-evaluation criteria identified in Part I of this paper. Next we consider how the conceptualization presented in Part II can be applied in practice.

III. Use of the Conceptualization of Meta-Evaluation

This third and final part of the paper is intended to provide practical guidelines and examples for conducting meta-evaluations. Specifically, the structure introduced in Part II has been used to generate and describe five meta-evaluation designs. Examples of real-world activities that match the designs are also presented.

Figure 5 summarizes the designs to be discussed within the logical structure for meta-evaluation that was presented in Part II. There are four proactive designs (1-4) that assist evaluators, respectively to determine evaluation goals, choose evaluation designs, carry them out, and use them to produce valuable results and impacts. Design 5 provides summative assessments of the overall worth of past evaluation efforts.

Design #l—for Pro-active Assessment of Evaluation Goals

Design #1 pertains to pro-active meta-evaluation studies that identify and rank alternative evaluation goals.

In delineating such studies, it is necessary to identify the audiences for the primary evaluation, to identify a range of possible evaluation goals, and to identify criteria for rating the goals. The audiences are those persons to be affected by the evaluation study that is the subject of the meta-evaluation. The alternative goals are the alternative reasons that members of the audience and the evaluation team have for conducting the study. Such reasons may be for decision making and/or accountability, and they may refer to specific questions about program goals, designs, processes, and results. The criteria for assessing evaluation goals include such variables as scope, importance, tractability, and clarity. Overall, the delineating activities for Design #1 should clarify audiences for the primary evaluation, alter-native evaluation goals, and criteria for rating the evaluation goals.

Steps for obtaining the information required by Design #1 include logical analysis and ratings of the alternative evaluation goals. The logical analysis can be done by the primary evaluators or by specially commissioned meta-evaluators. Their analyses should define each goal in terms of at least the following questions:

1. Who is to be served by the goal?
2. What question will be answered?
3. Why does the audience want to know that?
4. What action will likely be guided through achieving this evaluation goal?

Purpose of the Meta-Evaluation	Steps in the Meta-Evaluation Process	Objects of the Meta-Evaluation			
		Evaluation Goals	Evaluation Design	Evaluation Processes	Evaluation Results
Pro-active Meta-Evaluation to serve Decision making in evaluation work	Delineating the information requirements Obtaining the needed information Applying the obtained information	Design #1 for Pro-active Assessment of Evaluation Goals	Design #2 for Pro-active Assessment of Evaluation Design	Design #3 for Pro-active Assessment of the Implementation of a Chosen Design	Design #4 for Pro-active Assessment that Enhance the Quality and use of Evaluation Results
Retroactive Meta-Evaluation to serve Accountability in evaluation work	Delineating the information requirements Obtaining the needed information Applying the obtained information	Design #5 for Overall Retroactive Assessment of the Merit of a Total Evaluation Effort			

Figure 5
Five Meta-Evaluation Designs

One way of analyzing the alternative goals is through a matrix with labels for alternative evaluation goals as its row headings and the above questions as its column headings. Figure 6 illustrates the use of such a matrix.

Once the alternative evaluation goals have been analyzed, it is necessary to rank them. This is a matter of getting representatives of the primary evaluation team and of their audiences to rate the goals on each selected criterion (e.g., for clarity, scope, importance and tractability). A systematic way of doing this is through use of the Delphi technique (Cyphert and Cant, 1970).

After the alternative evaluation goals have been identified and rated, recommendations should be formulated concerning what evaluation goals should be adopted. Ultimately, the primary evaluators and their clients must choose the objectives that will serve as the basis for their evaluation study.

A study that was conducted for the Bureau of the Handicapped in the U. S. Office of Education illustrates the use of Design #1. This study was directed by Dr. Robert Hammond. The charge was to identify and rate alternative goals for evaluating programs for the educationally handicapped.

Hammond commissioned experts in evaluation and in education for the handicapped to write two position papers; one concerned what alternative evaluation goals should be considered, the other suggested criteria for use in rating the evaluation goals.

These papers were used as the basis for a national conference to identify and rate goals for national and state efforts to evaluate programs for the handicapped. About forty people were invited to attend this working conference. These persons were selected to be representative of work in the different areas of handicapped; of local, state, and national levels of education; of educational evaluation, and of different areas of the country.

Questions	Who is to be served?	What question will be answered?	Why does the audience want to know?	What action will be guided?
Goals: Assess and rate students' health needs	Administrators Health Personnel Funding agency	Are Students receiving dental, medical, nutritional, and recreational services	Federal funds are potentially available for upgrading health services	Securing and allocating of funds, and design of health services to serve the most critical health service needs
Identify and assess alternative plans for health services	Administrators Health personnel	What viable alternative health service plans exist? How good are they?	A proposal for federal funds is to be written	Choice of a particular health service plan
Determine whether a new curriculum is being implemented	Assistant superintendent	Are the procedures being implemented?	There are conflicting reports concerning whether the project is on track	Possible administrative actions to insure that the curriculum will be implemented
Determine whether a new arithmetic program has negative side effects	Board of education Administration PTA	What are the effects of the new program on students' performance in reading, interest in math, and attitude toward school	Concerns have surfaced that the program, while achieving its objectives, is having bad effects in other areas	Continuance, modification, or termination of the program

Figure 6
Matrix for Displaying and Analyzing Evaluation Goals

The conference lasted five days. The first day was devoted to reviewing and discussing the working papers and especially to choosing criteria for rating evaluation goals. The second, third, and fourth days were used in conducting three rounds of a Delphi study. Its purpose was to have the group expand the alternative evaluation goals, rate them on the selected criteria, and achieve a consensus concerning what evaluation goals should be recommended to the Bureau of the Handicapped. The final day was devoted to preparing the final report for the U. S. Office of Education.

To Dr. Hammond's credit and round-the-clock shifts of clerical personnel, the final report was distributed in final form during the last day of the conference. This fact plus the fact that the report reflected thoughtful working papers on evaluation goals and criteria and three rounds of a Delphi study is evidence that Design #1 can be employed to assess alternative goals for projected evaluation studies.

Design #2—for Pro-active Assessment of Evaluation Designs

Design #2 pertains to pro-active meta-evaluation efforts that identify and rank alternative evaluation designs.

In delineating such studies, one identifies alternative evaluation designs and criteria for rating the designs. Identifying evaluation designs starts with a survey of existing designs in the literature. If such a survey fails to turn up appropriate designs, it is necessary to invent new ones. Formulation of the designs includes matters of sampling, instrumentation, treatments, and data analysis. Standard criteria for rating evaluation designs include technical adequacy (internal validity, external validity, reliability, and objectivity), utility (relevance, importance, scope,-credibility, pervasiveness, and timeliness), and the prudential criterion of cost/effectiveness.

After the alternative evaluation designs and the criteria for rating them have been determined, it is necessary to apply the criteria to the designs. Campbell and Stanley's (1963) standardized ratings of experimental designs are useful in this area. The Buros Mental Measurement Yearbooks (1949-1965) are also useful for identifying and assessing published tests that might be a part of the designs. Finally, the alternative evaluation designs under consideration need to be ranked for their overall merits.

The description and judgment of alternative evaluation designs leads to a recommendation concerning what evaluation design should be chosen. This recommendation should be based on documentation of the meta-evaluation study. The documentation should include a reference to the selected evaluation goals, a description of the alternative designs that were considered, a listing of the criteria that were used to compare the designs, and a summary of the ratings of the designs. Finally, the recommended design should be justified in view of the available evidence.

An instance of Design #2 occurred when the National Institute of Education sought to adopt a design for evaluating regional laboratories and research and development centers. To achieve this purpose, NIE contracted with the Ohio State University Evaluation Center for the development and assessment of alternative evaluation designs (Stufflebeam, 1971 -c-).

The Center engaged two teams of evaluation specialists to generate alternative evaluation designs. These specialists were presented with an NIE policy statement (Fry, 1971) concerning what decisions should be served by the evaluation. The teams were oriented to the nature of activities in labs and centers. The teams were given criteria that they should meet in the development of their evaluation designs.

The teams then generated competing evaluation systems. Their reports (Scriven et al, 1971; Stufflebeam et al, 1971 -d-) were sent to lab and center personnel who rated the two designs. A panel of four experts was also engaged to evaluate the two designs. A hearing was held in Washington to obtain further input concerning the designs

Finally, the NIE staff reviewed the available information and chose one of the designs. Overall, the implementation of this meta-evaluation was conducted during two months and under a budget of $21,000.

A similar application of this design occurred when the Ohio State Department of Education engaged the Ohio State University Evaluation Center to identify

and assess alternative designs for a new state educational accountability system. The Center commissioned three teams to generate alternative accountability plans (Guba et al, 1972; Jaeger et al, 1972; Nystrand et al, 1972). A fourth team assessed the merits of the three plans (Worthen et al, 1972) and provided recommendations to the Ohio Department of Education.

Design #3—Pro-active Assessment of the Implementation of a Chosen Evaluation Design

Design #3 pertains to pro-active meta-evaluation studies to guide the implementation of a given evaluation design.

The delineating tasks in relation to Design #3 are extensive. Based on the results of a type #2 meta-evaluation, an evaluation design has been chosen. There are many administrative and technical decisions to be made in operationalizing the chosen design. The operational characteristics of the chosen evaluation design need to be explicated, and potential problems in the implementation of the design need to be projected. These characteristics and potential problems server as foci for periodic checks on how well the chosen evaluation design is being implemented.

A number of techniques are available for delineating the operational characteristics of evaluation designs. These techniques include "Work Breakdown Structure," Critical Path Analysis, and Program Evaluation and Review Technique (Cook, 1966). An additional technique called an Administrative Checklist for Reviewing Evaluation Designs is being introduced here. The Checklist appears as Exhibit 1. It reflects the problems that were described in Part I of this paper and is suggested for use in reviewing evaluative activity. These techniques are intended for use in delineating the operational characteristics, decision points, and potential problems that relate to the implementation of a given evaluation design.

Exhibit 1
An Administrative Checklist
for Reviewing Evaluation Plans

Conceptualization of the Evaluation

___	Definition	___	How is evaluation defined in this effort?
___	Purpose	___	What purpose(s) will it serve?
___	Questions	___	What questions will it address?
___	Information	___	What information is required?
___	Audiences	___	Whom will be served?
___	Agents	___	Who will do it?
___	Process	___	How will they do it?
___	Standards	___	By what standards will their work be judged?

Sociopolitical Factors

___	Involvement	___	Whose sanction and support is required, and how will it be secured?
___	Internal communication	___	How will communication be maintained between the evaluators, the sponsors, and the system personnel?
___	Internal Credibility	___	Will the evaluation be fair to person inside the system?
___	External Credibility	___	Will the evaluation be free of bias?
___	Security	___	What provisions will be made to maintain security of the evaluation data?
___	Protocol	___	What communications channels will be used by the evaluators and system personnel?
___	Public Relations	___	How will the public be kept informed about the intents and results of the evaluation?

Contractual/Legal Arrangements

___	Client/evaluator relationship	___	Who is the sponsor, who is the evaluator, and how are they related to the program to be evaluated?
___	Evaluation products	___	What evaluation outcomes are to be achieved?
___	Delivery schedule	___	What is the schedule of evaluation services and products?
___	Editing	___	Who has authority for editing evaluation reports?
___	Access to data	___	What existing data may the evaluator use, and what new data may he obtain?
___	Release of reports	___	Who will release the reports and what audiences may receive them?
___	Responsibility and authority	___	Have the system personnel and evaluators agreed on who is to do what in the evaluation?
___	Finances	___	What is the schedule of payments for the evaluation, and who will provide the funds?

The Technical Design

___	Objectives and variables	___	What is the program designed to achieve, in what terms should it be evaluated?
___	Investigatory framework	___	Under what conditions will the data be gathered, e.g., experimental design, case study, survey, site review, etc?
___	Instrumentation	___	What data-gathering instruments and techniques will be used?
___	Sampling	___	What samples will be drawn, how will they be drawn?
___	Data gathering	___	How will the data-gathering plan be implemented, who will gather the data?
___	Data storage and retrieval	___	What format, procedures, and facilities will be used to store and retrieve the data?
___	Data analysis	___	How will the data be analyzed?
___	Reporting	___	What reports and techniques will be used to disseminate the evaluation findings?
___	Technical adequacy	___	Will the evaluative data be reliable, valid, and objective?

The Management Plan

___	Organizational mechanism	___	What organizational unit will be employed, e.g., an in-house office of evaluation, a self-evaluation system, a contract with an external agency, or a consortium-supported evaluation center?
___	Organizational-location	___	Through what channels can the evaluation influence policy formulation

			and administrative decision-making?
___	Policies and procedures	___	What established and/or ad hoc policies and procedures will govern this evaluation?
___	Staff	___	How will the evaluation be staffed?
___	Facilities	___	What space, equipment, and materials will be available to support the evaluation?
___	Data-gathering schedule	___	What instruments will be administered, to what groups, according to what schedule?
___	Reporting schedule	___	What reports will be provided, to what audiences, according to what schedule?
___	Training	___	What evaluation training will be provided to what groups and who will provide it?
___	Installation of evaluation	___	Will this evaluation be used to aid the system to improve and extend its internal evaluation capability?
___	Budget	___	What is the internal structure of the budget, how will it be monitored?
Moral/Ethical/Utility Questions			
___	Philosophical stance	___	Will the evaluation be value free, value based, or value plural?
___	Service orientation	___	What social good, if any, will be served by this evaluation, whose values will be served?
___	Evaluator's values	___	Will the evaluator's technical standards and his values conflict with the client system's and/or sponsor's values, will the evaluator face any conflict of interest problems; and what will be done about possible conflicts?
___	Judgments	___	Will the evaluator judge the program; leave that up to the client; or obtain, analyze, and report the judgments of various reference groups?
___	Objectivity	___	How will the evaluator avoid being co-opted and maintain his objectivity?
___	Prospects for utility	___	Will the evaluation meet utility criteria of relevance, scope, importance, credibility, timeliness, and pervasiveness?
___	Cost/effectiveness	___	Compared to its potential payoff, will the evaluation be carried out at a reasonable cost?

The actual data gathering and analysis involved in implementing meta-evaluation Design #3 involve periodic reviews of the evaluation design and monitoring of the evaluation process. These review and monitoring activities are intended to determine whether the design has been adequately operationalized and how well the design is being carried out. Such data-gathering activities can be implemented by evaluation administrators through requiring the evaluators to make periodic vat and/or written progress reports. Another means of gathering this information is through employing external auditors to make periodic checks on the implementation of the evaluation design.

Feedback from meta-evaluation Design #3 includes two basic kinds of information. The first is a logging of the actual process of evaluation. This will be useful at the end of the evaluation project for interpretation of evaluation results. Another kind of feedback pertains to the identification of problems and recommendations for improving the evaluation activities. This type of feedback is important for the manager of the evaluation process.

In practice, there are many instances of meta-evaluations that check on and

guide the implementation of evaluation designs. Largely these pertain to self-assessment activities and sometimes to the employment of external consultants.

Design #4—for Pro-active Assessment of the Quality and Use of Evaluation Results

Design #4 provides for proactive meta-evaluation studies that enhance the quality and use of evaluation results.

In delineating the information requirements associated with this design type, three things must be done. The evaluation objectives should be noted; the meta-evaluation criteria of technical adequacy, utility, and cost/effectiveness should be spelled out in relation to the evaluation objectives; and the intended users of the primary evaluation results should be designated. Delineation of these matters provides a basis for obtaining the information needed periodically to assess the quality and impact of the evaluation information that is being gathered in the primary evaluation activity.

A number of things can be done to obtain information about the quality and impact of primary evaluation reports. Evaluation reports can be gathered and the information they convey can be rated for its validity, reliability, and objectivity. Records can be kept of primary evaluation expenditures. Records can also be kept of instances of use of the evaluation reports by the intended audiences. Also these audiences can be asked to rate the utility of the evaluation reports that they receive. Such information on the effectiveness of evaluation can be obtained by the evaluation manager or an external auditor.

It is to be noted that such meta-evaluation of the effectiveness of an evaluation might appropriately be conducted in conjunction with an effort to gather meta-evaluation data on the implementation of an evaluation design. While both meta-evaluation Designs #3 and #4 are implemented during the same time frame, feedback concerning the adequacy of implementation of an evaluation activity is relatively more important during the early stages of an evaluation project. Conversely, later in meta-evaluation projects feedback concerning the effectiveness of an evaluation is more important than is feedback about implementation of the primary evaluation design. This relationship between meta-evaluation Designs #3 and #4 is portrayed in Figure 7.

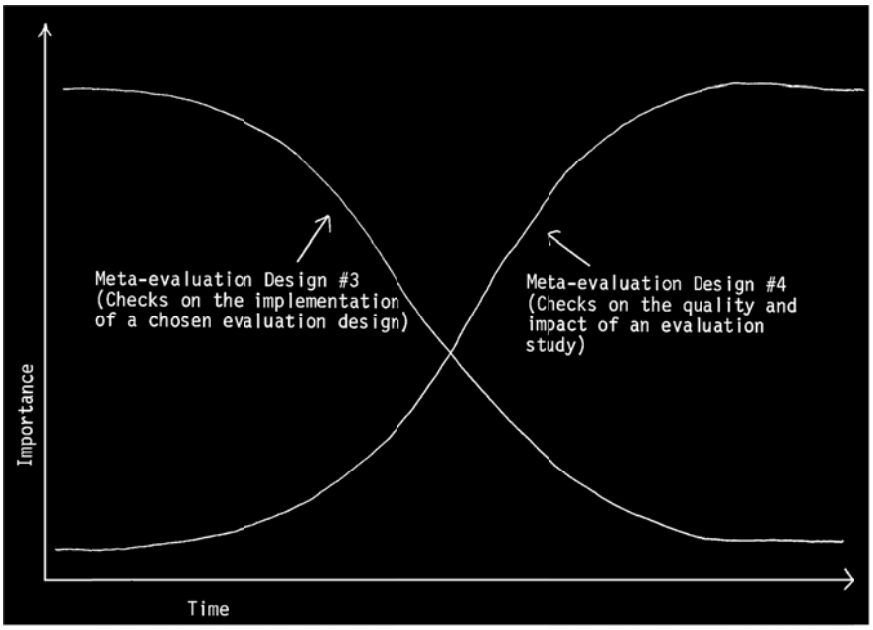
Feedback from meta-evaluation Design #4 includes periodic reports of the quality, impact, and cost/effectiveness of the evaluation work. The burden of these reports is periodically to rate the success of the evaluation results and to provide recommendations for improving the evaluation effort.

Design #5—for Retroactive Assessment of Evaluation Studies

With Design #5 we move to the area of retroactive meta-evaluation. In practice, the retroactive meta-evaluation of goals, designs, implementation, and results usually are combined into a single summative case study.

The first main step in implementing Design #5 is to determine the intents of the evaluator who conducted the study. What audience did he the effectiveness of an evaluation is more important than is feedback about implementation of the primary evaluation design. This relationship between meta-evaluation

Designs #3 and #4 is portrayed in Figure 7.



Information should be sought to answer a number of questions about the evaluation goals. What did the intended audience say they wanted from the evaluation? Were there other legitimate audiences? What information would they have wanted? Were alternative evaluation goals considered by the evaluator? What were they? Why were they rejected? Overall, how defensible were the evaluation goals? Data in response to these questions are needed for judging evaluation goals.

The evaluator should also compile information about the evaluation design that was chosen. How does it rate on criteria of technical adequacy, utility, efficiency, and feasibility? Were other designs considered? On balance was the chosen design better than others that might have been chosen?

Given the design that was chosen, the evaluator should next determine how well it was implemented. Was it carried out fully? If not, what difficulties accounted for the faulty implementation? Did the evaluator effectively counter the conceptual, sociopolitical, legal, technical, management, and ethical problems that were described in Part I? What did the implementation cost? Overall, how well was the evaluation design implemented, and what specific problems were encountered?

As a final issue, the evaluator should consider what results were produced. Were the objectives achieved? What information was produced? How good was it? Was it used? By whom? Did they use it appropriately? Overall, what impact was made by the evaluation, and how desirable and cost/effective was it?

The information obtained in response to the above questions should be combined into an overall report. This report should be written for and disseminated to the audiences for the primary evaluation that has been scrutinized. Through this practice, results of the primary evaluation can be viewed

and used in regard to both their strengths and weaknesses.

An example of Design #5 occurred when a research and development agency engaged a team of three meta-evaluators to assess the agency's evaluation system. The agency presented a conceptual framework to describe their evaluation system and charged the meta-evaluators to assess the agency's evaluation performance against the framework.

The framework appears in Figure 8. The horizontal dimension indicates that the agency's evaluation system should address questions about the system's goals, plans, processes, and achievements. The left-most vertical dimension references the levels of the parent agency, i.e., system, program, and project. The third dimension indicates that the evaluation system should be judged concerning whether audiences and evaluative questions have been delineated (the matter of evaluation goals), whether data collecting and reporting devices and procedures have been determined for answering the questions (evaluation design); whether evaluation data are actually being gathered and analyzed (implementation), and whether results are sound and used appropriately by the audiences.

The combination of the three dimensions of Figure 8 provide 48 cells that specifically focused the meta-evaluation work. Agency personnel were asked to generate documentation for what had been done regarding each of the 48 cells. For example, for cell 16 they were asked to produce data that their evaluators had obtained concerning the impacts of the agency on its target population and to describe how the impact data had been used. For cell 18 they were asked to describe how alternative program designs and criteria for judging them are identified in the agency. In cell 22 they were asked to produce procedures and instruments that their evaluators had used to judge program plans. In cell 26 they were asked to produce data regarding alternative program designs; and in cell 30 they were asked to produce evidence concerning the quality and use of their data about alternative program designs. What the meta-evaluators wanted then was information about evaluation goals, designs, implementation, and results) and they wanted it for each of the three levels in the agency and for the program variables of goals, designs, implementation, and results.

Organizational Levels	Evaluation Attributes	Program Attributes			
		Goals	Designs	Implementation	Results
Level 1 (e.g., System)	Delineation of questions and audiences	1[a]	2	3	4
	Operationalization of evaluation procedures and devices	5	6	7	8
	Implementation of evaluation procedures	9	10	11	12
	Use of evaluation data	13	14	15	16
Level 2 (e.g., Program)	Delineation	17	18	19	20
	Operationalization	21	22	23	24
	Implementation	25	26	27	28
	Use	29	30	31	32
Level 3 (e.g., Project)	Delineation	33	34	35	36
	Operationalization	37	38	39	40
	Implementation	41	42	43	44
	Use	45	46	47	48

[a]cell number

Figure 8
A Framework for Describing and Judging an Evaluation System

The agency personnel responded by preparing notebooks of information that were organized according to the dimensions of their evaluation framework. Included were three major parts on system evaluation, program evaluation, and project evaluation. Within each part were sections on the evaluation of goals, designs, process, and results. Each of these sections contained the information needed by the metaevaluators concerning the evaluation's goals, designs, implementation, and results. Thus, the notebook of information provided the initial basis for evaluating the agency's evaluation system.

The meta-evaluators prepared nine-items rating scales for each of the 48 cells. These scales were to be used for rating the quality of the agency's evaluation work in each of the 48 cells.

The meta-evaluators then visited the agency for three days. During that time they read pertinent documents, studied the contents of the specially prepared notebooks, and interviewed personnel at each of the three levels of the agency. Then the meta-evaluators independently completed the 48 rating scales. The three observations per cell provided a basis for determining inter-judge reliability and for analyzing the strengths and weaknesses of the evaluation system. Subsequently, the meta-evaluators developed a table that essentially was the agency's logical evaluation structure with mean ratings of quality in each of the 48 cells. This table was used as a basis for an initial report and exit interview with agency personnel.

Three main findings were presented during that session: (1) the agency was generally strong in identifying and assessing alternative plans, (2) the agency

was somewhat weak at all organizational levels in assessing results, and (3) the program level evaluation was almost nonexistent. The agency personnel were interested in the judgments of the system- and project-level evaluation but were startled and concerned about the poor showing of their program-level evaluation. They asked the meta-evaluators to provide recommendations in the written report concerning what could be done to change this situation.

Following the visitation, the meta-evaluators wrote and submitted their final report. It was focused on the 48 cell table of judgments that had been prepared on site. However, it was broader than that. Generally, it addressed ten questions about the agency's evaluation system:

1. whether it addresses worth and merit questions regarding goals, designs, implementation, and main and side effects;
2. whether it does so both pro-actively and retroactively in relation to decisions about the four question types;
3. whether the four question types and the key audiences for the evaluation are explicated at each organizational level;
4. whether explicit, sound procedures, and instruments have been (or will be) determined for answering the specified questions at each level. (The concern here is with the criteria of technical adequacy [reliability, internal validity, external validity, and objectivity]; utility [timeliness, relevance, importance, pervasiveness, credibility, and scope]; and the prudential criterion of the cost/efficiency of the evaluation.);
5. whether data required to answer the specified questions are being obtained at each organizational level;
6. whether data concerning the specified questions systematically are being organized and stored in retrievable form to meet accountability needs at each organizational level;
7. whether the evaluation system is having an impact on the decisions related to the four question types at each level;
8. whether the evaluation system has the capacity to identify and respond to emergent evaluative needs at each level;
9. whether the evaluation system is being implemented so as to enhance prospects for systematic evaluation beyond those short-term efforts supported through externally funded projects, and
10. whether a strong case can be made that the cost of the evaluation system is appropriate for satisfying the criteria enumerated above.

The report that was submitted in regard to the above questions pinpointed program-level evaluation as the weakest part of the agency's evaluation work. The report further speculated that the programs themselves were not taken seriously in the agency—that in fact the agency was only a holding company for miscellaneous projects. An unexpected effect of the report was that it lead to a reorganization of the total agency in order to strengthen both programs and the mechanism for evaluating them. This illustrates that meta-evaluation can play strong roles in effecting change—not just in evaluation but also in the enterprise being evaluated.

Another instance of Design #5 occurred when Ernest House, Wendell Rivers, and I were engaged by the National Education Association and the Michigan Education Association to evaluate the Michigan Education Department's statewide accountability system (House et al, 1974, Kearney et al, 1974; Stufflebeam, 1974).

The Michigan accountability program had been the center of much controversy, and I realized early that our study would be political, difficult to conduct, and suspect by whoever would oppose our findings. Before agreeing to participate, I met with members of MEA and NEA, with Rivers and House, and with a representative of the Michigan Department of Education to determine whether our team could conduct an independent study and whether we would have access to all the requisite data. After being satisfied on these two points, we designed, conducted, and reported the study.

A key feature of our design was an advance set of working agreements intended to summarize our plan and guarantee the independence and viability of our study. Our ten working agreements were as follows:

1. Charge
   The external evaluation panel consisting of Ernest House, Wendell Rivers, and Daniel Stufflebeam have been engaged by the Michigan Education Association and the National Education Association to evaluate the educational soundness and utility for Michigan of the Michigan Accountability Model with a particular focus on the assessment component.
2. Audiences (in priority order)
   a. NEA/MEA
   b. Decision makers in Michigan's educational system (State Board of Education and State Department of Education)
   c. The media (the public)
   d. Consumers (parentis, PTA, the public, etc.)
   e. Technical persons (especially in the area of educational measurement).
3. Report/Editing
   The panel will be solely in charge of developing and editing its final report. NEA/MEA may write and disseminate any separate statement (such as an endorsement, a rebuttal, a commentary, or a descriptive piece). It is understood that the panel's report is to be as short and direct as possible and to be designed to communicate with the audiences designated for the report.
4. Dissemination
   The external panel has the right to release its report to any members of the target audiences or other persons following the completion of the report. The panel's release of the report will imply no MEA/NEA endorsement. Further MEA/NEA may choose to endorse or not endorse the report depending on their judgment of the quality and appropriateness of the report. Should MEA/ NEA decide to disseminate their own document describing the report, their document will be identified as their own and not that of the committee. Only the committee's final report as edited by the committee will be distributed with the names of the committee on it.

5. Format of the Report
   The following items were identified as desirable ingredients for the panel's final report:
   a. citation of the agreements between the review panel and MEA/NEA.
   b. presentation of the major findings.
   c. presentation of minority opinions, if any.
6. Questions to be Addressed in the Report
   Specific questions to be addressed will include:
   a. validity and reliability of criterion-referenced tests.
   b. use of tests to evaluate staff.
   c. merit of the objectives on which Michigan assessment is based.
   d. involvement of teachers in developing both objectives and tests.
   e. the panel's recommendations for change and further study.
   f. comments about the balance of the state effort and appropriateness of expanding the scope of assessment especially given cost factors associated with the projections for improving or expanding Michigan assessment.
   g. quality of planning in the Michigan Accountability Program.
   h. cost benefit projections for the program.
   i. value of Michigan assessment outcomes and reports for different levels of audiences in Michigan
   j. problems of bias in the Michigan Accountability Program
7. Resources (budget) to Support the Evaluation
   Sufficient resources will be made available by MEA/NEA to the external review panel to support eight days of work per panelist to work on evaluation, whatever secretarial support is needed in conducting the evaluation and whatever materials and equipment are needed in the Lansing hearings. It is understood that if any of the panelists need to make long distance telephone calls in collecting opinions about the program from people in Michigan that the panelists will be reimbursed for such expenses provided that an accurate and complete report is made of the purpose of the phone call and who was contacted.
8. Delivery Schedule
   The panel is to deliver its final report on March 1 or as soon thereafter as is practicable.
9. Access to Data
   It is understood that the Michigan Department of Education will make available to the panel any and all data and reports required by the panel to do the job. This, of course, is restricted to those data and reports that are now available to the Michigan Department of Education regarding Michigan accountability.
10. Procedures
    Pursuant to the above conditions, the external three man panel will have control over the evaluation process that it must implement to responsibly respond to the charge to which it has agreed. In accordance with this position, the panel has agreed to implement the

following general process. Private interviews and hearings will be conducted solely by the panel with representatives of the Michigan Department of Education, representatives of MEA/ NEA, representatives of selected groups (teachers, administrators, board members, and educational action groups). The panel will also review documents made available to it by MEA/NEA and the Michigan Department of Education. Finally, the panel will conduct a hearing to obtain additional information concerning issues identified by the panel in the course of interviewing various client groups and studying various documents (Stufflebeam, 1974).

This concludes Part III and the discussion of meta-evaluation designs. An attempt has been made to present general designs that cover the different meta-evaluation assignments. Actual cases that relate to the designs have been described to demonstrate that meta-evaluations are real and not just theoretic. The designs are cryptic, and the examples few; it is hoped that others will extend and improve on these designs and examples.

Summary

The purpose of this paper has been to explore the topic of metaevaluation. Part I discussed the need to develop a technology for evaluating evaluation; described eleven meta-evaluation criteria; and delineated six classes of problems that plague evaluation efforts. Part II presented a definition, eight premises, and a logical structure for meta-evaluation work. Part III described how the structure might be used through describing and illustrating five meta-evaluation designs.

It is hoped that this paper will stimulate further actions. Hopefully, some of the ideas and devices will be of use to persons who evaluate evaluations. It is hoped that other persons might be stimulated by this paper to further delineate and operationalize metaevaluation concepts.

Given the poor quality of evaluation performance in education, and the lack of a research base to guide evaluators, it seems urgent to contrive ways of defining, assuring, and documenting the quality of evaluation work. This paper has been one attempt to move the field of evaluation toward a technology for evaluating evaluation.

Selected Bibliography

Adams, James A. A Manual of Policies, Organization and Procedures for Evaluation. Saginaw School District, 1970.

"A Study of the Status, Scope and Nature of Educational Evaluation in Michigan's Public K-12 School Districts." Ph.D. Dissertation, Ohio State University, 1971.

Assessment Council. "Handbook of Policies and Procedures for Evaluation." Columbus, Ohio: College of Education, Ohio State University, 1970. (mimes)

Bettinghaus, Erwin P., and Miller, Gerald R. A Dissemination System for State Accountability Programs; Part II: The Relationship of Contemporary Communication Theory to Accountability Dissemination Procedures. Cooperative Accountability Project, Colorado Department of Education. June 1973.

Bloom, Benjamin S. (ed.) Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: David McKay Co., Inc., 1956.

Bracht, Glenn H., and Glass, Gene V. "The External Validity of Experiments," American Educational Research Journal 5 (November 1968): 437-474.

Buros, Oscar Krisen (ed.) The Third Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1949.

(ed.). The Fourth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1953.

(ed.). The Fifth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1959.

ed.). The Sixth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1965.

Campbell, Donald T., and Stanley, Julian C. "Experimental and Quasi-Experimental Designs for Research on Teaching," Handbook of Research on Teaching. Edited by N. J. Gage. Chicago, Ill.: Rand McNally and Co., 1963. pp. 171-246.

Clark, David L., and Cuba, Egon G. "An Examination of Potential Change Roles in Education," Essay 6 in Ole Sand (ad.), Rational Planning in Curriculum and Instruction. Washington, D.C.: National Education Association, Center for the Study of Instruction, 1967. pp. 111-134.

Cook, Desmond L. Program Evaluation and Review Technique Applications in Education. Washington, D.C.: U.S. Government Printing Office, 1966.

Cook, Thomas. Evaluation Essay. Berkeley, California: McCutchan Publishing Co., 1974.

Cyphert, F. R., and Cant, W. L. "The Delphi Technique: A Tool for Collecting Opinions in Teacher Education." Journal of Teacher Education 31, 1970. pp. 417-425.

Fry, Charles. "SMI Institutional Support and Evaluation Policy" Working Paper, The Division of Manpower and Institutions, June 3, 1971. (mimeo)

Gephart, W. J.; Ingle, R. B.; and Remstad, R. C. A Framework for Evaluating Comparative Studies. In Henry Cody (ed.) Conference on Research in Music Education. U.S. Office of Education Cooperative Research Report N. 6--1388. May, 1967.

Glass, Gene V., and Stanley, Julian C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

Cuba, Egon G. "The Failure of Educational Evaluation," Educational Technology, Vol. IX, No. 5 (May 1969). pp. 29-38.

, and Stufflebeam, Daniel L. Evaluation: The Process of Stimulating, Aiding, and Abetting Insightful Action. Monograph Series in Reading Education, Indiana University. No. 1, June 1970.

Brickell, Henry M.; Lucas, Robert E.; Michael, William B.; Minor, Robert E.; Ruffer, David G.; Stufflebeam, Daniel L.; Walter, Franklin B. Ohio Accountability Project: Team One Report. A report presented to the Ohio State Department of Education. Prepared by the Ohio State University Evaluation Center. RF No. 3503A-1, July 3, 1972-December 1, 1972.

Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1965. (4th edition)

Hammond, R. "Context Evaluation of Instruction in Local School Districts," Educational Technology, 1969, 9 (1), pp. 13-18.

. "Priorities for Evaluating Programs," Community Education Journal. Vol. V, No. 2 (Mar.-Apr., 1975). pp. 13-17.

House, Ernest; Rivers, Wendell; and Stufflebeam, Daniel L. "An Assessment of the Michigan Accountability System" under a contract with the Michigan Education Association and The National Education Association. March, 1974.

Jaeger, Richard M.; Cahen, Leonard; Cohen, David; Jacobs, James, Linn, Robert; Mazur, Joseph; Wardrop, James. Ohio Accountability Project: Team Two Report. A report presented to the Ohio State Department of Education. Prepared by the Ohio State University Evaluation Center. RF No. 3503-A-1, July 3, 1972-December 1, 1972.

Kearney, C. Phillip; Donovan, David L.; and Fisher, Thomas H. "In Defense of Michigan's Accountability Program," Phi Delta Kappan. Vol. LVI, No. 1 (September 1974). pp. 14-19.

Krathwohl, David R.; Bloom, Benjamin S.; and Masia, Bertram B. Taxonomy of Educational Objectives, HandbookII: Affective Domain. New York: David McKay Co., Inc., 1964.

. "Functions for Experimental Schools Evaluation and Their Organization" in Glass, Gene V.; Byers, Maureen L.; and Worthen, Blaine R. Recommendations for the Evaluation of Experi-mental Schools Projects of the U.S. Office of Education. Report of the Experimental Schools Evaluation Working Conference: Estes Park, Colorado, December 1971. University of Colorado: Laboratory of Educational Research, February 1972. pp. 174-194.

Lake, Dale G.; Miles, Matthew B.; and Earle, Ralph B., Jr. Measurinq Human Behavior. Columbia University: Teachers College Press,]973.

Lessinger, Leon. Every Kid a Winner: Accountability in Education. Palo Alto, Calif.: SRA, 1970.

Merriman, Howard O. The Columbus School Profile. Columbus, Ohio: The Columbus Public Schools, May, 1969.

Metfessel, N. S., and Michael, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. Educational and Psychological Measurement, 1967, 27, pp. 931-943.

Nystrand, Raphael 0.; Ashburn, Arnold G.; Campbell, Ronald F.; Cresswell, Anthony; Garner, William; Greer, Robert 0. Ohio Accountability Project: Team Three Report. A report presented to the Ohio State Department of Education. Prepared by the Ohio State University Evaluation Center. RF No. 3503-A-1, July 3, 1972-December 1, 1972.

U.S. Department of Health, Education and Welfare. Office of Education. The States Report: The First Year of Title I, Elementary and Secondary Education Act of 1965. Washington, D.C.: U.S. Government Printing Office, 1967.

O'Keefe, Kathleen. "Methodology for Educational Field Studies." Ph.D. Dissertation, Ohio State University, 1968.

Provus, Malcolm. Discrepancy Evaluation. Berkeley, Calif.: McCutchan Publishing Corporation, 1971.

. "In Search of Community," Phi Delta Kappan. Vol. LIV, No. 10 (June 1973). pp. 658.

Reinhard, Diane L. "Methodology Development for Input Evaluation Using Advocate and Design Teams." Ph.D. Dissertation, Ohio State University, 1972.

Root, Darrell K. "Educational Evaluation Training Needs of Superintendents of Schools." Ph.D. Dissertation, The Ohio State University, 1971.

Scriven, Michael S. "The Methodology of Evaluation," Perspectives of

Curriculum Evaluation (AERA Monograph Series on Curriculum Evaluation, No. 1). Chicago: Rand McNally & Co., 1967.

. "An Introduction to Meta-Evaluation," Educational Product Report, Vol. 2, No. 5 (February 1969). pp. 36-38.

; Glass, Gene V.; Hively, Welis; and Stake, Robert E. "An Evaluation System for Regional Labs and R & D Centers." A report presented to the Division of Research and Development Resources, National Center for Educational Research and Development, U.S. Office of Education. Project No. 1-0857; Grant No. OEG-0-71-4558. August 31, 1971."

. "The Pathway Comparison Model of Evaluation." January 1972 (a) (mimes).

. "Goal-Free Evaluation. The Journal of Educational Evaluation-Evaluation Comment. December 1972 (b).

. "Maximizing the Power of Causal Investigations–The Modus Operandi Method." July 1973. (mimes)

. "Bias and Bias Control in Evaluation." The Evaluation Center Occasional Paper Series, Western Michigan University. In press.

Seligman, Richard. "College Guidance Program." Measurement and Evaluation and Guidance. Vol. 6 (June 1973), pp. 127-129.

Siegel, S. S. Nonparametric Statistics. New York: McGraw Hill, 1956.

Stake, Robert E. "The Countenance of Educational Evaluation," Teachers College Record, Vol. 68 (1967), pp. 523-540.

Stufflebeam, Daniel L.; Foley, Walter J.; Gephart, William J.; Guba, Egon G.; Hammond, Robert L.; Merriman, Howard 0.; and Provus, Malcolm. Educational Evaluation and Decision Making. Itasca, Illinois: F. E. Peacock Publishers, Inc., 1971. (a)

. "The Use of Experimental Design in Educational Evaluation," Journal of Educational Measurement. Vol. 8, No. 4 (Winter 1971. (b)

. Design of a Planning and Assessment System for the Division of Manpower and Institutions. Proposal submitted to the Office of Education by the Ohio State University Research Foundation, June 18, 1971. (c)

; Brickell, Henry M.; Cuba, Egon G.; and Michael, William B. "Design for Evaluating R & D Institutions and Programs." A Report presented to the Division of Research and Development Resources, National Center for Educational Research and Development, U.S. Office of Education. Project No. 1-0857, Grant No. OEG-0-71-4558. August 31, 1971.

. "Part I: A Conceptual Framework for the Evaluation of Experimental Schools Projects" in Glass, Byers, Worthen. Recommendations for the Evaluation of Experimental Schools Projects of the U.S. Office of Education: Recort of the Experimental Schools Evaluation Working Conference: Estes Park, Colorado, December 1971. University of Colorado: Laboratory of Educational Research, February 1972, pp. 128-135.

. "A Response to the Michigan Education Department's Defense of Their Accountability System." The Evaluation Center Occasional Paper Series, Paper NO. 1, Western Michigan University, August 27, 1974.

"Technical Recommendations for Psychological Tests and Diagnostic Techniques." Psychological Bulletin, 1S54, 51, Supplement.

Turner, Richard L. "Appendix G: Criteria for Comprehensive Evaluations and

the Appraisal of Evaluation Success in Experimental School Contexts," in Guba, Egon G.; Clark, David L.; McClellan, Mary C.; Sanders, James R.; and Turner, Richard L. The Design of Level III Evaluation for the Experimental Schools Program. A report presented to the U.S. Office of Education. Project No. R020862; Grant No. OEG 0-72-1867, September 30, 1972.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.

Wolf, Robert L. "The Application of Select Legal Concepts to Educational Evaluation." Ph.D. Dissertation, University of Illinois at Urbana-Champaign, 1974.

Worthen, Blaine R., Frazier, Calvin M.; Hood, Paul D.; Lidstrom, David C., Millman, Jason; Rogers, W. Todd, Shephard, Loretta A. "Critique of Three Proposed Designs for an Accountability System for the State of Ohio." A report presented to The Ohio Accountability Project, The Ohio State University Evaluation Center, Re: RF No. 3503-A-1, July 3, 1972-December 1, 1972. November 20, 1972.