

EVALUATION GOALS, OBJECTIVES, AND QUESTIONS

The kinds of outcomes considered legitimate in American schools appear to be an ever-changing phenomenon. Every governor's conference, political campaign, professional meeting, and state educational reform brings new objectives and a redistribution of priorities. Unfortunately, all too often there is no redistribution of revenues. Several seemingly contradictory trends are evident. On the one hand, there appears to be a push on the part of society to force the schools back to a basic skill development orientation. The often heated rhetoric about reading, math, and minimum competencies attest to this public concern. From another standpoint the schools appear to be taking over many of the educational responsibilities that were historically considered the prerogative of parents and other socializing agencies of society. Such areas as sex and health education, human relations (including marriage), and the strands of values, morality, and ethics are now addressed in the schools. Overlaying these trends is the desire to develop and enhance higher order thinking skills and problem--solving abilities.

The "accountability movement" is rampant in virtually all funding agencies, especially at the state and federal levels. Accountability requires documentation of program impact. The demand for responsive and relevant evaluation is, therefore, ever increasing.

With changes in goals and objectives, and accountability, come sociopolitical problems. Educational institutions, from the elementary grades through professional schools, tend to reflect changes in society. Social forces from a variety of political, legal, religious, or economic origins generally find manifestations in curriculum reform, modified instructional systems and professional training programs, or school assessment practices. Many of these forces are at work in today's schools. Such factors as civil rights, the feminist movement, recession/inflation, and consumer awareness impinge on school practice. The bottom line is that in evaluating these changes we had better be asking the right questions or we will get the wrong answers, or we may get answers for questions we didn't ask, or we might find out things we didn't want to know, or . . . well, you get the idea.

The importance of involving stakeholders in evaluation question preparation at the earliest possible date, therefore, cannot be overemphasized.

IDENTIFYING AND INVOLVING STAKEHOLDERS

A great variety of people and organizations will usually be interested in the results of any evaluation. They will vary in the degree of intensity of that interest, but listening to as many of them as possible will help the evaluator frame the right questions. The right questions most likely will lead to the collection of the most relevant information, which in turn will increase the likelihood that these results will be used. Nothing is more frustrating and cost-ineffective than the nonutilization of results. Unfortunately conducting evaluations for the sake of appearance occurs with too great a frequency. These "symbolic" evaluation data are often used as political weapons. Evaluations are most meaningfully done when the results are to be used in explaining ideas, theories, or concepts, or in specific decision-making situations.

The variety of potential stakeholders, decision makers, and audiences is well represented in a list compiled by Rossi and Freeman (1989, p. 423). The term stakeholder as used here refers to individuals who have a vested interest in the outcomes of the evaluation. Stated another way, the results of the evaluations will have consequences for the stakeholder. The consequences could be financial, emotional, political, or professional/vocational (Scriven, 1991).

- *Policy Makers and Decision Makers*: Persons responsible for deciding whether a program is to be instituted, continued, discontinued, expanded, or curtailed.
- *Program Sponsors*: Organizations that initiate and fund the program to be evaluated.
- *Evaluation Sponsors*: Organizations that initiate and fund the evaluation. (Sometimes the evaluation sponsors and the program sponsors are identical.)
- *Target Participants*: Persons, households, or other units who participate in the program or receive the intervention services under evaluation.
- *Program Management*: Group responsible for overseeing and coordinating the intervention program.
- *Program Staff*: Personnel responsible for actual delivery of the intervention (e.g., teachers).
- *Evaluators*: Groups or individuals responsible for the design and/or conduct of the evaluation.
- *Program Competitors*: Organizations or groups who compete for available resources.

- *Contextual Stakeholders*: Organizations, groups, individuals, and other units in the immediate environment of a program (e.g., local government officials or influential individuals situated on or near the program site).
- *Evaluation Community*: Other evaluators, either organized or not, who read and evaluate evaluations for their technical quality.

A given evaluation may involve only two or three of these groups, but an evaluator can be very surprised by how many groups and individuals may be interested in the results.

The evaluator will always be faced with resource--allocation conflict situations. The best professional judgment will need to be applied in deciding on which evaluation question(s) to target. A frequently followed realistic road is that of compromise, as long as professional integrity is not subverted. Even with a very large budget, employment of every known relative, and all the cooperation in the world, it is impossible to investigate or respond to all potentially relevant evaluation questions. It would seem reasonable to focus on a limited number of questions and do the best possible job on those.

KINDS OF EVALUATION QUESTIONS

The kinds of questions to be addressed by curriculum and program evaluators will, of course, be dictated by the information requirements of decision makers, and the nature and state of reforms or innovations that are proposed or have been implemented. Some questions might be considered *formative*, for example, how can we improve the materials used in the elementary mathematics curriculum? Or questions might be *summative* in nature; for example, should the current approach to the teaching of writing be continued? In any event, the evaluation question should be based on objectives or goals. We need goals and objectives to help us frame the right evaluation questions.

Following is a list of sample evaluation questions that might be asked. The list and kinds of questions are limited only by the creativity of the evaluator (Payne, 1982a).

Focus Category

Sample Evaluation Question

1. General Needs Assessment

Are the general system objectives in mathematics being met in our elementary schools?

- | | |
|----------------------------------|--|
| 2. Individual Needs Assessment | Are the career information needs of our graduating students being met? |
| 3. School Services | Are our school psychological services perceived as adequate by students? |
| 4. Curriculum Design | What effect has the implementation of the new way of organizing the mathematics courses over the school year had on student achievement? |
| 5. Classroom Process | Are teachers following the prescribed teaching techniques in using the new Muth Affective Education Program? |
| 6. Materials of Instruction | Is the drug abuse filmstrip/tape program more effective than the current combination of lecture and programmed materials? |
| 7. Monitoring of Student Program | Is our current performance and records system adequate in identifying those students in need of academic counseling? |
| 8. Teacher Effectiveness | To what extent has teachers' verbal reinforcement techniques resulted in a decrease in student retention? |
| 9. Learner Motivation | Has the tracking system based on post-high--school aspirations resulted in changes in learner motivation? |
| 10. Learning Environment | What changes in classroom climate, as perceived by students and faculty, have accompanied the introduction of the new position of assistant principal? |

- | | |
|-------------------------------|---|
| 11. Staff Development | Did last week's staff-development program on creating performance assessments result in improved teacher skills? |
| 12. Decision Making | To what extent have central office decisions in the last 24 months resulted in lower costs and improved student attitude toward school? |
| 13. Community Involvement | Is community involvement in the instructional program a good thing? |
| 14. Board of Education Policy | Are system policies effectively communicated to relevant personnel? |
| 15. School Outcomes | To what extent are major cognitive, affective, and psychomotor outcomes being accomplished on a schoolwide basis? |
| 16. Resource Allotment | Are projected allotments in the budget adequate for anticipated needs? |
| 17. Instructional Methods | Do students have a positive attitude toward the Goetz Word Processing Program? |

There is a tremendous variety of potential tasks represented here, and they would require a great variety of measurement techniques: among them traditional multiple-choice (or true-false and essay) achievement measures, questionnaires, surveys, attitude scales, observations, interviews, and perhaps cost-effectiveness analyses. The key to successful evaluation, however, is a systematic process.

But we are getting ahead of the story. Before deciding on how to get to where you are going and whether you enjoyed it once you got there, you must decide on where you want to go. Establishing goals and objectives helps in doing that. We are talking about a variety of educational outcomes. School outcomes can be specified at different levels of generality.

LEVELS OF SPECIFICITY IN EDUCATIONAL OUTCOMES

Educational outcomes like people, come in all shapes and sizes. There are big ones that at times are so ponderous that they don't say anything and can't move anywhere. There are others so small as to be almost microscopic. Many are so minuscule that you can't see them and are meaningless because they are so small, even nitpicky. "Truth," is as so often the case, probably falls somewhere in the middle. Outcomes that are useful undoubtedly have enough bulk to make themselves visible and make a statement but not be so small that they become intellectually invisible. One might conceive of outcomes as falling on a continuum of specificity. Figure 3-1 should help visualize the

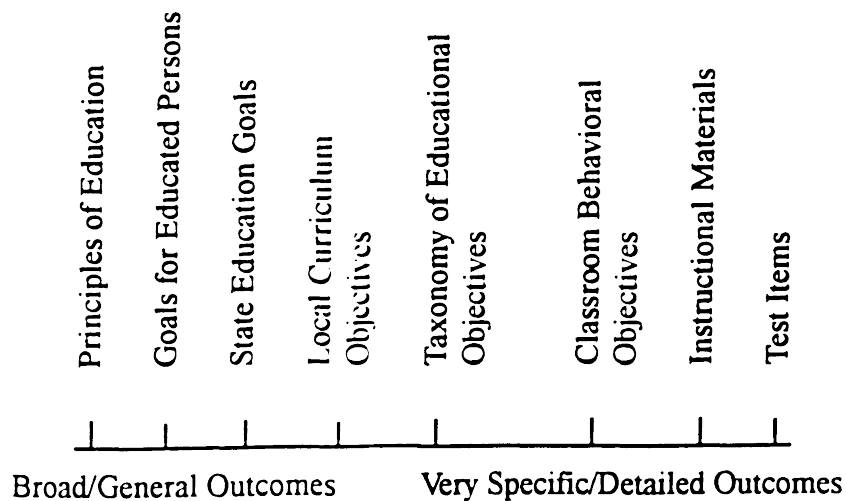


Figure 3-1 Degrees of Specificity of Educational Outcomes

individual differences in specificity. It contains a variety of terms that are frequently used to help focus and direct educational efforts. At the very general end we have educational goals like "Become a good citizen." In the middle (Taxonomy of Educational Objectives) (Bloom, 1956) we might have "Applies Archimedes principles of specific gravity in problem solving." At the specific end we have test items: "What domestic animal is most closely related to the wolf?"

Note that the spacing of the outcome-related terms is not even, since objectives and categories of objectives are not created equal. Figure 3-1 is not an equal interval scale. It can be seen that goals like those from national educational commissions would be left--the *far* left--on our continuum, and the ultimate in specificity is the test or performance item on the right. The test task is an actual sample of what we want the student to know or be able to do. If not an actual sample, it is as good an approximation as we can create.

The process of stating objectives is an iterative one; each level helps one understand the levels above and below it. There is lots of interaction. Developments at one level frequently have implications for other levels, and one obtains the most complete understanding--particularly once the major developmental lines have become clear--by working back and forth among the various levels. Thus it is clear that objectives can and must be stated at a variety of levels of specificity, for both curriculum building and evaluation.

One way of thinking of Figure 3-1 is in terms of implementing an evaluation project. One would begin with broad general goals, then develop evaluation questions, and finally use objectives in the development of program activities, and create or select evaluation instruments.

Illustrations of the larger educational goals are the hoped-for outcomes of our public schools explained by the National Governor's Association on February 25, 1990. These national education goals are as follows:

- Goal 1: By the year 2000, all children in America will start school ready to learn.
- Goal 2: By the year 2000, the high school graduation rate will increase to at least 90 percent.
- Goal 3: By the year 2000, American students will leave grades 4, 8, and 12 having demonstrated competency over challenging subject matter including English, mathematics, science, history, and geography, and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.
- Goal 4: By the year 2000, U.S. students will be first in the world in mathematics and science achievement.
- Goal 5: By the year 2000, every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship.
- Goal 6: By the year 2000, every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning.

These general statements help us see broad-based intents and may even have some implications for resource allocation. But something a little more specific is needed. We probably do not need, however, to go back to the 1960s and "behavioral objectives" when there was an effort to behavioralize everything short of respiration and blood flow in our classrooms.

PROGRAM DEVELOPMENT AND EVALUATION QUESTIONS

Rather than focusing on broad general content categories, a more systematic way of classifying evaluation questions is in terms of where they fit in the program or project development process. Pancer and Westhuer (1989) have proposed an eight--stage evaluation model that logically follows the development and implementation of a program from consideration of general standards or values to be met, to an assessment of the outcomes of the program. Using this general framework, Kaluba and Scott, in an unpublished 1993 paper, have identified not only what questions could be asked but also how one might go about collecting relevant data. The context for their illustration is taken from the task of revising an introductory statistics course to better meet student needs and enhance learning. The primary vehicle was to introduce an interactive computer--based tutorial program. Obviously the decision to use a computer--based tutorial program did not take place until there was general consideration of what the course and faculty were trying to accomplish (and why), what the needs were of students, and what alternatives existed. An outline of their application of the Pancer-Westhuer framework follows.

| <u>Stage</u> | <u>Questions To Be Asked</u> | <u>Data Sources</u> |
|--|--|--|
| (1) Determination of values with respect to statistics education | a. What level of statistics mastery is acceptable for successful course completion? b. Should this level be maintained across all sections? | Committee discussion Opinion of faculty |

| <u>Stage</u> | <u>Questions To Be Asked</u> | <u>Data Sources</u> |
|-------------------------------------|---|---|
| (2) Assessment of educational needs | a. To what extent are mastery levels achieved? | Committee discussion |
| | b. Is mastery the same across all sections? | Follow-up surveys of former course takers Professional organization opinion |
| (3) Determination of goals | a. What must be changed in order to meet these levels with consistency? | Committee discussion Faculty and student input |
| | (4) Design of program alternatives | a. What kinds of course programs could be used to produce the desired changes? Review theories Review comparable programs at other institutions |
| (5) Selection of alternatives | a. Which of the program alternatives should be selected? | Review source of funding Department feasibility assessment |
| (6) Program implementation | a. How should the program be put into operation? | PERT & GANTT charts |
| (7) Program operation | a. Is the course operating as planned? | Surveys, observations, peer review |
| (8) Program outcomes | a. Is the course having desired effects? | Achievement measures, attitude scales, case studies |
| | b. At desired level of cost? | |

Two points need highlighting. The first is the very important and intimate association between developmental stages and evaluation questions. The second point relates to an again intimate association between evaluation questions and data collection method and source.

Given that the nature of the evaluation question will be associated with the developmental stage of the project, how does one go about selecting and stating them?

DEVELOPING EVALUATION QUESTIONS

Cronbach (1982) suggests that evaluation questions emerge after the evaluator has engaged in a two--stage process. The first, or divergent, phase involves getting input from as great a variety of sources as possible. Obviously the evaluator wants input from the primary stakeholder(s), but in addition related audiences should be consulted. The project or program staff will have ideas about what should be expected outcomes. The "targets" themselves, e.g., students, can be very valuable resources, particularly the analyses of their needs. If external funding is involved, that source should be heard. Even critics or detractors of the proposed program should be given consideration. The evaluator may interview participants, examine records, and seek recommendations from the professional literature. In addition, the evaluator him or herself will have certain expectations that they have learned from experience. The end result of this process will be far too many candidates. There will be far too many questions that could be answered effectively and efficiently in a lifetime of evaluations. The evaluator is faced with a limited budget, time constraints, and restricted personnel resources. How does reconciliation take place?

Again, according to Cronbach (1982), a second, or convergent, phase of evaluation question development can be described. Cronbach suggests that the reducing and winnowing process can be accomplished by, in essence, judging individual question candidates according to each of the following two criterion questions.

1. Will the answering of this evaluation question significantly increase my knowledge and understanding about the phenomena being investigated?
2. Will this knowledge and understanding allow exertion of leverage about a decision?

In the first question we are asking to reduce the uncertainties surrounding the object of the investigation. The second question goes to the old adage that in a real sense--knowledge is power. If you want to change the belief

system, and eventually the behavior, of possible decision makers, provide them with the most useful, meaningful, and relevant information possible. One could conceive of a simple 2 x 2 table related to these variables.

| | | Reduction of Uncertainty (Knowledge) | |
|--|------|--|-----|
| | | High | Low |
| Likelihood of Increased Leverage | High | A | C |
| | Low | B | D |

Questions in cell A should probably receive our major attention and the lion's share of the resources. Then, depending on whether the evaluator believes that knowledge or leverage is more important, questions in cells B and/or C might be addressed. Questions categorized in cell D will probably not be addressed unless it can be accomplished with very low-level resource allocation.

The reduction of the question pile must be a joint effort among evaluator, major stakeholders (or a representative), project staff, and fiscal agent if appropriate.

What might a final evaluation question look like?

Formatting Evaluation Questions

Referring back to our illustration with the computer assisted tutorial project wherein the Pancer-Westhuer developmental process was used, we identified the following program goals:

1. To increase student performance in the statistics course.
2. To reduce the proportion of students dropping or failing the statistics course.
3. To enhance student attitudes about statistical methods.
4. To enhance student attitudes about the use of computers.
5. To increase instructor communication, satisfaction, and morale.
6. To improve the quality and standardization of testing.

What might an evaluation question derived from the first goal look like? Following is an example:

Will the implementation of a computer assisted tutorial statistics program enhance student learning?

Such phraseology, while providing a general framework for collecting and examining data, still has sufficient latitude to allow the evaluator to go in a number of different directions relative to what evidence will be used to finally evaluate the question. One might use a departmental exam covering the objectives of the course or a special set of problem-solving performance tasks. Individual faculty-constructed tests might be used. Perhaps a professional organization has a basic statistics skills competency exam that could be applied. In fact, the nature of the instrumentation might be included in the evaluation question. For example:

Will the implementation of a computer assisted tutorial statistics program raise scores on a comprehensive problem solving performance exam?

There are other things we could do to the question, such as insert the word significant in front of raise, but that might imply that we are here dealing with a statistical hypothesis where as there are many other ways to evaluate data than to use mathematical models.

The second program goal suggests both another possible format for the evaluation question as well as a philosophical issue with which the evaluator must wrestle. The second goal might be stated as follows:

Will the implementation of a computer assisted tutorial statistical program reduce the dropout and failing rate by 35%?

This absolute standard of 35% can be negotiated by the program coordinator and faculty, or a systematic standard setting procedure could be used (Popham, 1990, pp. 343-368). A comparative approach could also be used:

Will the implementation of a computer assisted tutorial statistics program result in fewer dropouts and failures than found in a traditional course?

There is the immediate implication that a formal comparative study will have to be undertaken if we are to find an appropriate answer for our question. There are, of course, a number of key terms that need definition such as fewer and traditional. There also may be an implication in the form of the question for the form of the final data collection design (e.g., post-test-only control group design).

As we anticipate decision-making (see Chapter 8), we must consider the standards and criteria for developing decision-rules for our evaluation questions.

STANDARD SETTING

Evaluation is just that--the making of a value judgment. The use of criteria and standards in evaluating outcomes of an evaluation study sets it apart from most other scientific activities. As most experts note, there must be "worth determination." Historically worth has been defined in statistical terms; for example, whether data fit a particular mathematical model. Recent trends focus on involving the stakeholders and/or evaluators in the process of setting outcome-based standards; for example, 50% of the students must master 75% of the outcomes.

There are vocal opponents and proponents of standard setting, particularly as regards the determination of student competence. Opponents argue that virtually all methods of establishing standards are arbitrary and it is difficult, if not impossible, to get judges to agree on applicable standards. Proponents cite research supporting good consistency in specifying standards, particularly when there is training involved and pilot test data are available to help guide decisions. For an extensive overview of issues and research results the reader is referred to Jaegar (1989).

A useful classification scheme for organizing some 38 different standard setting methods has been proposed by Berk (1986a) and modified by Jaeger (1989). An initial dichotomy is proposed: state vs. continuum. A state model assumes that competency is an all-or-nothing state; therefore, to be categorized as a master, a perfect test performance is required of the examinee. The procedure calls for adjusting backwards from 100% (e.g., to 90%) to set the standard. The continuum models assume that mastery or competence is continuously distributed. The standard-setting task is to search for all meaningful boundaries to establish categories. Continuum models can be test-centered or examinee-centered. All standard setting methods involve making judgments. This activity is implicit in all standard setting procedures.

Table 3-1 contains a summary of three approaches to standard setting.

TABLE3-1 Summary of Categories of Standard-Setting Methods

| Category | Description | Example |
|-----------------------------|---|--|
| State | Adjustments down from 100% performance criterion are made based on judgments about fallibility of test and characteristics of examinees. | Child will have demonstrated mastery of specified knowledge, ability, or skill when s/he performs correctly 85% of time (Tyler, 1973). |
| Test-Centered Continuum | Population of judges make probability estimates about item performances of borderline or minimally competent examinees. | Minimum standards were researched based on the National Teacher Examination (Cross, Impara, Frary, Jaeger, 1984). |
| Examinee-Centered Continuum | Judges familiar with examinees categorize them (e.g., master, borderline, nonmaster), test is administered, overlap in distributions is assessed. | Second-grade basic skills tests (language arts, mathematics) were used to compare teacher judgment and examinee performance (Mills, 1983). |

Source: Jaeger (1989).

An example of the test-centered approach should help illustrate what standard setting is all about. In this case it will be Angoff's modified procedure (Angoff, 1971). Assume that an instructor wants to set a basic passing score for a midterm exam in an introductory statistics course. The exam is composed of 40 items such as the following:

SAMPLE ITEM: A student obtains a raw score of 23 in a unimodal, moderately skewed distribution of test scores with a median of 23. What would be the student's z score?

- (a) Exactly zero
- (b) Greater or less than zero depending on value of the standard deviation
- (c) Greater or less than zero depending on the direction of the skewness (answer)

A group of experts (instructors or advanced doctoral students) are asked to make judgments about each item. The judging directions were as follows:

What percentage of minimally competent introductory statistics students will correctly respond to this item?

Judges (experts) were to select from one of the following percentages:

5%, 20%, 40%, 60%, 75%, 90%, 95%

Each judge's estimates were summed, thus yielding an "expected score" for a hypothetical minimally competent student. The expected scores were then averaged across judges. This "criterion" score could then be used to evaluate individual students or as a target against which to, say, assess that new computer--based tutorial program aimed at enhancing the competencies of students in statistics classes previously discussed.

TRAINING FOR SETTING STANDARDS

It was noted previously that in order for the collective judgment approach to work effectively, some preparation and training must occur. This is particularly crucial when high-stakes evaluations (tests) are involved (e.g., setting grade promotion score standards). Popham (1987a) suggests several guidelines for preparing judges. Among the information necessary for *informed judgment* is:

1. Delineation of consequences of decision-what are potential effects on the individual and society?
2. Description of examination-if possible have decision makers take exams to assess difficulty level.
3. Provisions of information on reliability and validity of exam-in particular, questions of bias need to be addressed.
4. Overview of phase-in time for exam and system-shorter time perhaps calls for more relaxed standards.
5. Description of examinee instructional preparation for exam-was it adequate and sufficiently comprehensive?
6. Overview of audiences with interest in results of application, of standard-an examination of possible vested interests in higher or lower standards.
7. Description of experts' recommendations-formal review of test by expert groups should be part of process.

8. Overview of field-test results and actual data on subgroups should be examined.
9. Assess standard-setting time-line alternatives-standards can be elevated or lowered depending on phase-in and preparation time.
10. Inform interested audiences about the process and products of standard setting-in particular, media representatives need to be prepared.

There are legal implications for setting standards, whether for high stakes testing or evaluation programs. We live in a litigious society and one never knows when the legal eagles will swoop down on the unsuspecting public as evaluators. Mehrens and Popham (1992) have cautioned professionals about protecting themselves when establishing standards. Their suggestions are common sense: e.g., use qualified judges, train them, use a sufficiently large number of them, provide impact as performance data to them, and allow for discussion. And whatever you do-document, document, document. But as is so often the case, common sense often cannot be found.

Standard setting is a very important, complex, and sensitive task that needs to be taken seriously by policy makers and testing experts alike. If taken seriously it will require a great deal of preparation, planning, organization of data, and-above all-patience.

By asking the important questions in the right way we can help insure that our results will be used.

EVALUATION QUESTIONS AND THE UTILIZATION OF RESULTS

The time to plan for the use of evaluation results is at the beginning of the development process. The likelihood that the evaluation results will be used significantly increases if the information needs of the stakeholders are discussed and the most relevant questions are asked and answered. Common sense can do about as much as anything to help insure that evaluation results will be used. Cousins and Leithwood (1986) reviewed 65 studies covering a 15-year period related to the use of evaluation results. Their conclusions confirm what common sense would suggest, namely that the likelihood of having evaluation results used will be increased if:

1. Evaluations are appropriate in approach, methodological sophistication, and intensity;
2. The decisions to be made are perceived as significant to users and of a sort considered appropriate for the application of formally collected data;

3. Evaluation findings are consistent with the beliefs and expectations of the users;
4. Users consider the data reported in the evaluation to be credible and relevant to their problems;
5. A minimum amount of information from other sources conflicts with the results of the evaluation.

Questions, criteria, stakeholders and utilization are all important parts of the evaluation enterprise. They are significant individually and collectively. The old adage that the whole is greater than the sum of its parts is definitely true in educational evaluation.

COGITATIONS

1. Will increasing the detail and specificity of an evaluation question increase its utility?
2. How do you reconcile the evaluation questions from different stakeholders?
3. What is the nature of the relationship between program/project development, and the kind and type of evaluation question(s) to be asked?
4. How do evaluation questions interact with the use of evaluation results?
5. What are five steps an evaluator can take to help make sure that any criterial standards that are set, are legally defensible?

SUGGESTED READINGS

- Bryk, A. S. (1983). *Stakeholder-based evaluation* (New Directions for Program Evaluation, No. 17). San Francisco: Jossey-Bass.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass. See in particular Chapter 7, "Choosing Questions to Investigate."
- Morris, L. L., & Taylor Fitz-Gibbon, C. (1978). *How to deal with goals and objectives*. Beverly Hills, CA: Sage.
- Patton, M. Q. (1986). *Utilization-focused evaluation*, (2nd ed.). Beverly Hills, CA: Sage. See Chapter 4, "Focusing Evaluation Questions," and Chapter 5, "Beyond the Goals Clarification Game."
- Scriven, M. (1974). Pros and cons about goal-free evaluation. In W.J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley: McCutchan.

EVALUATION METAPHORS

According to Perloff, Perloff, and Sussna (1976), the first recorded instance of evaluation took place in the Garden of Eden, and involved man, woman, and serpent. Apparently, the program objectives were not being met. This was an illustration of an individual evaluation. At the publicly sponsored program level perhaps the first recorded example was the evaluation that Pharaoh undertook relative to the ratio of costs to benefits when using the Hebrew laborers. The cost in the form of plagues was too high; the program, terminated.

There are, unfortunately, many school personnel who feel that evaluation is in fact a plague. Fear, threat, and anxiety surround some of the evaluations that take place in and around our schools. We feel that we cannot do a professional job if we do not evaluate students, but what about evaluating ourselves and our programs? Well-conceived evaluation programs can add a great deal of relevant information to the data needed for treating the ills of education and making good and rational educational decisions (and we need lots of those). Effective and efficient prescriptions are often difficult to find and, like health care in general, can be costly.

The foregoing two paragraphs contained two metaphorical referents, one biblical and one epistemological. Their use here hopefully enhanced understanding and comprehension of what the author was trying to say. Metaphors stimulate creativity and evaluations, and evaluators need that.

One of the exciting things about metaphors is that they can be used both as models or paradigms as well as evaluation data in and of themselves. Both uses of metaphors will be illustrated in this chapter.

Some metaphorical thinking was involved in creating the following evaluation models proposed by Wolf (1969).

Cosmetic method. You examine the program and if it looks good it is good. Does everybody look busy? The key is attractive and full bulletin boards covered with pictures and pamphlets emanating from the project.

Cardiac method. No matter what the data say, you know in your heart that the program was a success. It is similar to the use in medical research of subclinical findings.

Colloquial method. After a brief meeting, preferably at a local watering hole, a group of project staff members conclude that success was achieved, and no one can refute a group decision.

Curricular method. A successful program is one that can be installed with the least disruption of the ongoing school program. Programs that are truly different are to be eschewed at all costs.

Computational method. If you have to have data, analyze the hell out of it. No matter the nature of the statistics, use the most sophisticated multivariate regression discontinuity procedures known to humans.

In reality, unfortunately, these methods are quite popular, due in part to the complex nature of high quality education, being labor-intensive and often expensive. In addition, the press of political considerations will often preclude the conduct of rigorous "scientific" evaluation.

Metaphors should lead to models or designs. It is frequently helpful to formalize a somewhat complex process such as program evaluation into a model or develop a theory. Such a model will sometimes take the form of some conceptual paradigm, flowchart, or other schematic. There are probably as many different theories and models about program evaluation as there are authorities writing on the topic. Everyone has his or her own particular emphases, biases, and idiosyncrasies. Educators love theories and models. To see everything neat and clearly laid out gives one a sense of security. Unfortunately most program models are only first approximations to the real world, and any true similarity is very often a coincidence. Nevertheless, broad-scope outlines can help us differentiate or evaluate intents and identify potentially useful approaches. The value of these abstract representations rests on their usefulness in defining activities, examining relationships among the components or activities, and pointing toward new applications or research. Alkin (1969) has noted some characteristics of evaluation theories.

A theory of evaluation should: (1) offer a conceptual scheme by which evaluation areas or problems are classified; (2) define the strategies including kind of data, and means of analysis and reporting appropriate to each of the areas of conceptual scheme; (3) provide systems of generalizations about the use of various evaluation procedures and techniques and their appropriateness to evaluation areas or problems. (p. 2)

In general, a model based on a theory will aid in planning and implementing an evaluation program or system. On the other hand, over-reliance on an evaluation model can result in the routinization of what should be an ever-changing process. Such a danger is particularly acute if the evaluation model already has been well institutionalized. It is doubtful whether there are any real evaluation theories. Certain kinds of metaphors can help us "think through" an evaluation program and allow for better visualization of a framework through the mind's eye.

THE NATURE OF METAPHORS IN EDUCATION AND EVALUATION

A metaphor is one thing or act being implied when another is meant. It is one of the most serviceable figures in language. Take, for example, a common experience that most have lived through: "Poverty is the banana skin on the doorstep of romance."

This bit of wisdom from P. G. Wodehouse communicates an often encountered problem surrounding the place of financial security as it can hinder the beginnings of love and marriage. How often have we yearned for that extra few dollars to make an outing with a date or a dinner with a new loved one on a special occasion? Money can't buy happiness, but perhaps it can provide the wherewithal for the opportunity for one to experience it.

The value of the metaphor is not limited to its use in literature. To the extent that metaphors allow us to express ideas in less literal ways, they can help us conceptualize a process or product. The richness and vitality of words can help us "see" what is known but difficult to express or that which is felt and also difficult to express. Any tool that can facilitate communication within the educational setting is most welcome.

Before considering the metaphor as a framework for conducting evaluations, let us present some dramatic prose that uses the metaphor as a source of data useful to an evaluator as she assesses the impact of a program. The examples are taken from a paper by Hallie Preskill (1991) which describes the use of metaphors in her evaluation of the differentially staffed Saturn School of Tomorrow in St. Paul, Minnesota. This new from the ground-up school (grades 4-8) is so named because of the management concepts borrowed from the General Motors Saturn automobile plant. The Saturn School curriculum is based on the premise that learning should be student-driven and that students should participate in decisions related to *what* they learn, *how* they learn it, *when* they will learn, and *what* will be the characteristics of the learning activities that they will help to develop. Portfolio assessment is used, and no letter or numerical grades are employed. Student progress is documented in a personal growth plan. Following are three excerpts that catch the flavor of the first two years of implementation derived from 100 focus group interviews and 350 hours of observation:

We're like the Super Orient Express train traveling at 500 m.p.h. As we go along, the environment changes....the climate changes too....Some people get left off (we're moving so quickly). Some of the people who manage to stay on are held on by outreached hands and ropes that are thrown out to them....We've had a couple of people fall off. We haven't thrown anyone off yet....When people enter the train, they expect to see a dining

car...either they adapt to what we are or they go to the back and jump off. We're constantly redecorating the inside of the train....There's so much information coming at us....also traveling the same speed as the train....we can only pay attention to the information that is thrown up in the air and hits us in the face...some hits, some misses. If the train stopped, it would fall apart, we need to keep it moving, keep the energy. If we did stop, we'd end up looking like everything else.

During the school's second year, this metaphor was revisited. The teacher added,

We're still on the train but it's being remodeled in motion...taking and throwing out the old pieces and putting in new pieces and waiting for whoever is actually driving the train to identify themselves and say "Here I am, I will do this and this. "...people feel chaotic....It's still growing so fast, the changes around here are still happening fast and I think that it's going to slow down....but it continues to charge forward. We're on this journey and it's powerful, important....now on our journey we're looking at each other and fighting inside the boat, and we'll get caught in the current and we have to get the oars back in the water. We need to focus again on our common mission/vision and the extent to which we can paddle together....

The travel metaphor (train, journey, boat) is quite vivid. One can almost see the teachers, students, and administrators working together to create a map. The uncertainties, anxieties, fears, and frustrations are captured in the prose. The metaphor helps make the intangible (but something felt and experienced) almost become tangible and overt. These words sing to our imaginations. How can the concept of metaphors be used to help us create evaluation design? Following are four metaphors that have evolved into four general approaches to programmed project evaluations.

THE MANAGEMENT EVALUATION METAPHOR

There are six generally acknowledged school management functions: collecting information, planning, communicating, decision-making, implementing, and evaluating. Different educational administrators will obviously emphasize these six functions differently depending on the nature of the operation, available resources, and organizational structure. At some time, however, all functions come into play. At times one or two functions are dominant over the other. The characterization of management as an evaluation metaphor is predicated on the fact that (1) evaluation is decision--oriented, and (2) the major management functions are also included in the

evaluation process: e.g., both management and evaluation require goal identification and clarification, data collection, communication, and so on. As used here evaluation is viewed as focusing on decision-making facilitation. A prime proponent of this metaphor, although most refer to "models" rather than metaphor, is Daniel Stufflebeam, who along with Egon Guba developed the CIPP approach.

The CIPP lives not only in the world of acronyms and the minds of our countrymen but also in real life. The CIPP elements stand for four types of evaluation: C = context, I = input, P = process, and P = product. The definitive discussion of the CIPP model can be found in the book commissioned by Phi Delta Kappa (PDK) and authored by Stufflebeam and others (1971). For an abbreviated presentation see Stufflebeam (1983). In the PDK volume evaluation was defined as the "process of delineating, obtaining and providing useful information for judging decision alternatives." Note that the basic processes or functions of collecting, organizing, analyzing, and reporting are included in this working definition as they were described in the list of evaluation activities in Table 1-1. It is obvious that these activities would be part of any evaluation effort and simply points to commonalities among approaches. But what are the CIPP types of evaluation?

Context Evaluation: Under this heading, evaluation refers to activities undertaken during program planning aimed at defining need and the situation. In a real sense it is not truly evaluation because formal assessments of merit are not the primary focus. Needs assessments so prevalent in public education would fit well under this rubric. Efforts lead to specification or classification of goals and objectives. A major mode in context evaluation is the identification of the congruence between intended and actual operation. Development of a relevant data base is essential. A sample Context evaluation question might be: "What proportion of seventh grade students are reading at grade level?"

Input Evaluation: Procedures used in the name of input evaluation are aimed at identifying and assessing the capabilities of the proposed program or project and resources to address the "need" identified as part of context evaluations. The end product of this evaluation is a summary of alternative designs. Concern is with the dimensions of cost, benefit, and implementation time, what and how barriers are to be confronted, and an assessment of the overall design relative to total program goals. A sample input evaluation question might be: "What are the relative advantages and disadvantages of the Pappas, Durham, and Lynn techniques for teaching basic computational skills to at-risk elementary students?"

Process Evaluation: The focus here is primarily on implementation and a description of what goes on in the program. The overall strategy is to identify and monitor on a continuous basis various elements of program operations. Feedback to managers is critical, particularly with regard to personnel and materials. A sample process evaluation question might be: "Are teachers in School A following the prescribed procedure in using the new math methods?"

Product Evaluation: Here concern is with assessing general and specific outcomes. The CIPP framework is best characterized as an objectives-based model where the intent is to provide the decision maker(s) with as much relevant data as possible. Also treated here would be questions related to the degree to which context objectives had been met. A sample product evaluation question might be: "Are scores on the eighth grade hygiene test given in the spring meaningfully higher than those obtained in the fall?"

To help the reader gain some perspective on the CIPP model, a summary table has been prepared (Table 4-1), based on an audiotape presentation by Daniel Stufflebeam. The reader's attention is drawn to the kinds of questions raised in each cell. A comment about the two left-hand dimensions is in order.

The *goal* of evaluation—to determine the merit of some procedure, program, project, process, or product—is generally considered to be the same, no matter what the context. The *role* that evaluation may play, as noted in Chapter 1, may vary depending on the timing of the evaluation and the reason for collecting the data. Formative evaluation refers to assessments that are undertaken during the implementation of an ongoing project or program. If one were developing curriculum materials, formative evaluation might take place at several stages to check on the adequacy of developmental process and seek answers to questions related to usability, responsiveness to objectives, etc. A summative, or terminal, evaluation would focus on the end of the program or project accomplishments and the end-of-course achievement or final status of product. Data used formatively are applied by decision makers in adjusting program elements or procedures so that the desired outcomes will be obtained. For Stufflebeam, data used summatively are used for accountability purposes, the intent being to check on whether what did in fact happen was what was supposed to happen.

What does it feel like to CIPP? Following is a brief illustration of an attempt to use CIPP to organize a multidimensional evaluation.

TABLE 4-1 Overview of the CIPP Evaluation Model

| CONTEXT (Goals) | INPUT (Design) |
|--|--|
| DECISION MAKING | |
| <ol style="list-style-type: none"> 1. What needs are to be served? 2. What problems need to be solved in meeting needs? 3. What funding or other kinds of opportunities that might be used in solving problems or meeting needs are available? | <ol style="list-style-type: none"> 1. What procedural design should be chosen to achieve chosen objectives? 2. What kind of proposal to funding agency ought to be written? What are cost-effectiveness possibilities? |
| ACCOUNTABILITY | |
| <ol style="list-style-type: none"> 1. What goals were chosen when program was initiated? 2. Why were these goals chosen over other possibilities? | <ol style="list-style-type: none"> 1. What designs were proposed? 2. What alternative designs were rejected? 3. Why was the winning design chosen? |
| PROCESS (Activities) | |
| DECISION MAKING | |
| <ol style="list-style-type: none"> 1. Is the design being implemented as intended? 2. What are flaws in the design? 3. Has staff been adequately oriented and trained? 4. Is the staff supportive of program goals and design? 5. Do staff members know how to implement their roles? 6. Are there any particular procedural problems? | <ol style="list-style-type: none"> 1. What interim and final products were developed? 2. Is the program solving the problems it was designed to solve? 3. Are unanticipated effects produced by treatment identified? Seek answers related to questions as to whether to continue project, to recycle for another year, or to expand to broader population. |
| ACCOUNTABILITY | |
| <ol style="list-style-type: none"> 1. Record of actual treatment conducted--what types of treatment produced what kind of outcomes? 2. What decisions were made in changing treatment in project design so those who want to replicate can do so? 3. What steps were taken in helping people implement project design? | <ol style="list-style-type: none"> 1. What was overall outcome achieved by the program? 2. What were the side effects? 3. To what extent can we make inferences about what treatments actually produced the observed effects? 4. How valuable were the results from the project? 5. How cost effective were they in comparison to results produced by competing projects? |

Source: Adapted from Daniel Stufflebeam, *A Conceptualization of Evaluation*. Audiotape C2, American Educational Research Association, 1971.

AN ILLUSTRATION OF A CIPP EVALUATION: EVALUATING A GIFTED AND TALENTED PROGRAM

In the following case study the CIPP framework was used to organize, structure, delimit, guide, and manage the evaluation of an eight-week summer enrichment experience for 400 rising junior and senior high school students who had been identified as being artistically or academically talented. The program was held on the campus of a Southeastern college. The program (Governor's Honors Program, GHP) was targeted for evaluation in hopes of yielding data useful for evaluating the major goals of the program which were to (1) provide an enriching cognitive environment for academically and artistically talented students, (2) assist in developing appropriate teaching techniques, (3) assist in developing appropriate counseling techniques for the gifted, and (4) develop a research base for studying gifted and talented youth. It was hoped particularly that data could be gathered which would assist in faculty and student selection. Areas of the program investigated included the (1) nature and effectiveness of the instructional experiences, (2) post-program achievements of students, and (3) personality and life-history characteristics of attendees.

Each public and private high school was allowed to nominate a student in each of eight areas: art, drama, English, foreign language (French or Spanish), mathematics, music, science (physics, biology, chemistry), and social science. Vocational areas are also now included. Cut-off scores on a standardized academic aptitude test were lower for artistically talented youth.

Nominees ($n=3,800$) took a screening test, yielding approximately 1,100 semi-finalists, who were then interviewed. Based on procedures that varied from nomination area to nomination area, 400 finalists and some alternates were selected. Table 4-2 contains an overview of some of the evaluation activities associated with each element in CIPP. With regard to *Context*, one can see that the activities are mainly descriptive-this is an important function of evaluation if replication of the project or program is contemplated. For example, in GHP the selection process had never been documented. It's difficult to evaluate a process that has never been described. The reader will also note that ratings were obtained from students relative to the congruence or compatibility of their personal goals in attending the program relative to objectives generated by the faculty. The objectives of the area programs had not been previously documented. This is a good example of the kinds of contributions that an internal evaluator can make to a program or project.

The semantic differential used for *Input* included only "evaluative" adjective pairs. Stimulus concepts such as Governor's Honors Program, Academically Talented Students, Learning, and Dormitory Living were used.

TABLE 4-2 Sample CIPP Activities Associated with the Evaluation of Summer Enrichment Program for Gifted and Talented Youth

| Context | Input | Process | Product |
|---|--|---|---|
| <p>1. Examination of enabling legislation</p> <p>2. Description of selection procedures for each nomination area</p> <p>3. Student ratings of congruence of their personal objectives relative to those of their program area</p> | <p>1. Semantic differential given to both faculty and students</p> <p>2. Measure of creative personality given to students (Torrance's <u>What Kind of Person Are You</u>)</p> <p>3. Students took Cattell's <u>Sixteen Personality Factor Questionnaire</u></p> <p>4. Objective biodata form for students</p> <p>5. Faculty responded to measure of classroom management philosophy (<u>Pupil Control Ideology</u>)</p> | <p>1. Audio tapes of instructional sessions. These data subjected to Ober interaction analysis</p> <p>2. Administration of <u>Classroom Activities Questionnaire</u> which allowed students to rate nature of instructional experience (Based on <u>Taxonomy of Educational Objectives</u>)</p> | <p>1. Student ratings of self vs. program contribution to extent of mastery of program objectives</p> <p>2. Extensive follow-up survey of previous 10 years' worth of participants</p> <p>3. Post-semantic differential</p> |

The last was one of the most positively evaluated concepts based on pre--(mid-summer) vs. end--of--summer data. One measure of program impact was index by convergence of faculty and student semantic differentials.

Classroom instruction process tapes were volunteered for use in assessing *Process* and therefore obviously were not representative but nevertheless suggestive of the teacher's approach. These process data proved to be of particular interest when contrasted with the "classroom management" data collected as part of the input evaluation.

There were no commercially available standardized tests that could be used to evaluate *Product* outcomes nor was there time to develop any. Lack of readily available instrumentation calls for creativity. It was therefore felt that program's impact could be judged by examining student ratings of (1) their progress toward mastering their area objectives, and (2) the extent to which the program (vs. the student him\herself) had contributed to that mastery.

The evaluation was carried out over a six-month period, with a budget of \$15,000 and a staff of one full-time director, one full-time data-collection coordinator, one full-time data analyst, a half-time graduate assistant, and a half-time secretary. The project came in under budget by \$2,500 and the report was three weeks ahead of schedule. Such an accomplishment deserved at least a nomination for Most Efficient Project of the Year.

Some of the advantages of the CIPP model are outlined in Table 4-3. The main advantage of CIPP is its comprehensive framework which makes it easy to organize evaluation activities. Conversely it may be too structured for some evaluation tasks. Other metaphors that may come under the "management" rubric are the Discrepancy model (Provus, 1971; Steinmetz, 1976, 1977) and school-based models offered by Metfessel and Michael (1967) and Hammond (1972). The discrepancy idea is particularly attractive to many evaluators since it tends to emphasize the congruence between performance data and standards.

THE JUDICIAL METAPHOR

The legal system has also provided some ideas useful in developing an evaluation metaphor. The ideas of a judge, attorneys for the plaintiff and defendant, and blind justice weighing the veracity of evidence appeal not only to our sense of fairness but perhaps also a love of the dramatic. The most frequently cited evaluation mutation generated by the ideas of the law is the so-called adversary model. It takes different forms from a very structured court model to procedures embodied in congressional hearings (Kourilsky, 1973; Owens, 1973; Thurston, 1978; Wolf, 1979; Worthen & Owens, 1978). Although not representing a comprehensive, homogeneous, and integrated model per se, the adversary approach rests on the basic notions of systematic

TABLE 4-3 Advantages and Disadvantages of the CIPP Metaphor

| ADVANTAGES | DISADVANTAGES |
|---|--|
| <ol style="list-style-type: none"> 1. Comprehensive-is responsive to intents. 2. Each one of parts can be undertaken while waiting for product. 3. Meets needs of decision makers, administration, and managers. 4. Provides structure for focusing on evaluation tasks and questions. 5. Provides flexible framework. | <ol style="list-style-type: none"> 1. Too much structure may cause a variety of tunnel vision and miss unintended outcomes. 2. Can be complex and costly if fully implemented. 3. All decisions may not be able to be specified in advance. |

presentations, examination, and cross-examination found in legal proceedings. Central to the approach is an arbiter who may be a judge, a group of judges, or several decision makers. At least two evaluators (one proponent, one opponent) engage in dialogue, discussion, and debate. Their presentations and evidence are examined by each other and the arbiter. The arbiter also cross-examines. One important aspect of most adversarial proceedings is a public presentation. Such a requirement can lift the perceived veil of secrecy surrounding evaluation and decisions. The basic outline of the adversary model is presented in Figure 4-1 (Kourilsky, 1973).

There are a number of situations where the single recommendation or conclusion evaluation approach is inappropriate or at least less efficient and viable than a more complex and interactive model. The adversary approach should find applicability, however, in situations where the following conditions apply.

1. The focus is on a specific policy decision.
2. The issue(s) have significant fiscal and other resource allocation implications.
3. The decision maker would like to be involved in deliberating about alternatives.
4. There may be disagreements between expert consultants and/or evaluators and the decision maker(s) on relevance of data and interpretation.

5. The public is made up of diverse audiences and has an immediate vested interest in the outcome of the evaluation and decision.
6. The program or issue is controversial, and there is a polarization of views and values.

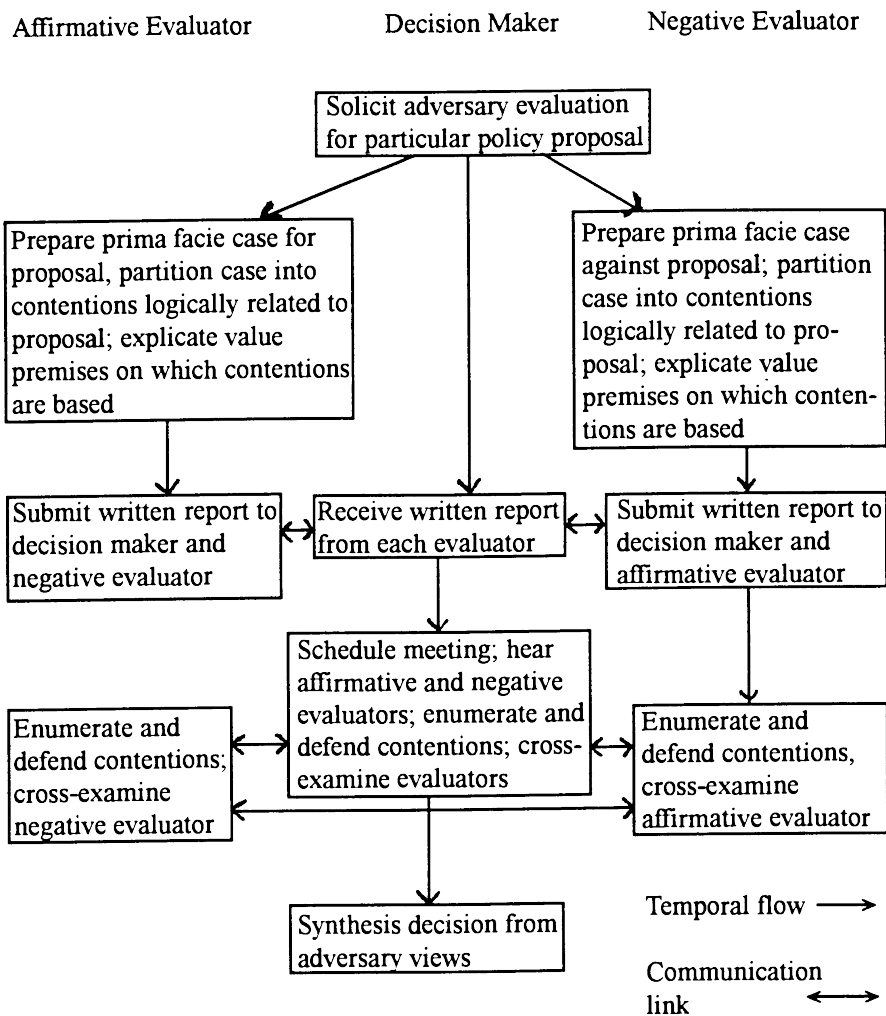


Figure 4-1 Representation of Adversary Evaluation Metaphor
 Reprinted by Permission of Copyright Holder: Center for
 Research on Evaluation, Standards and Student Testing.
 University of California at Los Angeles. Source: Kourilsky,
 1973.

In attempting to balance possible sources of bias, the needs and interests of a variety of stakeholders can be considered. These stakeholders may be educators and administrators at all levels, students, parents, teachers, taxpayers, and community groups. With increasing frequency business and

industry are becoming involved with the educational enterprise. They obviously have a vested interest in the quality of the output from our educational systems, since they are the consumers (employers). The adversary metaphor may be a useful vehicle for these groups to have input into the making of important policy decisions which may have significant price tags attached or reflect directly on the acquisition of employable skills.

As is the case with any model, there are advantages and disadvantages in using the particular approach. Table 4-4 contains a summary of these advantages and disadvantages for the judicial metaphor. Particularly helpful in developing this table were the writings of Popham and Carlson (1977), and Owens (1973), and Thurston (1978). Let there be no doubt that the approach has some potentially serious shortcomings. The adversary approach is not the final word in evaluation models. It does not have applicability in a great variety of situations but does seem particularly timely in this day and age of accountability. Public disclosure of evaluation data and decisions could go a long way toward allaying society's fear, anxiety, and hostility about public education.

TABLE 4-4 Advantages and Disadvantages of the Judicial Evaluation Metaphor

| ADVANTAGES | DISADVANTAGES |
|--|---|
| <ol style="list-style-type: none"> 1. Variety of data may be introduced 2. Tends to force higher quality of evidence 3. Can operate to diminish unwitting bias 4. Variety of points of view and opinions presented 5. Opportunity to examine opposing data and views 6. Tends to force assumptions to surface 7. Applicable to situations having complex outcomes 8. Facilitates communication | <ol style="list-style-type: none"> 1. Need for equally able and motivated adversaries 2. Expensive method 3. Heavy burden on arbiter/judge/decision maker 4. No control over decision maker having hidden agenda 5. Not all propositions amenable to adversarial approach 6. Judicial model can generate false confidence 7. Possibility of extremism 8. Can deteriorate to courtroom melodrama with emphasis on who wins |

Be aware of another problem in using the judicial metaphor. Thou shalt not bear false witness! In biblical times contracts, covenants, or business promises were sealed by the calling of at least two witnesses to the agreement. All witnesses, before testimony was given if there was a dispute, were

cautioned to tell nothing but the truth and to conceal nothing that was pertinent to the case. It was a sin for a witness to withhold evidence in his possession (Leviticus 5:1, Proverbs 29:24). If false witness was detected it drew the same penalty upon the false witness as the accused. Ancient judges didn't fool around! The point was that the veracity of both judge/decision maker and witnesses is crucial to the effectiveness of this metaphor.

THE ANTHROPOLOGICAL METAPHOR

Anthropologists have been variously described as scientists who investigate humanity and human culture. They examine strategies for living that are learned and shared by people or members of living groups. They follow general procedures that involve (1) entering and establishing themselves in a community, (2) developing hypotheses, (3) collecting and synthesizing evidence, and (4) drawing conclusions. If one conceives of a classroom, school, or school system as a "culture," then the anthropological approach to investigation emerges as a metaphor for evaluation. Ethnography has emerged in the last several decades as an extremely valuable tool for the educational anthropologist turned evaluator. Ethnography being the documentation and description of social and cultural groups (Fetterman, 1984).

It is difficult to identify a single approach to represent the anthropological metaphor. Throughout this section the descriptions/terms *anthropological*, *qualitative*, *responsive*, *goal-free*, and *naturalistic* will be used interchangeably. One might also engage Stake's description and judgment matrices in his "countenance" model (1967), or the Lincoln and Guba naturalistic approach (1985) in the consideration of qualitative approaches. Further explorations might lead to Scriven's so-called "goal-free" approach (1972, 1973). Let's stop a moment and consider the nature of goal-freeness as it reflects significant philosophical images of the anthropological metaphor.

We begin with the premise that if one were interested in what impact a project or program has had (intended and unintended), one should look at the outcomes of the program. Scriven (1972, 1973) has posited a goal-free model. Using this orientation, an evaluator does not begin with the rhetoric of the project or program, but rather focuses attention on results. There is an assumption that beginning from a goal or objectives base may result in tunnel vision for an evaluator. In addition to the increased likelihood of identifying unanticipated or side effects, goal-free evaluation also concerns itself with an assessment of the quality of program goals and objectives themselves. Obviously if goals or objectives are not worthwhile their attainment would not be meaningful. Goal-free evaluation is sometimes referred to as "responsive evaluation" and goal- or objectives-based evaluation as "preordinate evaluation" (Stake, 1976). Schermerhorn and Williams (1979)

reported a study that attempted to compare indirectly the effectiveness of these two general approaches to evaluation. Preordinate evaluation was characterized by an emphasis on prespecified intents, expected outcomes, already established criteria for success, and the use of standardized data-gathering instruments and procedures. Responsive evaluation, on the other hand, was described as methodologically fluid, relying heavily on observation data and providing descriptions of activities rather than intents. The two approaches were contrasted by having a panel rate evaluation reports generated by the two approaches across the following three dimensions: interest, value of information, and efficiency. The study found that the case study developed using the responsive evaluation format was favored relative to the three dimensions. In addition, the study found, not unsurprisingly, that the case-study approach was much more expensive. The researchers concluded, however, that the responsive technique should be used to supplement more goal-oriented evaluations.

The implication of the anthropological metaphor is perhaps not so much for the overall design of the evaluation as it is for the activities that occupy the evaluator's efforts and time. Patton (1987, p. 7), for example, notes that a qualitative/naturalistic evaluation would be concerned with:

- Describing the program or project implementation in detail;
- Analyzing program or project processes;
- Describing participants and the nature of their participation;
- Describing program impact cognitively, affectively, and behaviorally;
- Analyzing strengths and weaknesses of the program or project based on a variety of data and sources.

These general intents have been operationalized in the responsive evaluation approach advocated by Robert E. Stake (1975, 1983). Although likely to result in some decrement in measurement precision, the validity and usefulness of a responsive evaluation more than justifies its application. A responsive evaluator is concerned with producing a "product" or what Stake calls a *portrayal*. A portrayal is a verbally rich description of the program or project reflecting multiple realities that the evaluator has experienced. The key to the gaining of the data needed to generate a portrayal is the use of qualitative methods in naturally occurring situations. The closer the data are to the source in context, the more meaningful the judgments. Most experts expect the evaluator to make judgments of value, worth and merit, and not rely on some individual outside the environment to interpret the data. Among the methods a responsive/naturalistic/qualitative evaluator might use are: ethnography, case study, investigative journalism (another interesting

metaphor), oral history, participant (interactive) or nonparticipant (noninteractive) observation, field study, or connoisseurship/criticism (Eisner, 1976, 1991, 1992). Eisner's interesting ideas, which many have likened to art and literary criticism, rest on expert judgments.

Stake's ideas have been gathered together in the "event clock" contained in Figure 4-2. Central to the responsive evaluation approach is observation and feedback, with the cycle repeated as many times as necessary. Stake notes that the clock in Figure 4-2 may operate in the usual and expected clockwise direction, or it may run counterclockwise or even cross-clockwise. This, of course, further confuses the already disoriented evaluator. Although the language is different, the "events" of Figure 4-2 are comparable to the "steps" of Figure 1-1.

Responsive evaluations probably are more complex and cognitively and emotionally more exhausting than traditional, more quantitative assessments. It's very easy to sit back and punch a bunch of test scores into a computer, run a standard statistical analysis package, and table some means and F-values. The human dimension is demanding both as reflected in focusing data sources and evaluator-as-observer-data-collector. As opposed to the quantitative (objectives-oriented "preordinate") evaluator, a qualitative/responsive/naturalistic evaluator is likely to spend (1) less time in instrument development, (2) much more time observing the program or project and gathering judgments, and (3) much less time formally processing data, although sometime qualitative (such as that derived from observation, open-ended questionnaires, or interviews) will be subjected to extensive and time-consuming content analyses. The results of these analysis will sometimes be summarized with frequencies and percents for graphic display.

The next section contains an attempt to use some of the "responsive" ideas of the present discussion in evaluating a new, locally developed teacher evaluation system. But before baring that soul, let us summarize the advantages and disadvantages of the responsive/qualitative/naturalistic anthropological metaphor, collected in Table 4-5. A major appeal of the metaphor is the closeness of data (from observation) and interpretation (judgment of worth), thereby enhancing validity. When using structured paper-and-pencil devices, one frequently has the feeling that the leap from marks on the paper to "true" meaning is almost one of total faith. Qualitative methods tend to yield higher inference data, i.e., more subjectivity is involved in interpreting them. Do not overlook the loss of measurement precision, however, that can also accompany the use of qualitative methods. The "human instrument" can be unpredictable and unreliable.

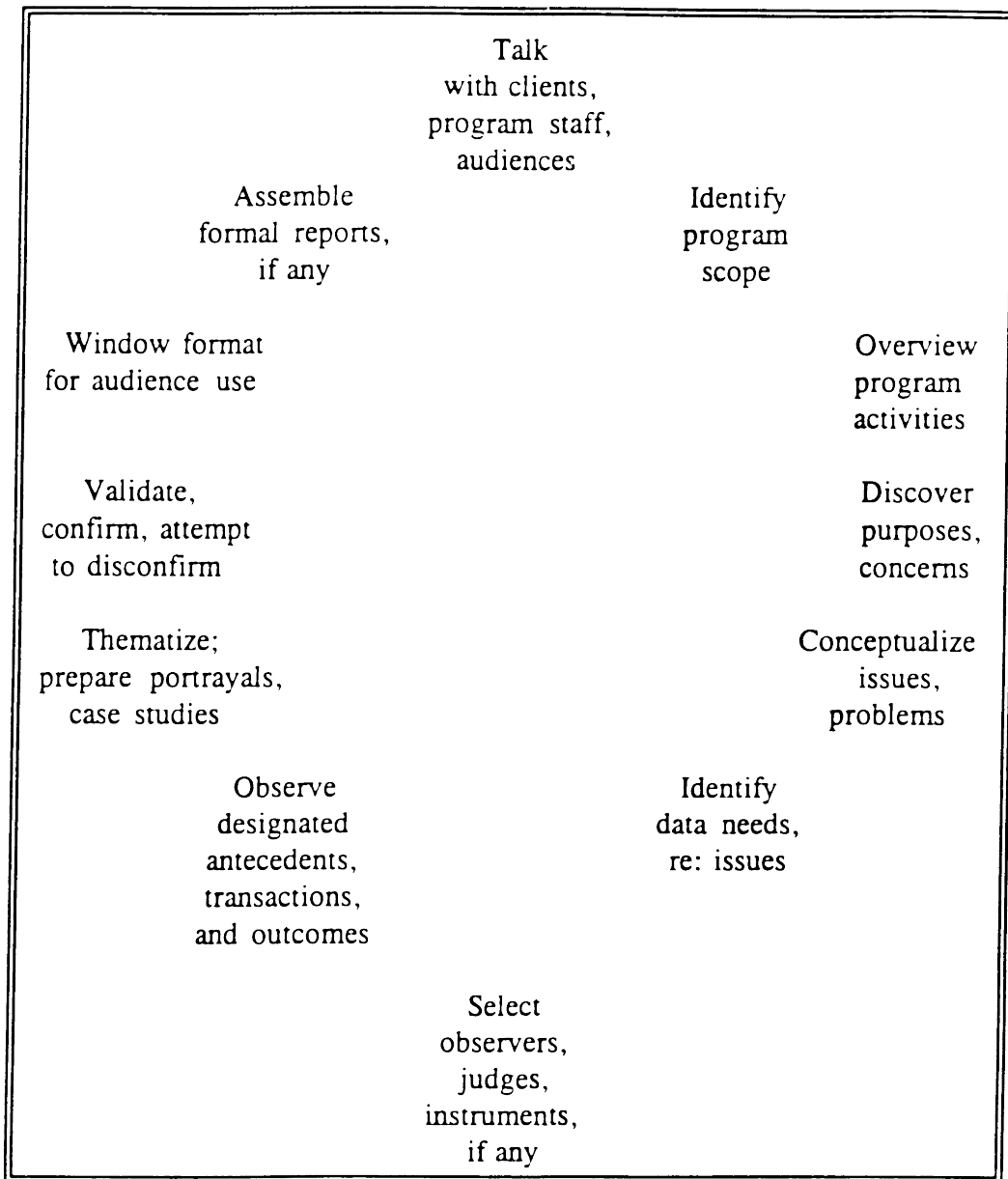


Figure 4-2 Events in Responsive Evaluations (Stake, 1983; reprinted by permission of author)

Following is an example of an attempt to use the "responsive" philosophy to evaluate a locally developed teacher evaluation system.

TABLE 4-5 Advantages and Disadvantages of Anthropological Metaphor

| ADVANTAGES | DISADVANTAGES |
|---|--|
| <ol style="list-style-type: none"> 1. Potentially greater validity 2. Greater responsiveness to stakeholders within context 3. Great heuristic value and likelihood of new insights 4. Encourages multiple data types and sources 5. Less likely to miss unintended effects 6. Nature of data is inherently credible and persuasive 7. High degree of flexibility 8. Emphasizes real and complex nature of evaluation context | <ol style="list-style-type: none"> 1. High degree of reliance on subjectivity 2. Problems or reliability of human observers 3. Data collection and analysis likely to be labor intensive 4. Potentially very expensive |

**AN ATTEMPT AT RESPONSIVE EVALUATION:
EVALUATING A TEACHER EVALUATION SYSTEM**

Following is a description of an application of Stake's "responsive" philosophy in evaluating a newly created teacher assessment system. Efforts were made to produce descriptions (Payne & Hulme, 1988) that would be maximally useful to the stakeholders, local teachers, and administrators. Qualitative data collection methods were used primarily to solicit views about various important dimensions of the new program, with a focus on how to improve the system. We have here, hopefully, an ethnography of a developmental project. An attempt was also made to triangulate data collection where feasible (use multiple data sources relative to the same variable).

Teacher evaluation is a powerful tool that can result in significant improvement in student learning and school climate. If managed poorly, however, it can lead to devisiveness, increased anxiety, and "evaluation-fear," and possibly the destruction of teacher morale. The evaluation of a new teacher evaluation system, therefore, provides a tremendous opportunity to generate data for formative applications aimed at improvement and the medication of instructional ills. (How's that for a metaphor?)

The Grass-roots Teacher Evaluation System

Authorities have identified several teacher evaluation systems (McGreal, 1983; Darling-Hammond, Wise, & Pease, 1983). These range from the highly structured (Medley, Coker, & Soar, 1984) to the artistic and almost mystical (Eisner, 1982). The system described here was developed from a clinical supervision perspective. It emphasized the following activities:

- Preobservation conference
- Observation of teaching (short and extended)
- Feedback and analysis
- Goal-setting
- Observation of teaching
- Post-observation conference and evaluation.

The term *grass-roots* is used here to describe the development of the system. That catch phrase is used intentionally, since the design, development, and implementation of the system comprised a total effort where in all system educators were represented and/or had direct input. The intent was to develop a system that would meet the following purposes: (1) accountability; (2) improvement of instructional effectiveness; (3) encouragement of professional growth; (4) collaboration; (5) planning; and (6) corroboration of employment decisions.

A committee of 17 teachers and 9 administrative personnel developed the evaluation procedures and instrumentation. The total evaluation system included assessments of counselors and media personnel in addition to teachers. Only data on teachers will be presented in this report. The system involved a three-year cycle for each teacher that included orientation, assessment, and evaluation phases. The assessment phase included both long-term and short-term classroom observations. The evaluation phase was only for end-of-cycle teachers.

The pilot implementation also involved: (1) workshops with leadership personnel, particularly principals, aimed at enhancing conferencing and observation skills; (2) the refinement of a generic teaching model based on teacher competencies; (3) publication of a newsletter for teachers, Keeping Informed on Teacher Evaluation (KITE); and (4) central office meetings with outside consultants to refine the system. Teachers could develop goal plans for the year and present data from a variety of sources to support their performance evaluations. The major theme of the system was "improvement through both formal and informal staff development." The system was high-inference and judgmental as suggested by Popham (1987b).

The Setting

The pilot project took place in a fast-growing Southern community (bedroom for Atlanta) where (1) student enrollment was almost 50,000, (2) there were almost 3,000 teachers on staff, and (3) the per-pupil expenditure was \$2,458 a year. Four schools were involved in the implementation: an elementary (n=83), middle (n=60), high (n=60), and vocational (n=11), with a total of 214 teachers.

Instrumentation

The following are considered to be the psychometric lawnmowers used to trim what had evolved from the grass roots.

Administrator activity log

All principals, assistant principals, and, where applicable, leader teachers were requested to maintain daily logs of their relevant activities and the amount of time spent in each activity. The logs were summarized weekly over four seven-week blocks. Content analyses of the logs were undertaken and fed back to principals.

Teacher assessment instrument

Teachers and principals responded to an eight-scale summary instrument in October and again in May. Each scale represented a critical teacher activity. The eight scales were as follows: Knowledge of Subject, Planning, Implementing, Evaluating, Classroom Management, Professional Growth, Professional Reliabilities, and Interpersonal Skills. Judgments were made using four categories: Exceeds Expectations (E), Meets Expectations (M), Needs Improvement (N), and Unsatisfactory (U). Although global judgments were being made, each scale had two or more specific indicators to aid the evaluators in synthesizing their judgments (e.g., Implements activities in a logical sequence). No performance standards were specified for the evaluation because of the formative nature of this pilot implementation.

Teacher survey

Inasmuch as pre-project evaluation data might have sensitized the teachers to the innovation, a 30-item retrospective survey form was developed and administered at the end of the school year (Rippey, Geller & King, 1978). The response scale was Better This Year, No Difference, and Better Last Year. Following are two sample items:

The amount of anxiety I feel about being evaluated.

My involvement in the evaluation process.

Teacher interviews

In an effort to triangulate on teacher perceptiveness of the effectiveness and efficiency of the systems, four teachers were selected at random from each of the pilot schools and interviewed with a semistructured questionnaire. The content of this questionnaire was derived from the teacher survey. Five general questions guided the interviewers (nonpilot teachers) after a session about interview techniques.

Results

Evaluation Question 1: What Changes Need to be Made in the Procedures and Implementations?

Initial content analyses of administrator logs yielded four categories: Activity, Reactions, Concerns, and Suggestions. The amount of time associated with each activity was tallied for each team member in each school. It was hoped that these data would reveal how the implementation of the new teacher evaluation system impacted on the activities of, tasks of, and demands made on personnel charged with operationalizing the system. Table 4-6 contains a summary of the activity data in terms of average number of hours per week for each of the four quarters. The per-person averages are based only on the number of individuals actually reporting data for a particular activity. In the interest of brevity only the eight most time-consuming activities are reported.

TABLE4-6 Summary of Results of Content Analyses of Administrator Logs for the Activity Category (Average Hours Per Week Per Person) by Quarter

| ACTIVITY | PER PERSON AVERAGE BY QUARTER (Hours per Week) | | | |
|-----------------------------------|--|----|---|----|
| | 1 | 2 | 3 | 4 |
| 1. Meet with leadership team | 7 | 2 | 3 | - |
| 2. Meet with central office staff | 9 | 7 | 5 | 5 |
| 3. Teacher orientation | 5 | 2 | 3 | - |
| 4. Observation | 6 | 11 | 6 | 7 |
| 5. Teacher conferences | 4 | 9 | 9 | 15 |
| 6. Presentation to peers | 2 | 2 | 7 | - |
| 7. Paperwork | 4 | 6 | 6 | 8 |
| 8. Individual work | 2 | 3 | 2 | - |

It is interesting to note how the major activity changes from the first period to the last period. At the outset large amounts of time are given over to meetings with central office personnel to work on issues related to implementation of the system and how data collection requirements for the evaluation were to be met. During the second period administrators were involved with making teacher classroom observations for assessment purposes. The last two periods reflect the end product of the process, namely, teacher conferencing for purposes of communicating evaluations. It is also obvious that the aggregate amount of time involved is very large. In fact, it works out that the three major activities contributing to implementing the evaluation system (Teacher Orientation, Observation, and Teacher Conferences) required an aggregate average of almost 20 hours per week. No meaningful differences were noted between the four levels of schools. The only trend was as one would expect, that as the number of faculty increase, so do time demands. The increase was geometric rather than linear.

Content analyses of the Reactions, Concerns, and Suggestions basically followed the chronology of the implementation.

Evaluation Question 2: What Is the Impact of the Evaluation System on Communication Between Teacher and Evaluator?

Percent agreement in the use of the four evaluation categories for the October and May data points is summarized in Table 4-7. The overall percent agreement for October was 57 and in May increased to 65. Although not dramatic, the change was in the hypothesized direction. The largest single change for a competency was for Instructional Techniques-Implementing, where the input of principal observation data probably had greatest impact.

Analyses of the principal and teacher use of each of the four evaluation categories yielded some interesting results. In the fall data, the contribution to the overall 57 percent agreement came from 14 percent of the Exceeds Expectations (E) category and 43 percent from the Meets Expectations (M) categories. In the spring, the proportion changed to 25 percent for E and 40 percent for M. There was no contribution from the Needs Improvement and Unsatisfactory classifications.

Not unexpectedly teachers tended to evaluate themselves more favorably than the principals did at both data points. If the four categories are quantified and averaged (E=4, M=3, etc.), the following picture of means emerges:

| | October | May |
|---------------------|---------|------|
| Teacher self-rating | 3.45 | 3.53 |
| Principal rating | 3.17 | 3.29 |

TABLE 4-7 Percent Agreement Between Principal and Teacher Evaluations for October and May Data Points

| Teaching Competency | Percent Agreement | |
|--|-------------------|-----|
| | October | May |
| Knowledge of Subject | 67 | 69 |
| Instructional Techniques--Planning | 61 | 62 |
| Instructional Techniques--Implementing | 46 | 75 |
| Instructional Techniques--Evaluating | 60 | 75 |
| Classroom Management | 63 | 61 |
| Professional Growth | 48 | 66 |
| Professional Responsibilities | 57 | 57 |
| Interpersonal Skills | 53 | 56 |
| Total | 57 | 65 |

These data suggest an average increase in the evaluations from both groups as well as a decrease in the differences between the group means across time. The convergence is interpreted as reflecting enhanced communication between principal and teacher.

Evaluation Question 3: How Do Teachers Evaluate the Evaluation Process?

Item analyses of the teacher survey form led to the elimination of 4 of the original 30 items. The survey had a Kuder-Richardson internal consistency reliability estimates of .98. The responses (This Year, No Difference, Last Year) were converted to ratings of 3, 2, and 1, and averaged. The mean teacher survey score was 62.93 (S=11.37). This mean expressed as percent of the maximum possible is 81%. This statistic is interpreted as supporting this year's evaluation over last year's evaluation procedures.

Responses to individual teacher survey items added special insights into teacher opinions. The following three items were highest rated in terms of the "Better This Year" rating:

The Extent of My Input into the Evaluation Process (64%).

The Extent to Which I Was Able To Share Feelings with My Supervisor About My Job (60%).

The Forms Used To Summarize My Teaching Evaluation (77%).

It is obvious from an examination of the first two items that an important contribution of the new system was to provide the teacher greater active involvement and participation in the overall evaluation process. Teacher "ownership" will obviously enhance the likelihood that the system will be institutionalized. This conclusion is confirmed by qualitative data gathered from interviews. With regard to the evaluation form, an apparent conflict exists. Survey data indicate that overall the teachers liked the form, but interviewer data suggest that the use of the Exceeds Expectations, Meets Expectations, Needs Improvement, and Unsatisfactory evaluative categories were disliked.

Evaluation Question 4: What Suggestions Do Teachers Have for Improving the System?

Five open-ended question probes were used to interview 16 teachers. They were interviewed by teachers not members of their faculty. Following is a selected summary of this free-response data.

1. Describe the Usefulness of the Evaluation in Helping You Do a Better Job.

Almost all teachers were positive. They noted that the evaluation provided explicit objectives, important criteria, and structure for immediate feedback, teacher organization, and more frequent visitation. Great value was seen in providing reinforcement, confirmation, and positive input. It also provided greater self-awareness and was a great improvement over the old checklist.

2. To What Extent Did the Evaluation Experience Help You Look at the Total Teaching Process?

Most teachers were positive, noting that the process made them more conscious of their own teaching and provided well-rounded descriptions of the most important areas of teaching. For some, the process helped clarify important criteria and tied the whole process of teaching together. Several stressed that it encouraged increased dialogue between faculty and administration and among teachers.

Many teachers felt that it didn't substantially change what they did. Weaknesses were noted in that too much time was required of evaluators if they really were to do an effective job. A special education teacher noted that there was a great discrepancy between the teaching model assumed by the instrument and her actual job duties.

3. How Much Confidence Do You Have That Your Supervisor Helped You Improve as a Teacher?

Most were positive, saying that the criticism was helpful because it was constructive and that positively phrased comments increased their own self-confidence, making them want to improve continually. Comments and dialogue were more helpful than letter ratings. Several said that they had great respect for their evaluator because observations were tailored to the individual; others said increased frequency of visitations added validity to the evaluations.

Several teachers said they had confidence in their principal, but that his evaluation was not responsible for their improvement. Concerns were expressed at the secondary level that although they had high ratings their confidence in the evaluation would be strengthened if the department head's input were utilized. They noted that department heads might need training in supervision but that their subject area expertise was very important. A few teachers said that they didn't hear enough of what they were doing well. Several expressed concern that the evaluation process relied heavily on the fairness and competence of the evaluators, and they questioned that as the process spread, would all others be as qualified as this year's group? Several also expressed concern and confusion as to the role of evaluation of both assistant principals and the counselor. Especially concerning the counselor, they questioned whether her role would change since she now serves as an administrator.

4. To What Degree Did Being Evaluated Help You Set Goals for Your Teaching?

Positive and negative comments were balanced. On the positive side, the process was helpful in giving feedback on whether goals were met. Some said it gave structure for their own personal inventory and that writing formalized goals kept them on track. Others said the seven areas provided implicit goals. Many teachers said they didn't set formal goals. Some felt uncomfortable because in their competitive school situation they felt obliged to set goals; that meant they weren't truly optional. A few felt concern that it was unfair that the first time they heard of a weakness was during a formal evaluation. If they had been observed first without judgment they could have set goals to correct weaknesses, and that way the negative evaluation wouldn't have gone into their permanent record.

Summary

Although these are limited data, they do reflect a positive impact of the program, particularly when taken in concert with the quantitative data previously presented. It is obvious that the processes of supervision and

evaluation need not be irreconcilable as suggested by data from McCarty, Kaufman, and Stafford (1986). If the appropriate balance is struck between the gathering of data relevant for decision-making and that for staff improvement, a truly valuable evaluation experience can be had by all.

Epilogue

So often an external evaluator presents his or her findings, conclusions, and recommendations to a client and then hears no more about the project. It was gratifying in the present case to find that four significant actions were taken by the superintendent and central office staff as a result of the evaluation. They are as follows:

1. Due to the fact that 50% of the teacher evaluation scale was not being used and that interview data suggested a strong dislike for the scale, the rating dimension (E, M, I, and U) was eliminated from the instrument.
2. The basic evaluation instrument with its eight competencies and total of 38 indicators was retained but will be used as a basis for individualized goal-setting via a professional development plan.
3. Teacher evaluation is obviously a labor-intensive activity (see Evaluation Question 1). The data of the present study influenced school leadership personnel to establish a 1:15 supervisor-to-teacher ratio with the inclusion of peer helpers.
4. Efforts are being increased to refine a generic teaching model tied to the operational objectives-driven curriculum.

Although not a pure example of a "responsive" evaluation, it is hoped that the foregoing description captures the flavor of using "softer" data collection methods, -e.g., interviews, surveys, logs, and observation-to program impact.

A wise evaluator once said, "Reap as you have sown." In the present harvest the reaping was not too grim (and that's no fairy tale), but a more verdant product might have been gathered if better lawnmowers could have been found or created. From the initial seeding came interesting and promising growths, but as the grass grows, so do the weeds. It is frequently difficult to separate one from the other. One must be careful not to fertilize incorrectly (or overfertilize or misfertilize) as the seeding may be of discontent rather than enthusiasm. This low-budget evaluation was only partially responsive to Stufflebeam's Standards (see Chapter 2). Lack of time and resources did not allow for the development of maximally responsive instrumentation. For the lack of a good lawnmower, too much grass was lost!

THE CONSUMER METAPHOR

The role of the evaluator as a consumer surrogate, as suggested by Scriven (1974b), reflects a very strong summative philosophy. Before a consumer makes a purchase of some consequence, such as an automobile, VCR, or home, a period of comparative shopping is involved. Advantages and disadvantages are examined, perhaps weighted, and responsiveness to needs is assessed. Costs are weighed relative to benefits, both initial and maintenance. At some point an overall summative global assessment of merit is made. Such is the general approach taken in using the consumer metaphor as a framework for evaluating programs, projects, and particularly products.

Consumers of Products

If the evaluation focus is on a product, Scriven (1974a) has provided a useful checklist that could be used for evaluation purposes. Scriven identifies 13 dimensions in the evaluation of a product that need to be considered. Following is a brief list of the elements (somewhat rephrased) in Scriven's product checklist. Each item would have a scale attached to it. The reader is referred to the original document for the full scale.

1. Need. Priority given to number of individuals affected and social significance.
2. Market. Size and importance of market to be served and dissemination plan.
3. Field Trial Performance Data. Adequacy of try-out and likelihood of generalization.
4. Consumer Performance Data. Extensiveness of data on product performance for major consumer groups.
5. Comparative Performance Data. Comparison of performance data across competitors.
6. Product Performance over Time. Evidence that effects of product hold up over time.
7. Side Effects. Evidence of nature and seriousness of side effects in using product.
8. Implementation Performance Process. Provision for procedures for identifying fidelity of implementation in using product.
9. Internal Validity of Product Use. Description of nature and effectiveness of method used to establish internal validity of product.

10. Statistical Significance. Nature, appropriateness, and results of determining statistical significance.
11. Educational Significance. Documentation of variety of methods used to establish and extent of "educational meaningfulness" of product impact.
12. Cost-Effectiveness. Extent to which product is cost-effective and results of cost analyses.
13. Extended Support. Extent of support and follow-up services relative to product use, including staff development and updating.

Although Scriven suggests that such an evaluation framework (or an augmented one) could be used formatively, primary use of the approach is to produce an aggregate global product merit index that could be used to make comparative evaluations. Product evaluation profiles could be generated and standards applied. The categories are not mutually exclusive, so the reader should be aware of possible interactions such as statistical significance and internal validity. Some data collection designs lend themselves more readily to statistical analysis procedures than others. For more on product evaluation see Chapter 11.

Consumers of Programs and Projects

The consumer metaphor has been used extensively by a variety of governmental agencies ranging from the federal level to the local level. Typical of the federal use is the Program Effectiveness Panel (formerly known as the Joint Disseminating Review Panel). The intent is to solicit for review programs and projects that have already demonstrated their effectiveness. If favorably evaluated they would be entered into a network for dissemination to local systems. The school systems would, therefore, have confidence that the program or project is likely to be effective due to prior expert assessment. Programs "approved" are collected together in a publication called Educational Programs That Work (National Diffusion Network, 1993).

The Program Effectiveness Panel (60 members) does not do evaluations, but evaluates the evaluation results of programs and projects applying for inclusion into the National Diffusion Network. School systems may apply for federal dissemination monies for these "validated" projects. Judgments about a given application for recognition (limited to 15 pages) is based on three sets of criteria: Results (0 - 50 points), Evaluation Design (0 - 40 points), and Replication (0 - 10 points). An application should fall into one of four categories: (1) Academic achievement-changes in knowledge and skills; (2) Improvements in teacher attitudes and behaviors; (3) Improvements in

students' attitudes and behaviors; and (4) Improvements in instructional practices and procedures. For further elaboration of the guidelines and procedures the reader is referred to a guidelines reported by John Ralph and M. Christine Dwyer (1988).

A system similar to the Program Effectiveness Panel approach is used by the State of Georgia to help local education agencies operationalize their ideas about how to respond to local needs, but may also have implications for other schools and systems in the state. Grants ranging from several thousand to several hundred thousand dollars are awarded each year. Figure 4-3 outlines the process. It begins with the development of a concept paper that is responsive to the needs of students and schools as perceived by the State Department of Education. These "State Priorities for Educational Improvement" are used to guide state programs and resource allocations. Following are some recent state priorities:

- Enhance teacher morale and enthusiasm.
- Increase the rate of school completion by students.
- Enhance the readiness of children at-risk for entry into kindergarten.
- Develop a plan to make the individual school a community resource center.
- Develop an instructional program of continuous progress for the primary school years (K - 3).
- Develop a plan at the school building level to increase the effectiveness of that school.

Obviously there is something there for everyone. Stated another way, if what you want to do doesn't relate to one of the above categories, perhaps you shouldn't do it. Once a relevant priority has been selected at the local level, the next step (see Figure 4-3) is to create a five-page concept paper that includes the following elements:

- Evidence documenting that the state priority to be addressed by the proposed program focuses on a need or problem in the local school system.
- A general description of the approach to solving the problem identified in the priority.
- A list of the specific components included in the approach to the problem.

- A description of the characteristics of the school system (small, rural, urban, etc.) and the proposed group that will receive the intervention (at-risk students, elementary students, remedial classes, etc.).
- The specific expected outcomes of the program.
- Plans for evaluating the success and effectiveness of the intervention.
- An explanation of how the project could be adopted by other Georgia school systems at a reasonable cost.

Upon receipt of a favorable rating high enough to be selected, a system would create a full blown proposal and a budget would be negotiated. Consultants would be used to help refine the innovation and evaluation design. Evaluation activities come into play at two stages of the process. The first year on-site audit is a formative evaluation by an external team of content and technical experts. The second year on-site is summative and at that time a decision is made as to whether the program or project is worthy of dissemination throughout the state. Limited funds are available to systems who want to adopt a particular program. Sites where projects were developed can also become training sites for adopters or this function may be handled by a regional center.

Table 4-8 contains a summary form completed by the on-site validation team. This would accompany a brief narrative as well as a recommendation for continued funding at the end of year 1 or a recommendation for state validation at the end of year 2.

The relevance of this process for the consumer metaphor is obvious. The evaluation process is being completed for the state consumers of the programs and projects. It has to be cost-effective!

It's nice to have the shopping done for you, but there are some disadvantages. Table 4-9 contains some advantages and disadvantages of the consumer metaphor. One of the clear and present dangers of this metaphor is that it may stifle local initiative and leave a system open to undue influence by commercial vendors. On the other hand, having the evaluation done by experts who have access to greater resources is a definite advantage.

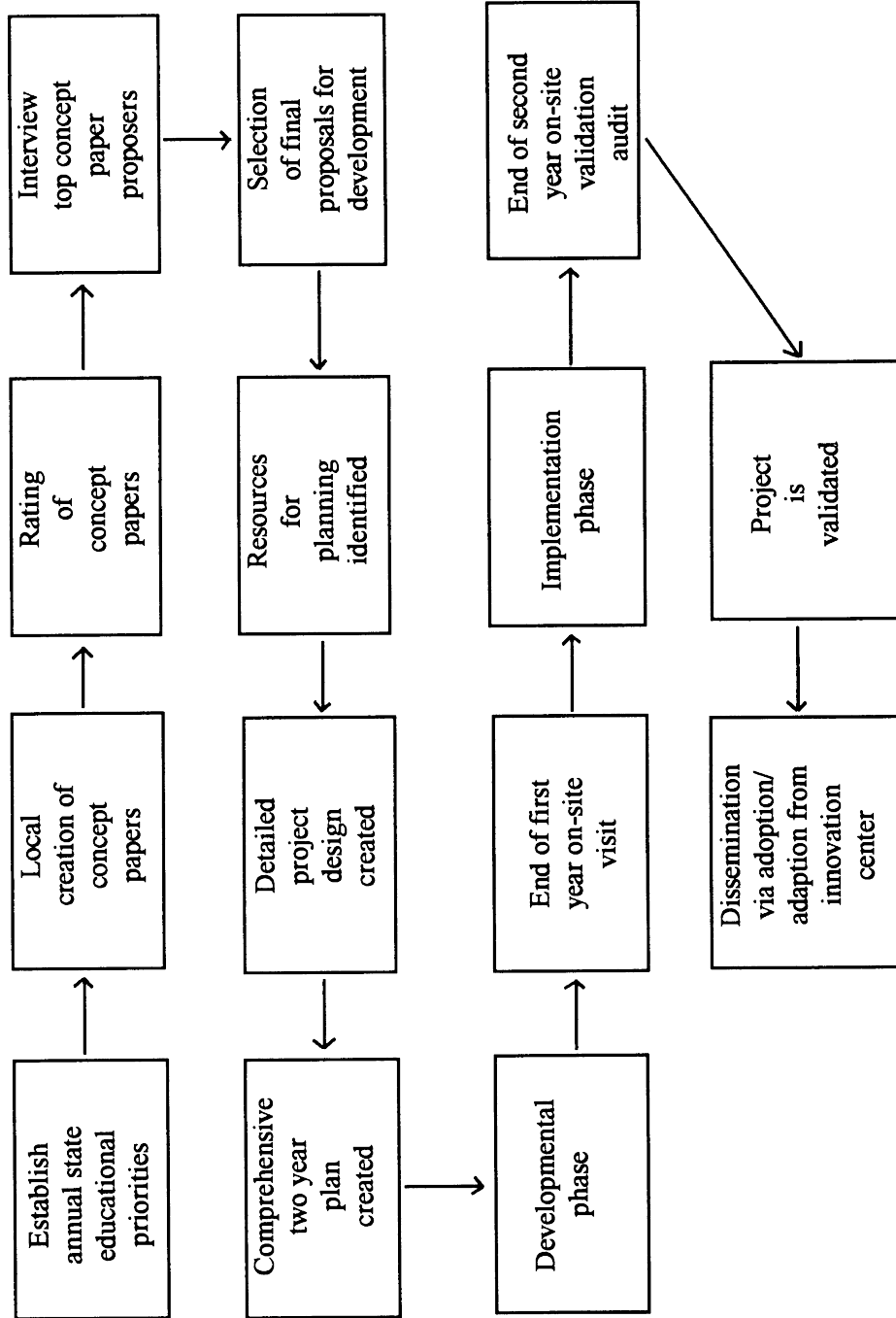


Figure 4-3. Life History for Innovative Project Development and Validation

TABLE 4-8 Summary On-Site Evaluation Form for State Projects

| I. Information & Overview | Section Rating (Circle Appropriate Rating) | |
|---|--|----------------|
| A. Project Information Complete | Acceptable | Not Acceptable |
| B. Project Abstract Clear | Acceptable | Not Acceptable |
| II. Effectiveness/Success | | |
| A. Purpose & Objectives | Acceptable | Not Acceptable |
| B. Program Activities | Acceptable | Not Acceptable |
| C. Evaluation Design | Acceptable | Not Acceptable |
| D. Results & Analysis | Acceptable | Not Acceptable |
| III. Exportability | | |
| A. Educational Significance | Acceptable | Not Acceptable |
| B. Target Populations | Acceptable | Not Acceptable |
| C. Staffing & Training Requirements | Acceptable | Not Acceptable |
| D. Materials, Equipment, & Facilities | Acceptable | Not Acceptable |
| E. Minimum Adoption Requirements | Acceptable | Not Acceptable |
| F. Replication Costs | Acceptable | Not Acceptable |
| G. Special Problems in Replication | Acceptable | Not Acceptable |
| Final Recommendation | | |
| Program or practice is recommended for validation. | YES | NO |
| Federal Program Effectiveness Panel submission is encouraged. | YES | NO |

Comments and special recommendations of the Validation Team should include mention of materials or procedures of special merit.

TABLE 4-9 Advantages and Disadvantages of the Consumer Metaphor

| ADVANTAGES | DISADVANTAGES |
|--|---|
| <ol style="list-style-type: none"> 1. Provide <u>independent</u> assessment relative to developer. 2. Is cost-effective relative to consumer. 3. Helps establish <u>standards</u> for product quality. 4. Sensitizes consumer to producer hype and dangers of anecdotal advertising. 5. Decreases likelihood that "untested" program will be foisted onto the consumer. | <ol style="list-style-type: none"> 1. Evaluation accomplished separate from consumer/practitioner. 2. Requires high degree of expertise. 3. May require considerable resources. 4. Possible bias of evaluator. 5. May inhibit evaluation and creative product development at the local level. 6. Can be expensive, with cost passed on to consumer. |

METAPHOR SELECTION: IN PRAISE OF ECLECTICISM

All this discussion of models is wonderfully enlightening, but what can the evaluator do when faced with a decision about which metaphor to use? That selection will depend on a lot of different variables including, but not limited to: (1) financial resources, (2) nature of evaluation object, (3) personalities of evaluator and major stakeholders, and (4) nature of "political" environment surrounding the decision to be made. Quite frankly, the evaluator must employ an approach that is within his or her "comfort zone," a kind of psychological state that allows him or her to feel confident in completing the tasks and working with the stakeholders. If the decisions to be made are in support of management, and are aimed at assessing the objectives and tasks associated with planning, structuring, implementing, or recycling programs or projects, then the CIPP or a CIPP-like metaphor makes sense. On the other hand, if the evaluation appears to call for a great variety of data that reflects on the expenditure of considerable amount of money on a sensitive issue about which the public has great concern, then the open nature of the judicial metaphor may be most appealing. Not to be overlooked is the very important role that could be played by the summative evaluator as consumer surrogate. And finally, the stakeholder may want an inside-out look at a project and an assessment of what really happened as opposed to what may have been intended. If that is the case, then a more naturalistic/participant observer metaphor might be desired. Using this anthropological approach may also more likely lead to the uncovering of unintended side effects than any of the

other methods. One can use the general metaphors for ideas about the *general* approach to a particular evaluation problem. They represent "ways of thinking" about evaluation. Once a general approach has been identified, concepts from any of the metaphors might be woven into an evaluation fabric. Data collection methods, for example, will require a variety of techniques running from quite subjective to quite objective. Any techniques could be used as part of any of the metaphorical framework. The four metaphors presented in this chapter reflect differences in emphasis on a variety of philosophical elements such that the "approach" to evaluation will be different or at least reflect different emphases or "flavors." They have been tried and found to work. It is obvious that the metaphors follow rational principles, if not the so-called scientific method. But science should be molded to meet our needs. The value of the metaphor is to help us think through the entirety of the evaluation task. One should select, then, a metaphor that comes closest to the intent of the evaluation, considering resources available, nature of decision to be made, data collection methods likely to be used, nature of stakeholders, and comfort zone of the evaluator.

Back in 1979 Willis documented the existence of 58 evaluation models. After much inbreeding and mutating, one wonders how many may exist today. Beware of friends bearing metaphors!

COGITATIONS

1. What are the advantages to having multiple evaluation metaphors?
2. Metaphors other than those presented in this chapter are possible: for example, investigative journalism, photography, and architecture. Can you think of others?
3. How would proponents of each of the four metaphors in the present chapter approach the task of selecting social science textbooks for a school system or a new language arts program? What approach makes sense in working with art education at the middle school level or in establishing interventions for at-risk pre-school students?
4. How might you methodologically address the disadvantages of each of the four evaluation metaphors presented in the present chapter? In other words, what specific techniques would more likely be used by one metaphor or another?

SUGGESTED READINGS

- Eisner, E.W. (1976). Educational connoisseurship and criticism: Their form and functions in educational evaluation. *Journal of Aesthetic Education*, 10 (3-4), 135-150.
- House, E.R. (1978). Assumptions underlying evaluation models. *Educational Researcher*, 7 (March), 4-12.
- Ralph, J., & Dwyer, M.C. (1988). *Making the case: Evidence of program effectiveness in schools and classrooms*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.
- Scriven, M.S. (1974). Evaluation perspectives and procedures. In W.J. Popham (Ed.), *Evaluation in education* (pp. 1-94). Berkeley, CA: McCutchan.
- Smith, N.L. (1985). Adversary and committee hearings as evaluation methods. *Evaluation Review*, 9(6), 735-750.
- Stufflebeam, D.L. (1983). The CIPP model for program evaluation. In G.F. Madaus, M.S. Scriven, & D.L. Stufflebeam (Eds.), *Evaluation models* (pp. 117-141). Boston: Kluwer-Nijhoff.