DESIGNING EDUCATIONAL PROJECT AND PROGRAM EVALUATIONS

Evaluation in Education and Human Services

Other books in the series:

Madaus, G. and Stufflebeam, D.:
Education Evaluation: Classic Works of Raphy W. Tyler
Gifford, B:
Test Policy and Test Performance
Osterlind, S.:
Constructing Test Items
Smith, M.:
Evaluability Assessment
Ayers, J. and Berney, M.:
A Practical Guide to Teacher Education Evaluation
Hambleton, R. and Zaal, J.:
Advances in Educational and Psychological Testing
Gifford, B. and O'Connor, M.:
Changing Assessments
Gifford, B.:
Policy Perspectives on Educational Testing
Basarab, D. and Root, D.:
The Training Evaluation Process
Haney, W.M., Madaus, G.F. and Lyons, R.:
The Fractured Marketplace for Standardized Testing
Wing, L.C. and Gifford, B.:
Policy Issues in Employment Testing
Gable, R.E.:
Instrument Development in the Affective Domain (2nd Edition)
Kremer-Hayon, L.:
Teacher Self-Evaluation

DESIGNING EDUCATIONAL PROJECT AND PROGRAM EVALUATIONS

A Practical Overview
Based on Research and Experience

David A. Payne
University of Georgia

With a chapter on Qualitative Methods
contributed by Mary Jo McGee-Brown,
University of Georgia

Springer Science+Business Media, LLC

Library of Congress Cataloging-in-Publication Data

Payne, David A.
Designing educational project and program evaluation : a practical overview based on research and experience / David A. Payne.
p. cm. -- (Evaluation in education and human services)
Includes bibliographical references and index.
ISBN 978-94-010-4602-2 ISBN 978-94-011-1376-2 (eBook)
DOI 10.1007/978-94-011-1376-2
1. Educational evaluation. 2. Evaluation research (Social action programs) I. Title. II. Series.
LB2822.75.P39 1994
379.1'54--dc20 93-38469
CIP

Printed on acid-free paper.

For Beverly . . .

. . . with love by design.

CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

LIST OF EXHIBITS

EXHIBIT

PREFACE

The consensus contemporary definition of evaluation rests on the concepts of assessment of merit and worth determination. In these days of reform, educators are continually faced with the challenges of evaluating their innovations. Both common sense and accepted professional practice would suggest a systematic approach to these evaluation challenges. The author has attempted to present just such a systematic approach. The philosophy of the volume is to engage the assessment of merit concept in a systematic decision-making framework. Harried and often harassed educational evaluators and administrators are increasingly required to respond to requests for informed evaluative decisions.

It is hoped that the following extracts from traditional and current literature, and illustrations of evaluation methods from the field will prove useful. In particular it is hoped that practicing educational program and project evaluators, public school administrators, and state department of education personnel will find applicable ideas and suggestions in the following book. In addition it is hoped that the book will find an audience with those teachers or administrators who wish to engage in what historically was called "action research," those local, relatively small scale studies of processes, procedures, or products.

The book begins with an overview of the generic evaluation process. In particular, research and evaluation activities are contrasted. Chapter 2 is devoted to the criteria for judging the effectiveness of evaluation practice. These criteria can be used to help design an evaluation or evaluate an already completed study. Chapter 3 addresses the all-important topic of evaluation goals and objectives. Interfaced with evaluation questions is the setting of standards or establishment of criteria against which the questions are evaluated. Chapters 4, 5, and 6 basically are concerned with the approach, framework, or design of an evaluation study. Chapter 4 contains a discussion of four major philosophical frameworks or metaphors and the implications of these frameworks for conducting an evaluation. It is hoped that the reader will consider Chapters 5 and 6 together. These chapters describe predominately quantitative and qualitative designs, respectively. Viewing the evaluation process as requiring multiple methods and approaches emphasizes the need for a responsive evaluation to be eclectic. Design, implementation, and operational issues related to instrumentation (Chapter 7), management and decision-making (Chapter 8), and reporting and utilization of results (Chapter 9) are next addressed. The final chapter of the book (Chapter 10) considers the evaluation of educational products and materials.

Each chapter concludes with a series of Cogitations. These are offered to help the reader think about and through the ideas of the chapter. They might also be used as "advance organizers" as well as a base for review purposes.

The Suggested Readings of each chapter were selected to represent introductory to intermediate level content, and classical as well as contemporary literature.

The transcription of the manuscript for this book was an onerous and fatiguing task. The author apparently suffered a traumatic experience in elementary school and as a result his cursive script was too primitive to be even considered hieroglyphics by contemporary standards. Further trauma in graduate school generated a fear of computer composition. It fell to the Queen of the Keyboard, Janet Goetz, to bring order out of chaos. Her manuscript manipulations represent a major contribution to this venture. Very able assistance was also provided by Rachel Anderson and Michelle Bennett.

Several colleagues also directly contributed to whatever is valuable in this volume. Dr. Mary Jo McGee-Brown's chapter on qualitative methods is both illuminating and useful for either beginning or experienced evaluators. The author gratefully acknowledges her contribution. Dr. Carl J Huberty provided his usual scholarly insight and moral support during this venture, particularly regarding the design of evaluation studies. Dr. Gerald Klein of the Georgia State Department of Education provided many opportunities for real live evaluation experiences and material related to the Consumer Metaphor described in Chapter 4. Margaret Lipscomb of the Eisenhower Medical Center in Augusta, Georgia, graciously permitted the abridged use of her unpublished work related to evaluating text materials presented in Chapter 10. The senior editor at Kluwer, Zachary Rolnik, provided many excellent ideas and suggestions. To these and many others, a heartfelt "thank you!"

1

EDUCATIONAL EVALUATION: DEFINITIONS, PURPOSES, AND PROCESSES

Scene I
(Evaluator's Office)

It was a sultry August afternoon, although an occasional puff of air would bend the Bermuda grass and cause the leaves to wave and pine needles to fall. Summer school was over, and the campus was deserted. It was almost an academic ghost town. Once in a while you could catch a glimpse of an assistant professor scurrying from library to computer center in search of the publication quota for the year. I was about to leave for the day and perhaps catch nine holes before dinnertime when the daydream of a modest 200-yard drive off the first tee was shattered by the ringing of a phone. The call was from the science coordinator of the local school system.

Science Coordinator: Dr. Stufflestake?

Evaluator: Yes, ma'am.

SC: My name is Dee Pressed. Your name was given to me by one of your colleagues as someone who occasionally is involved in evaluating curriculum projects.

E: Yes, ma'am.

SC: Is that correct?

E: Yes, ma'am.

SC: Well, our system has received a developer/demonstrator grant from the state education department to finish the revision of our K--12 science curriculum, Project Hawthorne, which was initiated last year. We have also been charged with the responsibility for evaluating the program.

E: Yes, ma'am.

SC: As project director, I am inviting you to join us in the evaluation phase of the project. Do you have time to devote to such a project during the coming year?

E: Yes, ma'am.

PD: Does it sound like something you would be interested in working on?

E: Yes, ma'am.

PD: Can you say anything besides "Yes, ma'am?"

E: I surely can. Let's see.... When does the evaluation design have to be completed?

PD: Tomorrow!

E: (*Showing great insight*) That doesn't give us much time, does it?

PD: No, sir. Could you meet with a writing team at 6:30 this evening in the board of education conference room?

E: Yes.

PD: (*Quickly interjecting*) Don't say it.

Scene II
(Later That Night)

The conference room is filled with the stale smell of cigarettes, old coffee, French-fries, Big Macs, and sweaty bodies. Scene opens with Dr. Stufflestake entering, somewhat tardy.

Project Director: ...and so I called my friend Seymour Clearly, the curriculum director over in Red Clay County. He said we could use their high school science classes if we don't disrupt the program too much. But there is one tiny problem... .

State Department Representative: Oh?

PD: It seems that they treat science topics in a different sequence than we do.

SDR: That's not my problem!

Evaluator: That's really going to cause difficulties in trying to match the time when specified objectives are to be evaluated in the experimental and control schools!

SDR: That's not my problem!

E: (*In exasperated tone*) That means that we will have to develop 2,300 test items before school opens!

SDR: That's not my problem!

Science Education Professor: Let us not be inhibited by the prospect of creating a multitude of inquiry mechanisms prior to the initiation of the academic calendar. Such an event need not result in the promulgation of psychophysical reactions that can be traced to an interaction of the endocrine system and selected neocognitive apperceptions.

SDR: Whad he say?

PD: Don't panic.

Scene III
(Sweat Shop)

A crisp November Saturday afternoon. Sounds of cheering football fans drift across campus. Item writers are huddled over tables in a College of Education closet, upon which are spread thousands of behavioral objectives. Stacks of reference books, science texts, human anatomy charts, climate graphs, and stem, leaf, and tree diagrams are seen around the room.

1st Item Writer: (*Timidly*) Excuse me, but we are finding it difficult to meet all the item specifications, particularly for the elementary students. It seems that analogy items for first graders would be unfair even if they will only be given to the control group. And is it really necessary for the multiple-choice items to always have five alternatives, with a direct question stem, and be free of negative statements?

Overseer: (*Disgustedly*) You ask too many questions. Shut up and write.

2nd Item Writer: (*With even greater timidity*) I was visiting one of our schools a couple of days ago and some of the teachers commented on the quality of the items.

O: Oh, yeah!

2nd IW: (*In sotto voce*) Yes! For example, one teacher said that the green square in the color discrimination test is really a gray rectangle. What do you think that means?

O: That the teacher was color blind.

2nd IW: And she said that the question that asks the student to point to all the "man-made" lights led a girl to point to all the light sources including the sun and stars saying, "God made them and he's a man."

O: What happened then?

2nd IW: A long discussion about the ERA.

Scene IV
(Evaluator's Office)

A mid-December Monday morning. Evaluator opens weekend accumulation of mail which contains following memo.

DATE: Sunday, December 7

TO: All Science Project Personnel

FROM: Project Director and Central Office

SUBJECT: Problems

We have just learned that our printer in central duplicating has decided to take his annual three-week vacation beginning Friday. We will consequently have to reschedule all testing for the rest of the year. Also, would teachers please refer to their master coding sheets when assigning codes for each unit test? Our last readout indicated that the third grade experimental group contained 324 teachers and 12 students.

We have scheduled a meeting of all irate parents of control students for 11:45 P.M. on Sunday.

P.S. The tests which were stolen from Mr. Jones' locker last Friday were returned Monday by the thief who said the reading level was too high.

P.P.S. School Facilitators-Please request teachers not to return their answer sheets in 3x5" envelopes as it tends to drive the optical scanner operator bananas.

Scene V
(Evaluator's Office)

It's a dark, dreary, cold, and rainy Friday morning in March following a hectic week of project trials and tribulations in addition to harassment from administrators. The evaluator is deeply engrossed in the latest issue of the Journal of Obscure Statistics that carries an article by one of his colleagues, Dr. Sig Nificant, concerning the discovery of a negative variance. The phone rings.

Computer Center Rep.: Dr. Stufflestake?

Evaluator: Yes.

CCR: This is Jim Nastike from the Computer Center. I'm afraid we have some bad news.

E: What do you mean *we* have some bad news.

Jim Nastike: Well, ah...ah...as you may or may not be aware, the turnover in operations personnel during the late evening hours here at the center has become a problem.

E: Mmmm.

JN: Well, ah...ah...it seems that your last file tape (*gulp*) containing the aggregated data from the fall and winter testings were accidentally erased by one of our new technicians.

E: You erased *what*?!

JN: Your fall and winter data. (*Sound of phone crashing to floor and body falling off chair*)

(Curtain)

Doing "real--life" evaluations of the type described in the foregoing "morality play" (with apologies to Aeschylus) is likely to result in much stress, calamities, fizzles, and occasional disasters. If the aspiring evaluator is not prepared for such frustrations, then loss of physical and mental health may result. But despite the problems encountered in doing evaluations in naturally occurring educational settings, a great deal of satisfaction and many positive outcomes can be realized. One of the joys of doing an evaluation, and it can be a very exciting and satisfying experience, is represented by the problem--solving challenges. "Here is an opportunity for me to do something worthwhile--I can make a difference." Be assured that doing efficient and effective evaluations will take a great deal of hard work, require the expenditure of considerable time and effort, and necessitate the use of a very broad range of competencies. The "new generation" evaluator, if he or she is to function optimally, must be even more of a generalist than was his or her predecessors. In addition to being master of some fairly sophisticated quantitative skills (research, measurement, and statistics), the neo-evaluator must be a little bit of a sociologist, economist, social psychologist, anthropologist, and philosopher. The evaluator must be blessed with a strong self-concept, high tolerance for ambiguity, and, quite frequently, the patience of a United Nations arbitrator. But what is evaluation all about?

THE NATURE OF PROJECT AND PROGRAM EVALUATION

People are always evaluating. We do it every day. We buy clothing, a car, or refrigerator. We select a movie or subscribe to a magazine. All these decisions require data-based judgments. Data take many forms. Sometimes we rely on our own experience or the opinions of others. Sometimes we require more formal information like that derived from experiments and controlled studies. Educators make decisions about the effectiveness of curricula and/or programs, the progress of individual students toward specified goals, and efficiency of instructional methods. The most generally accepted definition of educational evaluation involves the idea of the assessment of merit, or the making of judgments of value or worth (Scriven, 1991)[1]. The process employs both quantitative and qualitative approaches. One theme of this book is that the making of informed value judgments requires the availability of reliable and valid data, and the exercise of rational decision making. This is as true about programs, projects, and curricula as it is about individuals. Good evaluations require sound data! So let's hear it for sound data *and* evaluations.

As used throughout this book the terms project and program are *not* interchangeable. A *project* is viewed as an isolated, probably one-time effort to "try to make a difference" by using an innovation. That "innovation" might relate to a method of teaching science lab skills with a portable equipment cart or an approach to improving student attitudes toward learning for the elementary students in a particular rural school. In essence the evaluator, or more correctly a client or stakeholder, wishes to find out if the innovation is of value, e.g., made a positive impact on student writing skills. If it is, then it may be incorporated as a regular part of a program. A *program* is seen as less transitory than a project, probably more complex and broader in scope. A specific project may be one of several innovations synthesized into addressing the needs of at-risk elementary students, whereas the program is viewed as being a multifaceted and focused approach to solving general educational problems. As used here the term *program* addresses a problem and the term *project* addresses a specific purpose. We might further stretch our complexity analogy to describe a curriculum as a collection of programs, e.g., social studies, reading, science, mathematics, which has scope and sequence. There are some who say that even if a particular innovation does not yield replicable results, it was still worth the effort to try something new. It gets everybody fired up, the creative juices flowing, and enthusiasm coursing through our veins.

---

[1]References are collected at the end of book.

Educational evaluation is probably of greater concern today than at any time in history due to the massive amounts of knowledge that our citizens must transmit and process, as well as to the complexity of this knowledge. Evaluative techniques adequate for assessing the effectiveness of small units of material or simple processes are significantly less satisfactory when applied to larger blocks of information, the learning of which is highly complex and involves prerequisite learning, sequential behaviors, and perhaps other programs of study. Educational institutions from state to local level are emphasizing problem solving. The traditional use of experimental and control groups (as examined by contrasting gross mean achievement scores in a pre--post treatment design study), although generally valuable, tends not to provide sufficiently detailed information upon which to base intelligent decisions about program effectiveness, validity, efficiency, and so on. Consumers and evaluators alike have lamented the failure of many evaluation designs, particularly those in government research proposals, to meet even minimal requirements. The desire or need to compromise evaluation designs results in far too many "no significant differences." The practitioner seeking information about the success of his or her innovative program is "inviting interference." This is a situation incompatible with control. If we lack control of the treatment or data collection, the experimental designs and methods of data analyses are considerably less applicable. Most applied studies are done in natural settings, and natural educational settings are anything but controlled. But it is in these relatively unstructured and uncontrolled situations that evaluation and decisions must be undertaken. The field of evaluation is developing in response to the requirements for decision making in these kinds of environments.

Occasionally there is sufficient control of the sampling unit to allow for the application of an experimental or quasi-experimental design. In most cases, however, evaluators must use their creativity to find contrast or benchmark data to use in assessing program or project impact.

THE ROLES OF EVALUATION

Evaluation will play many roles, contingent on the demands and constraints placed on it (Heath, 1969). Three broad functions of evaluation are:

1. Improvement of the program during the development phase. The importance of formative evaluation is emphasized. Strengths and weaknesses of the program or unit can be identified and enhanced or strengthened. The process is iterative, involving continuous repetition of the tryout--evaluation--redesign cycle.

2. Facilitation of rational comparison of competing programs. Although differing objectives pose a large problem, the description and comparison of alternative programs can contribute to rational decision making.

3. Contribution to the general body of knowledge about effective program design. Freed from the constraints of formal hypothesis testing, evaluators are at liberty to search out principles relating to the interaction of learner, learning, and environment.

These potential contributions of evaluation to the improvement of quality and quantity in education, have been described by Scriven (1967) as "summative" and "formative" evaluation. He notes that the goal of evaluation is always the same, that is, to determine the worth and value of something. That "something" may be a microscope, a unit in biology, a science curriculum, or an entire educational system. Depending upon the role the value judgments are to play, evaluation data may be used developmentally or in a summary way. In the case of an overall decision, the role of evaluation is summative. An end-of-course assessment would be considered summative. Summative evaluation may employ absolute or comparative standards and judgments.

Formative evaluation, on the other hand, is almost exclusively aimed at improving an educational experience or product during it's developmental phases. A key element in the formative technique is feedback. Information is gathered during the developmental phase with an eye toward improving the total product. The evaluation activities associated with the development of Science--A Process Approach, the elementary science curriculum supported by the National Science Foundation and managed by the American Association for the Advancement of Science, are illustrative. During the several years of the program's development, sample materials were used in centers throughout the country. Summer writing sessions were then held at which tryout data were fed back to the developers. A superior product resulted. Teacher materials were improved and student learning activities were changed to adapt better to their developmental level. The summative-formative distinction among kinds of evaluation reflects differences, for the most part, in intent rather than different methodologies or techniques.

The suggestion has been made that summative and formative evaluations differ *only* with respect to the time when they are undertaken in the service of program or project development. There are, however, other dimensions along which these two roles could be contrasted. A very informative and succinct summary has been created by Worthen and Sanders (1987) and is reproduced in Table 1-1.

Most projects will use both approaches. Obviously an end-of-year summative evaluation can be formative for the next year. As projects and programs mature, the amount of time devoted to the type of evaluation will shift, with movement from more formative to more summative.

The use of evaluation in the investigation of merit might imply that evaluation should be viewed as a research effort. As a matter of fact, Suchman (1967) has formalized this idea and describes the process as "evaluative research." But there are dangers in treating the two processes as equivalent.

TABLE 1-1 Differences Between Summative and Formative Evaluation

Basis for Comparison	Formative Evaluation	Summative Evaluation
Purpose	To improve program	To certify program utility
Audience	Program administrators and staff	Potential consumer or funding agency
Who Should Do It	Internal evaluator	External evaluator
Major Characteristic	Timely	Convincing
Measures	Often informal	Valid/reliable
Frequency of Data Collection	Frequent	Limited
Sample Size	Often small	Usually large
Questions Asked	What is working? What needs to be improved? How can it be improved?	What results occur? With whom? Under what condition? With what training? At what cost?
Design Constraints	What information is needed? When?	What claims do you wish to make?

From: Educational Evaluation: Alternative Approaches and Practical Guidelines by Blaine R. Worthen and James R. Sanders. Copyright© 1987 by Longman Publishing Group. Reprinted by permission.

EVALUATION IS NOT RESEARCH

Many experts view evaluation as the simple application of the scientific method to assessment tasks. In this sense, which parallels Suchman's use of the term, evaluative becomes an adjective modifying the noun research. The emphasis is still on research, and on the procedures for collecting and analyzing data that increase the possibility of supporting claims, rather than simply asserting, the worth of some social activity. It is perhaps best not to equate the two activities of research and evaluation because of differences in intent and applicability of certain methodologies. The following parallel lists are a brief but general comparison of these two activities.

ACTIVITY	RESEARCH	EVALUATION
1. Problem selection and definition	Responsibility of investigator	Determined by situation and constituents
2. Hypothesis testing	Formal statistical testing. Usual in highly quantitative studies	Sometimes
3. Value judgments	Limited to selection of problem	Present in all phases of project
4. Replication of results	High likelihood	Low likelihood
5. Data collection	Dictated by problem	Heavily influenced by feasibility
6. Control of relevant variables	High	Low
7. Generalizability of results	Can be high	Usually low

Some important differences between research and evaluation are evident in these contrasting emphases. Many further differences are implied. It is argued by some scientists that the primary concern of research should be the production of new knowledge through the application of the "scientific method." Such information or "conclusions"would be added to a general body of knowledge about a particular phenomenon or theory. A high proportion of the research studies in the physical, biological, and behavioral sciences are aimed at contributing to a particular theory or, at the very least, are derived from theory. Evaluation activities are generally not tied to theory except, perhaps, to the extent that any curriculum project is founded on a particular

theoretical position. Evaluation studies are generally undertaken to solve some specific practical problems and yield decisions, usually at a local level. There is little interest in undertaking a project that will have implications for large, widely dispersed constituencies. Control of influential variables is generally quite restricted in evaluation studies. It is for this reason that routine application of experimental designs--as described, for example, by Campbell and Stanley (1963)--may be inappropriate. Research in the behavioral sciences is, in a restricted sense, concerned with the systematic gathering of data aimed at testing specific hypotheses and contributing to a homogeneous body of knowledge.

One of the contributions that an evaluation can make in and above an assessment of merit or value is pure description. The documentation of what has gone on in implementing a project or program is important so that (1) we can better understand and monitor the fidelity of implementation, and (2) there exists a basis for generating replications if the project or program proves valuable.

A question remains about the ways educational evaluation differs from pure research, or the straightforward evaluation of learning. Following is a list of variables that may clarify the emphases relatively unique to evaluation:

Nature of goals. The objectives of evaluation tend to be oriented more to process and behavior than to subject matter content.

Breadth of objectives. The objectives of evaluation involve a greater range of phenomena.

Complexity of outcomes. Changes in the nature or life and education, and the increased knowledge we now possess about the teaching-learning process, combine to require objectives that are quite complex from the standpoint of cognitive and performance criteria. The interface of cognitive, affective, and psychomotor variables further complicates the process of identifying what must be evaluated.

Focus of total evaluation effort. There is a definite trend toward increasing the focus on the total program, but this is in addition to the continued emphasis on individual learners.

Context of education. Evaluation should take place in a naturalistic setting, if possible. It is in the real-life setting, with all its unpredictable contingencies and uncontrolled variables, that education takes place. We must evaluate and make decisions in the setting in which we teach.

It perhaps makes most sense to conceive of evaluation, as Cronbach and Suppes (1969) have, as "disciplined inquiry." Such a conception calls for rigor and systematic examination but also allows for a range of methodologies from

traditional, almost laboratory--like experimentation to free-ranging, heuristic, and speculative goal-free evaluation. As much as we want to be as scientific as possible, we must realize (1) the very real practical boundaries in most evaluation settings, and (2) the very real political influences that can and will be brought to bear on the evaluation. Sometimes it seems that *everybody* has a vested interest in the results. Interest is one thing, undue influence is another. If all of these evaluation roles and functions are to be addressed, some general framework is needed to help guide the process.

ACTIVITIES IN THE EVALUATION PROCESS

There will probably never be total agreement on the nature of the activities and sequence of steps in the evaluation process. The kind of evaluation questions being asked, availability of resources, and time lines are some of the factors that would dictate the final form of the process. Basically, the process boils down, with some exceptions, to an application of the principles of the "scientific method," but not always being used in a linear fashion. Some evaluations might simply require the retrieval of information from records in files, while others might require pilot or field studies. Such studies might be as simple and informal as sitting with a student and listening to him or her work through a new unit on long division, or it might be something as complex as a 20 percent sample achievement student survey study of all major physics objectives at the ninth grade level.

Figure 1-1 contains a brief outline of the usual activities in conducting an evaluation. Only the major activities are identified. The dotted lines indicate that information may be shared between blocks (activities/processes), and that decisions are continuously being modified and revised. The activities presuppose that a needs assessment has been conducted and that an innovative project or program has been proposed or put in place (Kaufman & English, 1979). The sequence of activities in Figure 1-1 may be followed directly and exactly if summative evaluation is the role being played by evaluation, or periodically and systematically repeated if formative evaluation is the primary intent. The sequence of activities may change depending on not only the requirements of the project or program evaluation but also the approach or methodology used by the evaluator. A traditional objectives-based evaluator (Tyler, 1942; Stufflebeam, 1983; Provus, 1971) would be likely to follow the steps in Figure 1-1 more or less in sequence. A goal-free or responsive evaluator (Scriven, 1972, 1978; Stake, 1983) might skip the first, second, and/or third steps and just begin gathering data via observations, questionnaires, interviews, and so on. At some point, however, most if not all of the activities would need to be addressed.

All of the activities are important, but one of particular interest in developing a comprehensive evaluation program is Standard Setting. The specification of criteria may be the most important part of the evaluation process. The question asked is: "On what basis do I make a value judgment?" The criteria might relate to an individual (Did Rick learn 75% of today's vocabulary words about plants?) or group or institution (Did 80% of the

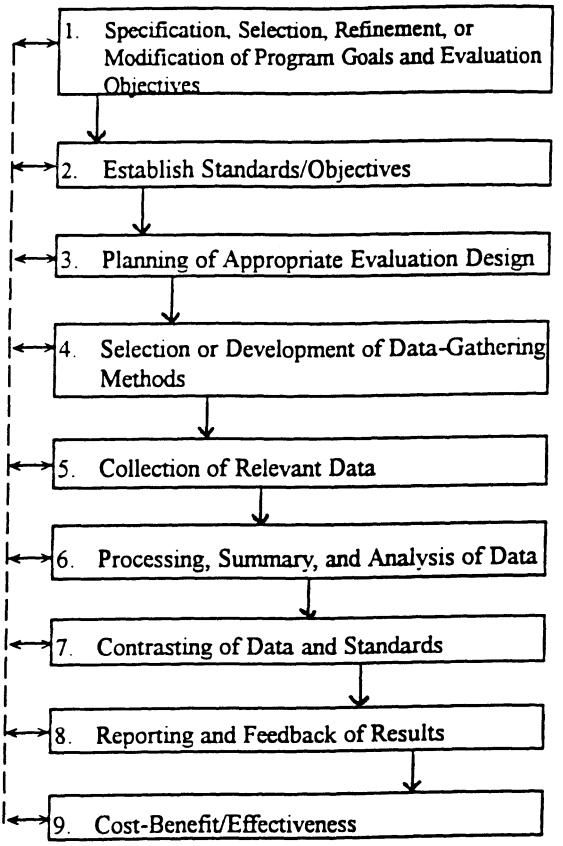


Figure 1-1 Overview of Usual Activities in the Evaluation Process

students in grade 5 in the county learn 80% of the capitals of 50 major countries?). We then gather data to evaluate the objectives.

Standards may be set prior to data collection if the instrumentation is known or selected. It might take place after data collection but before decision-making. One of the dimensions that might be used to differentiate research and evaluation is the nature of the decision-making method.

Traditional research studies tend to rely on mathematical models (e.g., statistical tests) to make or help make data-based decisions. Evaluations may use statistical procedures but may also employ subjective and judgmental approaches.

Another step, but a frequently overlooked one in the evaluation process, is cost analysis (see Chapter 8). There are costs associated with effective educational programs and projects, both monetary costs and costs in terms of human resources. The cost-benefit question (Did it benefit the individual or society?) and cost-effectiveness question (Was the investment worth the dollars expended?) can be answered after the overall evaluation has taken place. An evaluator may be faced with the problem of finding that method A of teaching the dangers of drug abuse is as effective as the current approach, method B, but takes half the classroom time. Unfortunately, the data revealed that method A costs a third again as much as method B. Cost versus effectiveness questions are difficult to resolve.

The major thrust of this book will be to take these activities and hopefully describe and discuss them to such an extent that an evaluator can Do It!

We have been doing evaluations of all types, kinds, and sizes for many decades. What have we learned from these experiences?

EVALUATION TRUISMS

The methods, modes, models, and motivations for conducting evaluations have changed over the last 50 or so years since the rebirthing of evaluation by Ralph Tyler (1942). As the field of evaluation has evolved certain truths have become self-evident. These GEMS (Golden Evaluation Merit Statements, created for the Society for the Preservation and Encouragement of Sound Evaluation Practice) are not nearly as erudite as the theses derived by Dr. Lee J. Cronbach and his associates at Stanford University (Cronbach et.al., 1980), but hopefully they represent credible guidelines that may help focus thinking about the evaluation enterprise.

Evaluation Is A Way of Thinking One's philosophy, world view, and theory of how we "know" will greatly influence the approach taken to evaluation. Technical matters are not the only consideration. An evaluator must also reflect on the goals of society and what should be the objectives of the human community. Of necessity, therefore, political issues will intrude into evaluation. These issues can frustrate but also help clarify what we are trying to find out (House, 1983a,b).

Evaluations Should Be Naturalistic It is the view of many that since most educational projects and programs are problem-focused that their evaluations should take place in the context of where the problem is identified. One of the lessons of educational research is that a newly devised treatment or method will not work (generalize) to all settings. This is particularly true for educational innovations and interventions as these new approaches were developed to meet specific needs. Specific needs require specific evaluations. There is an old-fashioned term that is applicable here--termed action research. We have asked teachers, for example, to try out informally new ideas and see how they work. The "see how they work" part is a mini-evaluation. The term naturalistic, in addition to being applied to the setting of the evaluation, could also be applied to the evaluation procedures used--observations, interviews, and so on--, where less artificial means for data collection could be used (Guba & Lincoln, 1981).

The Design of Evaluation Studies is Evolutionary The fact that evaluations should take place in naturally occurring situations can lead to significant design problems. Because the situation is natural--a school or classroom, for example--, changes are always expected and experienced. The evaluation design must be flexible. What do I do about the "treatment teacher" who is going to be on maternity leave for six weeks? What do I do about the 50% student sample that was out with flu on the final data collection date? These and other frightening occurrences can cause an evaluator to become unstable. Fortunately a good design will include provisions for meeting unanticipated problems. Certain objectives may become unrealistic due to treatment failure or lack of availability of data may require a change in criterion measures. Expect the unexpected!

The Complexity of Contemporary Evaluation Requires Multiple Models and Methods The just noted change in criterion problems highlights the need for alternative data types and data sources. The marriage of quantitative and qualitative methods is in process. Courtship is in progress, but the merger has not been consummated. The birth of mixed methods allows for greater responsiveness to evaluation questions which in turn allows for greater responsiveness to stakeholder needs. In hope of not totally destroying the metaphor, the cross-fertilization of not just methods but also philosophy allows for the fuller and richer assessment of an evaluation question. Triangulation (multiple methods, common target) is now the keystone in the evaluation arch that must support the weight of an innovative program or project (Greene, Caracelli, & Graham, 1989).

Effective Evaluation Requires Continuous Involvement and Commitment from Concept to Implementation At the center of a fruitful evaluation (our nuptial metaphor continues) is utilization, since evaluation without utilization of results is a tragic waste of time, effort, and resources. One thing an evaluator can do to help insure utilization of results is to maximize the involvement of individuals who have the greatest investment in the outcome(s) of the project or program--the stakeholders. Stakeholders,--for instance, parents, teachers, and administrative personnel--must be included in the framing of the evaluation questions. Their input, for example, during the implementation of a nongraded K--3 instructional program is important not only important from the basic communication courtesy standpoint, but it really helps make the evaluator's job easier. Their suggestions should go right to the heart of the purpose for creating the innovation and conducting an evaluation of it (Patton, 1990).

A different set of "truisms" recently has been espoused by Scriven (1993). Drawing on his own and the experiences of others in the "war for truth in evaluation," Scriven has deduced some theses concerning what we have (or perhaps should have) learned from our mistakes in doing program evaluation.

His 31 theses are presented here to provoke readers to think about the evaluation process as they read through this book and the Suggested Readings. Most entries are obvious, self-evident, and self-explanatory, but some will require thoughtful consideration and reference to the source. For his theses Scriven uses seven organizing categories corresponding to chapters in his monograph.

The Nature of Evaluation

- Program evaluation is not a determination of goal attainment.
- Program evaluation is not applied social science.
- Program evaluation is neither a dominant nor an autonomous field of evaluation.

Implications for Popular Evaluation Approaches

- Side effects are often the main point.
- Subject matter expertise may be the right hand of education program and proposal evaluation, but one cannot wrap things up with a single hand.
- Evaluation designs without provision for evaluation of the evaluation are unclear on the concept.
- An evaluation without a recommendation is like a fish without a bicycle.

Implications for Popular Models of Program Evaluation

- Pure outcome evaluation usually yields too little too late, and pure process evaluation is usually invalid or premature.
- Noncomparative evaluations are comparatively useless.
- Formative evaluation is attractive, but summative evaluation is imperative.
- Rich description is not an approach to program evaluation but a retreat from it.
- One can only attain fourth-generation evaluation by counting backward.

Intermediate Evaluation Design Issues

- Merit and quality are not the same as worth or value.
- Different evaluation designs are usually required for ranking, grading, scoring, and apportioning.
- Needs assessments provide some but not all of the values needed for evaluations.
- Money costs are hard to determine-but they are the easy part of cost analysis.
- Program evaluation should begin with the presuppositions of the program and sometimes go no further.
- Establishing statistical significance is the easy part of establishing significance.
- "Pulling it all together" is where most evaluations fall apart.

An Advanced Evaluation Design Issue: Beyond Validity

- Validity does not ensure credibility.
- Validity and credibility do not ensure utility.
- Even utilization does not ensure utility.
- Program evaluation involves research and ends with a report, but research reports are negative paradigms for evaluation reports.

An Advanced Evaluation Management Issue: Bias Control

- Preference and commitment do not entail bias.
- The usual agency counsel's criteria for avoidance of conflict of interest select for ignorance, low contributions, indecisiveness, or some combination thereof.
- Program officers are biased toward favorable findings.
- External evaluators are biased toward favorable findings.
- Peer review panels are unreliable, fashion-biased, and manipulable.

Parting Perspectives

- The most difficult problems with program evaluation are not methodological or political but psychological.
- Evaluation is as important as content in education programs.
- Routine program evaluation should pay for itself.

The practice of educational evaluation is expanding. As it continues to extend both the scientific and interpersonal parameters of application guidelines are needed. One need only visit the committee meetings of the American Evaluation Association, walk the halls of their convention or eavesdrop on seminars, workshops and paper sessions at the annual conclaves to see the extent of professional refinement. At a recent meeting of the Association, five guiding principals for evaluation practices were presented to the membership for their consideration. Although in preliminary form these yet unofficial principals revolve around the need for evaluators to:

1. Conduct systematic data based inquiries,
2. Provide competence performance to stakeholders,
3. Ensure that evaluations are conducted honestly and with integrity,
4. Respect the security, dignity, and self--worth of evaluations respondents, program participants, clients, and other stakeholders,
5. Strive to articulate and take into account the diversity of interests and values in general and public welfare.

Well, we've read the script; now it's time to make the movie.

COGITATIONS

1. What does the concept and practice of evaluation mean to you relative to your work?
2. If you were asked to evaluate this text (for a large fee and with unlimited resources) how would you approach the task formatively and summatively?
3. What characteristics of the "scientific method" are in common with research and evaluation? Which characteristics best differentiate the two activities?
4. What are three instances in which an evaluation has had a major impact on your life? In what way were they evaluative?
5. How does one "assess merit" or "determine the value" of a program, project, or product?
6. Why should evaluation designs usually be considered as "tentative"?
7. Why should evaluation be carried out in naturally occurring settings?

SUGGESTED READINGS

All of the following are introductory in nature but vary in detail and intensity from heavy (Stufflebeam, and Shadish, Cook, and Leviton) to light (Berk and Rossi).

Berk, R.A.,& Rossi, P.H. (1990). *Thinking about program evaluation.* Beverly Hills, CA: Sage. A brief but insightful and provocative paperback.

Jaeger, R.M. (1992) (Ed.). *Essential tools for Educators. (The program evaluation guide for schools.* Newbury Park, CA: Corwin. A series of excellent brief manuals for evaluating programs in special education, counseling, reading and language arts, mathematics, and for at--risk students.

Kosecoff, J., & Fink, A. (1987). *Evaluation basics: A practitioner's manual.* Beverly Hills, CA: Sage. A kind of "how-to-do-it" guide--getting started.

Patton, M.Q. (1986). *Utilization-focused evaluation.* Beverly Hills, CA: Sage. Light, but right on target. If the results aren't used, it was a meaningless evaluation.

Payne, D.A. (1974). *Curriculum evaluation: Commentaries on purpose, process, product.* Lexington, MA: D.C. Heath.

Popham, W.J. (1988). *Educational evaluation.* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall. Who said textbooks can't be informative as well as entertaining?

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation--a systematic approach.* (5th ed.). Newbury Park, CA: Sage. Comprehensive and interdisciplinary.

Royse, D. (1992). *Program evaluation* (An introduction). Chicago: Nelson-Hall. An excellent overview of the process with helpful coverage of technical and pragmatic issues.

Scriven, M. (1991). *Evaluation thesaurus.* (4th ed.). Beverly Hills, CA: Sage. All you wanted to know but were afraid to ask.

Shadish, Jr., W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation. Theories of practice.* Newbury Park, CA: Sage. A solid foundation with insightful perspective. Will help develop framework.

Stufflebeam, D.L.,et al. (1971). *Educational evaluation and decision making.* Itasca, IL: F.E. Peacock. The CIPP model is described in excruciating detail.

Talmadge, H. (1982). Evaluation of programs. In H.E. Mitzel (Ed.), *Encyclopedia of educational research* (5th ed.) (pp. 592-611). New York: Free Press. A nice succinct overview of the general dimensions of program evaluation.

Tuckman, B.W. (1985). *Evaluating instructional programs*. (2nd ed.). Boston: Allyn & Bacon. Some very practical suggestions and illustrations.

Worthen, B.R., & Sanders, J.R. (1987). *Educational evaluation (Alternative approaches and practical guidelines)*. New York: Longman. Lots of very useful suggestions, checklists, and advice on how to do it.

2

CRITERIA FOR EFFECTIVE AND ETHICAL EVALUATION PRACTICE

One of the characteristics of an emerging profession is the public and private concern for issues related to the law, ethics, and professional standards. The concern for these issues focuses on both protecting the consumer and evaluator and producing meaningful evaluations, the results of which are appropriately used.

Values of every sort permeate these ethical, legal, and professional standards issues.

THE PLACE OF VALUES

Values play an important role in program evaluation from at least two standpoints. The first point at which values are asserted, or should be asserted, is the identification of those objectives and goals that have evaluative priority. A decision about which objectives are most important should be made (Stake, 1970). Second, judgments are continually being made as performance data are contrasted with objectives. Significant value judgments come into play in assessing how great a discrepancy between expected and observed data is needed before action is required. This is the standard or criterion-problem.

Judgments are involved at many different points during the completion of an evaluation study. In fact, the decision to do a study is itself a value judgment. In addition, several other judgments must be made. The role of judgment will depend on the amount of objective data available for decision making. The following list suggested by Brownell (1965) highlights some points on which judgments must be made.

- Determination of appropriate grade level for evaluative study.
- Selection of appropriate subjects.
- Length of study.
- Identification of objectives in common and those specific to curricula involved.
- Determination of type of study to be undertaken (cross sectional, longitudinal, comparative, and so on).
- Decisions about nature of data to be collected.
- Selection of data-gathering instruments available or decision to develop original devices.

- Selection of appropriate control mechanisms aimed at uniformity of treatment.
- Selection of appropriate analytic procedures.
- Interpretation of findings.

When the teachers have taught, the students have studied, the administrators have administered, the supervisors have supervised, and the consultants have consulted, the practical limitations of the climate for evaluation and common sense will,despite recent extraordinary technological developments, play the most influential role in the design and implementation of an evaluation program.

THE IMPORTANCE OF ETHICS

The presence of values alerts us to the responsibility for exercising ethical professional behavior and judgment. Many issues are included in this sphere. A large number of these crucial issues relate to potential problems surrounding data--gathering instruments and procedures. The *invasion of privacy* issue is important since the use of instruments that in fact do invade a respondent's privacy or are perceived to invade privacy can result in either suspicious data or withdrawal of the subject(s) from the study. An allied problem deals with confidentiality. Evaluators frequently tell participants that their data will be kept in confidence, hoping that the declaration of confidentiality and promised anonymity will help insure more objective and candid responses. If that confidence is abused, the entire project could be lost. There is an incident described in the context of a large midwestern statewide project dealing with the assessment of leadership personnel. The major data base came from teacher evaluations of local building administrators. The instruments were returned unsigned, but the principal collected them instead of allowing the school librarian to take charge of them. There were sufficient demographics on the forms (e.g., years of teaching experience, degree held, subject taught) that the principal could identify individual teachers. When word got out, all hell broke loose. Needless to say, that school was lost from the project and, perhaps even worse, the future likelihood of any evaluation taking place in that school may have been permanently damaged. Although it was the principal who acted unethically, the evaluator still had the ultimate responsibility to control his management procedures so that the "peeping principal" should have been blinded.

It is now the policy of most school systems to require the use of *informed consent* forms for research and evaluation studies. This protects the participants but may operate to sensitize them to the treatment, thereby decreasing or eliminating a possible treatment effect for the "experimented" group. Conversely, knowledge about the purpose of the study obtained

through the process of informed consent could bring about the John Henry effect. This effect comes about when the control group performs as well as or better than the experimental one because they tried harder, perceiving themselves as being disadvantaged or discriminated against.

Another area of concern rests on the distribution of treatments over school groups. If I have a new textbook or science curriculum that I think is better than anything available, do I not have an ethical responsibility to share it with all, not just the experimentals? To some extent this problem can be handled by working out a treatment time schedule so that everyone will at sometime experience the treatment, but it will be so organized that I can gather relevant contrasting data from experimentals and controls.

Evaluators have responsibilities to sponsors, participants, and audience. Straton (1977) suggested that we address some general questions while considering those responsibilities. He suggested that we ask the following questions.

1. To what extent is the evaluator a servant of the sponsor or an audience of an evaluation study and to what extent an autonomous professional consultant?
2. Does the evaluator have responsibilities to audiences not recognized by the sponsor?
3. Is there some subordinate set of values or professional code of behavior to which an evaluator should respond when a conflict develops between the interests of two groups concerned with the evaluation study or between the evaluator and one of these groups?
4. Are there areas within which the evaluator has absolute authority, such as in study design, instrument development, or data analysis?
5. Should an evaluator have unlimited access to existing information and sources of information?

It should be obvious that these questions are of such a magnitude of concern that they need to be resolved before the evaluation gets under way.

The Abuse of Control/Contrast Groups: Another Ethical Problem

The abuse of control groups stems not so much from the treatments applied or assumed to be applied (or not applied), but from the misuse of data derived from inappropriate control groups. There are accidental and convenience control groups just as there are accidental and convenience samples. Failure to equate through appropriate selection procedures or statistical adjustments can obviously lead to invalid decisions. (We are here assuming that evaluation is decision oriented rather than conclusion oriented.)

The other abuse heaped on control groups surrounds the failure to document the specific parameters of whatever treatment it was that was applied. Obviously, if we don't know what happened to the control group, replicability is lost. Also, any rational basis for interpreting group differences is lost without a description of the treatments.

A primary consideration in evaluating the abuse of control groups relates to several specific ethical issues (Katz, 1972; Sjoberg, 1975; Diener & Crandall, 1978; Conner, 1980). The denial of a potential benefit to a group or individual is a serious consideration. If they are eligible they have a right to be "treated." The key here, however, is the word potential. It is assumed that the major reason for doing the evaluation is in fact to determine if the innovation, program, or project will produce worthwhile effects. Once the potential benefits have been demonstrated to be actual, then we can apply the "new" treatment across the board to all eligible subjects or groups.

Most new treatments once conceived and parameterized are not field-tested across the board. This would obviously be true of major curricular reforms. There eventually exists a need to secure limited funding to "tryout" the innovation. If federal or state funds are secured, there perhaps exists another potential ethical issue namely, the use of resources for a limited group to the exclusion of a larger potential clientele. Again, the resolution may rest on the fact that the benefits of the innovations have yet to be demonstrated.

A final ethical issue surrounds the answer to the question: Do members of the control groups need to be informed that they are part of an evaluation study? The importance of this question will obviously interact with the nature of the treatments. If the treatments will have an effect that causes the members of either group to be less acceptable in a cognitive, affective, or physical sense, then informed consent would be required. If the experimental treatment being imposed is simply another version of the treatment being experienced by the controls then the answer to the question takes on less importance. Full disclosure in the majority of school program evaluation projects is probably not necessary and in fact may inhibit a valid evaluation.

All of these ethical concerns and procedures are aimed at protecting the general welfare of our clients and participants. Children, teachers, parents, and schools are not guinea pigs to be experimented with indiscriminately, but if they are to reap the benefits of improved programs and procedures, they must also contribute relevant data obtained under controlled conditions. Conditions controlled to protect them as well as provide for a valid evaluation.

PROFESSIONAL STANDARDS AND METAEVALUATION

Scriven (1991) defines metaevaluation as "...the evaluation of evaluation--indirectly the evaluation of evaluators--and represents the ethical and scientific obligation when the welfare of others is involved" (p. 228). He goes on to note that metaevaluation is a professional imperative and obligation. It can be accomplished formatively during the design state or summatively upon a completed evaluation.

Sanders and Nafziger (1976) have identified 11 general criteria that are crucial in the evaluation of evaluation. They are briefly outlined below. Each heading represents a subset of criteria.

General Design Criteria

A wise man once said, if you don't know where you are going, you may end up somewhere else.

Scope: Are all significant aspects of the evaluation being addressed in the evaluation plan? These would include inputs, processes, as well as outcomes. The scope should be broad enough so that not only are criterial concerns focused on but also variables that might adversely affect the evaluation.

Relevance: Are all the data that are being collected responsive to the information needs of the audiences and relevant to the objectives of the program? The question here is really one of data validity for the intended use of the information.

Flexibility: Is the plan open enough such that changes in objectives, audiences, or evaluation data could be accommodated? One characteristic of evaluation that takes place in naturally occurring situations, particularly those that take place over a long period of time, is that many planned and unplanned changes occur. There are changes in personnel and objectives in particular that necessitate changes in data specification and collection procedures.

Feasibility: Practical considerations should relate to such factors as schedules, budget, personnel availability, and data availability. Overall

concern is for a frugal, diplomatic, prudent, and realistic design. Failures in programs or in evaluation can frequently be traced to this subject of criteria. Be careful of special interest groups (Renzulli, 1972).

Data Collection and Processing Criteria

An evaluation is only as good as its data.

Replicability: The information-gathering devices and procedures should be selected and applied in such a way as to ensure reliability. Methods should be built into the design to check on replicability, particularly if less than totally objective data are to be collected (e.g., observations or qualitative information).

Objectivity: The basic concern here is for control of biases. Bias may have impact on data source, method of collection, procedures for processing, or interpretation and reporting. Some control could come from the use of external data collectors or processors, panels to aid in data interpretation, or the use of already demonstrated unbiased instrumentation.

Representativeness: When information needs are complex and broad of scope, methods need to be applied so that representativeness is achieved. Sampling procedures in general, and item-examinee matrix sampling methods in particular, could be successfully applied in response to this criterion.

Reporting, Presentation, and Communication Criteria

The following two criteria generally relate to the utility of the evaluation. If an evaluation is to have utility, practical and scheduled reports must be made available to relevant audiences.

Timeliness: This straightforward criterion requires that the evaluation data be communicated to the decision maker(s) on schedule. Do relevant audiences get the data when they need it?

Pervasiveness: Do the audiences get the data when they need them and do all the relevant audiences get the required data? The object of the evaluation is both the audiences and those who may be affected by the data.

Propriety or Prudential Criteria

Standards are needed to help insure that the evaluation is conducted in an ethical and legal manner, with due regard to the rights of data sources and audiences.

Ethical Consideration: This criterion includes a whole set of considerations related to rights of privacy, full and frank public disclosure, confidentiality, and the use of human subjects. Not only are the ethics dictated by the law involved but so are the ethics dictated by one's profession. Also considered here would be fiscal responsibility.

Protocol: Not only is professional courtesy involved for ethical reasons but failure to follow such procedures can sabotage an entire evaluation because of destroyed cooperation.

These "Elegant Eleven" criteria are very helpful in seeing the big picture, but what is needed is a more detailed set of criteria or guidelines. These have been provided by the Joint Committee on Standards for Educational Evaluation (1981, 1994).

The *Standards* are for both users and producers of evaluations. The 1981 document describes 30 separate standards which are collected under four general headings: Utility, Feasibility, Propriety, and Accuracy. The Joint Committee foresaw many potential benefits in developing the Standards. Among these were (a) development of a common language base, (b) creation of a general set of rules for dealing with a variety of problem areas, (c) provision for a set of working definitions and a framework for research, (d) declaration to the public of professional standards, and (e) provision for a basis for self-regulation and accountability. In addition to accountability the *Standards* can be used to help define, contract, budget, staff, report and utilize an evaluation. Each of the 30 standards contain (a) a description of the standard, (b) an overview, (c) guidelines for use, (d) pitfalls in using the standard, (e) caveats (potential mistakes), and (f) an illustrative case. Following is an overview of the 1994 *Standards*.

Utility Standards: The utility standards are intended to ensure that an evaluation will serve the practical information needs of given audiences. These standards are:

U1 **Audience Identification**. Audiences involved in or affected by the evaluation should be identified, so that their needs can be addressed.

U2 **Evaluator Credibility**. The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that their findings achieve maximum credibility and acceptance.

U3 **Information Scope and Selection**. Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs and interests of specified audiences.

U4 **Valuational Interpretation**. The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.

U5 **Report Clarity**. The evaluation report should describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood.

U6 **Report Timeliness and Dissemination**. Evaluation reports and significant findings should be disseminated to clients and other right-to-know audiences, so that they can be used in a timely fashion.

U7 **Evaluation Impact**. Evaluations should be planned, conducted, and reported in ways that encourage follow-through by members of the audiences, so that the chances of the evaluation being used are improved.

Feasibility Standards: The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

F1 **Practical Procedures**. The evaluation procedures should be practical, so that disruption is kept to a minimum and needed information can be obtained.

F2 **Political Viability**. The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

F3 **Cost Effectiveness**. The evaluation should produce information of sufficient value, so that the resources expended can be justified.

Propriety Standards: The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

P1 **Service Orientation**. Evaluations of programs, projects, and materials should be designed to assist organizations to provide services of high quality, so that needs of learner development are met.

P2 **Formal Obligations**. Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.

P3 **Rights of Human Subjects**. Evaluations should be designed and conducted, so that the rights and welfare of the subjects are respected and protected.

P4 **Human Interactions**. Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation, so that participants are not harmed or threatened.

P5 **Full and Frank Reporting**. The evaluation should be full and fair in its presentation of strengths and weaknesses of the object being evaluated, so that strengths can be built upon and problem areas addressed.

P6 **Disclosure of Findings**. The formal parties to an evaluation should ensure that oral and written evaluation reports are open, correct, and honest in their disclosure of pertinent limitations and findings, so that the right to know by persons affected by the evaluation, and any others with expressed legal rights to see the results, is respected and assured.

P7 **Conflict of Interest**. Conflict of interest, frequently unavoidable, should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.

P8 **Fiscal Responsibility**. The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that there is no question about how evaluation resources are spent.

Accuracy Standards: The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the object being evaluated.

A1 **Object Identification**. The object of the evaluation (program, project, material) should be sufficiently examined, so that the form(s) of the object being considered in the evaluation can be clearly identified.

A2 **Context Analysis**. The context in which the program, project, or material exists should be examined in enough detail so that its likely influences on the object can be identified.

A3 **Described Purposes and Procedures**. The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.

A4 **Defensible Information Sources**. The sources of information should be described in enough detail, so that the adequacy of the information can be assessed.

A5 **Valid Measurement**. The data-gathering procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is sufficiently valid for the intended use.

A6 **Reliable Measurement**. The data-gathering procedures should be chosen or developed and then implemented in ways that will assure that the information obtained is sufficiently reliable for the intended use.

A7 **Systematic Data Control**. The data collected, processed, and reported in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed.

A8 **Analysis of Quantitative Information**. Quantitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

A9 **Analysis of Qualitative Information**. Qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

A10 **Justified Conclusions**. The conclusions reached in an evaluation should be explicitly justified, so that the audience can assess them.

A11 Impartial Reporting. Reporting procedures should guard against distortion by personal feelings and biases of any party to the evaluation, so that the evaluation reports fairly reflect the evaluation findings.

A12 Metaevaluation. The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, on completion, audiences can closely examine its strengths and weaknesses.

The *Standards* can set the standard for evaluation practice. Practicing card-carrying evaluators have seen and felt the impact of their presence.

Prevalence of Ethical Evaluation Problems

Both evaluators and consumers need to be concerned about potential violations of ethical and professional evaluation standards. Two recent studies have indicated the extent od selected violations. Brown and Newman (1992) surveyed three groups of evaluators: those with little or no knowledge, a moderate knowledge, and an experienced group. Following is a list of the most frequently reported problems:

1. Evaluator changes the evaluation questions to match the data analysis.
2. Evaluator promises confidentiality when it cannot be guaranteed.
3. Evaluator makes decisions without consulting with the client when consultation has been agreed to.
4. Evaluator conducts an evaluation when he or she lacks sufficient skills or experience.
5. Evaluation report is written so partisan interest groups can delete embarrassing weaknesses.

The authors also found that utility and feasibility standards, in general, were perceived as those most frequently violated. The more serious violations, however, were of the propriety and accuracy standards.

A recent unpublished 1992 survey by Dr. Michael Morris of the University of New Haven sheds additional light on the extent of evaluation standards. His data from 459 members of the American Evaluation Association revealed that 65% of the respondents said that they had encountered ethical problems in their work. Following is a list of the four areas where respondents reported they had encountered the most frequent problems together with a specific incident.

1. **Reporting Findings**: Evaluator is pressured to alter his/her presentation of the findings (62%).
2. **Misinterpretation and Misuse of Findings**: Results are suppressed, ignored, or not used by client/stakeholder (32%).
3. **Identifying Key Stakeholders**: At the contracting stage it's clear that a client/stakeholder has already decided what the evaluation results "should" be or has a morally dubious reason for wanting the evaluation to be conducted (55%).
4. **Disclosure Agreements**: Although not pressured by someone to violate individual confidentiality, the evaluator is concerned that reporting certain results would represent such a violation (31%).

The overlap in the results of the two surveys is obvious. Evaluators and consumers need to be sensitive to potential ethical and legal issues.

Many things can be learned from using the *Standards*. Following is an example of such a lesson.

AN EXAMPLE LESSON FROM THE *Standards*

It was noted earlier in this chapter that the *Standards* were guidelines that could be used before evaluations, during implementation of the evaluations, and after the evaluation as metaevaluation. Following is a description of the author's experience reflecting all three applications of the *Standards*.

College professors are often accused of living in "ivory towers." The teacher of evaluation methods is not shielded from such indictments (Payne, 1982b). So it was with great expectations that the author accepted a contract to evaluate a computer-assisted instruction program recently instituted in a local minority high school. The task was indeed challenging. Despite the practical problems of conducting an evaluation in a fast-paced and complex multicultural educational setting, I was enthusiastic.

Following is a narrative account of that evaluation. After the summary is a discussion of lessons learned by doing the evaluation. An attempt is made to place the evaluation in the context of the *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee on Standards for Educational Evaluation, 1981).

Background and Setting

The use of microcomputers in the public schools is increasing every year, ranging from tutoring to simulation, managing, teaching programming, and doing drill and practice. Students benefit in areas such as learning efficiency, learning effectiveness, pacing, convenience, and enrichment. Theoretically, microcomputers should free teachers to use their time and talents with students more effectively and eliminate redundant and repetitious tasks. Several studies have also shown a positive impact on cognitive achievement as well as such affective educational outcomes as attitude toward school and specific school subjects, and self-concept (Kulik, Bangert, & Williams, 1983). Such outcomes were anticipated when the Atlanta Partnership of Business and Education, Inc. (a nonprofit public/private change-facilitation agency) and Control Data Corporation joined forces in the installation of a computer-assisted instruction (CAI) program at an inner-city high school. The installation included 20 off-line Control Data 110 microcomputer units, primary flexible-disk drives, and display terminals. Software came from the multipotential PLATO (Personal Learning and Training Opportunity) population of 8,000 lessons drawn from 60 disciplines. The intent of the evaluation reported here was to assess the impact of this two-semester installation.

Methodology

Because of practical limitations and administrative problems, two separate data collection designs were required for the two semesters. In the fall semester, a retrospective control group design was employed (Howard et al., 1979). In the spring semester a conventional posttest--only control group design was used with matched groups, which resulted in a quasi-control group design.

Sample

Fall semester students included 346 users and 358 nonusers who were primarily ninth and tenth graders. They averaged about 15 hours of CAI experience over the semester. Courses included language skills, physics, general math, biology, chemistry, social studies, and algebra. Spring semester data were collected from 406 users and 126 nonusers averaging 17 hours of CAI in the following courses: survival english, general and human biology, economics, algebra, U.S. history, political behavior, and math. Overlap between the two semester user/user and nonuser/nonuser groups was 77%.

Variables measured were as follows: Academic Self-Concept (9 items) (Brookover, LePere, Hamachek, Thomas, & Erickson, 1965); Locus of Control (20 items) (Crandall, Katkovsky, & Crandall, 1965); General Attitude

Toward School (15 items) (Instructional Objectives Exchange, 1972); Student Opinions About Computers (19 items); Teacher Evaluation of CAI/PLATO (23 items); Attitude Toward Specific Subject (15 items) (Remmers & Silance, 1934); Final Exam scores (0--100); and Final Course Grade (A=1,...,F=5). The General Attitude Toward School and Academic Self-Concept variables were not measured during the spring semester, since analyses of these variables in the first semester did not prove sufficiently productive or informative.

Analyses

Analysis of covariance was the major statistical procedure applied to the fall data. For the first semester a retrospective approach to data collection was applied. Students were asked to respond, for example, to questions about their attitudes toward school using two frameworks. An item like "Each Morning I Look Forward to Coming to School" was used in the context of "How I Feel Now" versus "How I Felt at the Beginning of the School Year." Correlations between the "Now"and "Beginning"responses over 15 weeks for Self-Concept, Locus of Control, General Attitude Toward School, and Attitude Toward Specific Subject were .74, .57, .30, and .67, respectively, for the user group. For the nonuser group they were .65, .65, .42, and .59. The later scores were used as the covariate. Spring analyses involved using a conventional analysis of variance technique.

Results

Fall semester data yielded the following results:

1. PLATO users of language skills, chemistry, and algebra had significantly higher final-exam scores and end-of-semester grades than nonusers.

2. Students and teachers evaluated PLATO software and CDC hardware very positively. Students were particularly enamored of the freedom to control the pace of their own learning (96%). Teachers preferred the clarity of objectives in the material (78%).

3. Although the statistical results were mixed, evidence suggested that students' feelings about their responsibility for and control of their academic progress and attitudes toward specific subjects may be positively affected by even brief exposure to CAI.

Spring semester data suggested the following conclusions:

1. There were no significant differences between final course grades of PLATO users and nonusers. Significant final-exam score differences in favor of the PLATO user group were found for students in algebra and mathematics.

2. Students were overwhelmingly positive in their evaluation of the CAI software and hardware. Some 84% said that their "basic skills have been improved."

3. A significantly higher academic self-responsibility score was found for PLATO users in Biology relative to nonusers. This effect was significantly intensified when PLATO was used twice a week rather than once a week.

4. Two hours a week of PLATO had a significantly more potent positive effect on Attitude Toward Specific Subject than one hour a week.

5. Greater gains in internal locus of control and attitude toward school subject were observed for the user group relative to the nonuser group from fall to spring semester.

Despite numerous methodological problems, which are discussed later, it appears that the CAI had a modest positive impact on the students and to some extent on the teachers by enhancing their "computer literacy." The results are similar to those reported by Jenkins and Dankert (1981).

Lessons from the Evaluation and Standards

The *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981) is a set of 30 criteria grouped into four functional standards categories: utility, feasibility, propriety, and accuracy. These standards have proven useful over the years in assisting in the design of evaluation studies as well as in metaevaluation. The intent in the present section of this article is to illustrate briefly how the standards influenced or could have influenced the CAI evaluation.

Utility Standards

The utility standards deal with such important issues as audience identification, evaluator credibility, and several issues related to evaluation report clarity, dissemination, and timeliness. In the present project there was a problem in identifying who the audiences were: the hardware/software vendor, the students, the teachers, the school administrators, the board of education, or the intermediary sponsoring/facilitating agency. Although the vendor (Control Data) was picking up the tab for the evaluation and insisted on having considerable input into the evaluation design, it was the sponsoring agency that pulled all the pieces together. The school personnel had a vested interest because a positive evaluation might mean that the board would provide money to keep the project going. As a result several reports were generated to meet the information needs of the various audiences.

The timeliness issue could not be dealt with effectively. It is often the case, particularly in education settings, that budget decisions must be made before the close of the fiscal year. As a result, a full report of evaluation results are not available for input into the decision-making process. This was the case with the present project. The board, after looking at some dollar figures and negotiating with the vendor, decided it was too expensive to continue CAI at the project level. A modest installation however, is, in operation.

Feasibility Standards

The feasibility standards are concerned with practicality (particularly with regard to data collection), political viability, and cost-effectiveness. Every attempt was made to minimize disruption of the ongoing program in terms of implementing the CAI program also with regard to data collection. Practicality needed to be balanced against accuracy of data. The evaluator, in consultation with the teachers and school-level project coordinator, developed, adopted, or adapted instrumentation. The teachers whose students used the CAI had the responsibility for administering the inventories and working among themselves to develop the departmental achievement tests. Reliance on many different "data collectors" also led to problems of incomplete data sets, since motivation and investment in the project varied from teacher to teacher.

Propriety Standards

The standards in this category reflect the important legal and ethical issues in conducting an evaluation. The protection of human rights and the public right to know and have access to relevant data and results are of paramount concern.

A scheme was developed so that students could code their own survey forms using a combination of the number of letters in their mothers' first names, an aggregate of student birth date (e.g., $1+14+70 = 85$), and their usual weight. It was felt that the gain in anonymity outweighed the potential loss in data due to inconsistent coding. Unfortunately, a tryout of the method proved to be a bust. The old reliable "Give me your name and I will protect it" plea was successfully used so that data could be correlated.

Periodic meetings were held with teachers and members of the board of education to provide progress reports and to stimulate interactions about the project, particularly as regards the responsiveness of the data being gathered and reported.

Accuracy Standards

The largest number of metaevaluation criteria deal with technical adequacy. The reader will recall that accuracy issues are related to standards such as adequacy of reliability, validity, data control, and data analysis. It is frightening to see how inadequacies in implementation can lead to analysis problems. Following are some shortcomings in program implementation that led to less-than-optimal data analysis and interpretation.

1. The treatment had been under way three weeks before the evaluator was consulted.
2. Different classes used CAI for differing amounts of time, both within and across subject areas.
3. Different classes used CAI in different ways.
4. Nonuser (control) classes were identified on the basis of school officials' judgments.
5. Students occasionally had to share terminals, as there was not enough hardware or software to go around.
6. Monitoring of students' use of hardware and software was inconsistent at best.
7. Instrumentation used was adopted from material originally developed primarily with white middle-class populations--quite different from the participants in this project. Internal consistency analyses did reveal, however, acceptable levels of reliability for the students completing the instruments in the project. Item language was adapted as necessary.

All these threats to internal and external validity would probably cause Professors Campbell and Stanley to blanch in horror. At least the evaluator was aware of them and could ameliorate overly zealous interpretations of project data and impact accordingly.

A Final Note

One of the dramatic lessons for this evaluator relates to how evaluation designs *evolve*. They are not simply adopted. It is an iterative process, with evaluator and stakeholders interacting at many critical points. Compromises are made as practical problems present themselves. For example, the evaluator is not called in until after the project is under way: Scriven calls this the "point of entry problem." Evaluation was also seen as evolutionary in the present project as variables were dropped from the second-semester data collection and some analyses were based only on first-semester results. The net effect was to limit and tighten the study and to reduce use of teacher and student time.

Another interesting phenomenon occurred when standards in one category interacted with Standards in another to create a hybrid problem. For example, the lack of control over data collection (failure to meet a feasibility standard) led to a unit-of-analysis problem (failure to meet an accuracy standard). A large amount of missing data forced us to abandon use of the teacher as the unit of analysis. An individual-by-course unit was therefore used. As noted previously, a potential achievement-criterion problem was avoided by having departmental exams.

Despite competing "publics,"the complex demands of practicality, and the uniqueness of the CAI experience for these students, a modestly meaningful (and, yes even satisfying) evaluation was carried out. We all learned something from the experience.

In summary, to be forewarned is to be forearmed. Walk softly, carry a copy of the *Standards*, and protect yourself at all times.

COGITATIONS

1. How can a consumer of evaluation results use the Program Evaluation Standards? How about the evaluator?
2. What role do ethics play in the evaluators interactions with clients, stakeholders, and targets?
3. How and where do the personal values of the evaluator come into play in conducting an evaluation?
4. What kinds of ethical problems are most likely to occur in public school evaluation?

SUGGESTED READINGS

Anderson, S.B.,& Ball, S. (1978). *The profession and practice of program evaluation*. San Francisco: Jossey-Bass.

Evaluation Research Society Standards Committee (1982). *1982 Evaluation Research Society Standards for program evaluations*. (New Directions for Program Evaluation, No. 15). San Francisco: Jossey-Bass.

House, E.R. (1990). Ethics of evaluation studies. In H.J. Walberg & G.D. Haertel (Eds.). *The international encyclopedia of educational evaluation*. New York: Pergamon.

Newman, D.L.,& Brown, R.D. (1992). Violations of evaluation standards. *Evaluation Review*, 16(3), 219-234.

Perloff, R.N.,& Perloff, E. (1980). *Values, ethics, and standards in evaluation*. (New Directions for Program Evaluation, No. 7). San Francisco: Jossey-Bass.

3

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 with 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

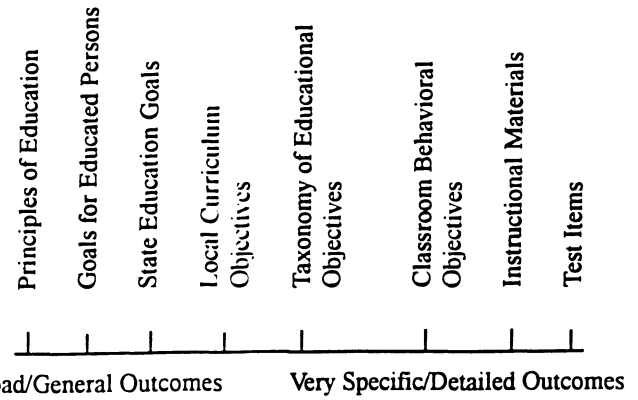


Figure 3-1 Degrees of Specificity of Educational Outcomes

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.

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

Stage	Questions To Be Asked	Data Sources
(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.

TABLE 3-1 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 unimodel, 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.

4

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	
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?	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	
1. What goals were chosen when program was initiated? 2. Why were these goals chosen over other possibilities?	1. What designs were proposed? 2. What alternative designs were rejected? 3. Why was the winning design chosen?
PROCESS (Activities)	**PRODUCT** (Results)
DECISION MAKING	
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?	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	
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?	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
1. Examination of enabling legislation 2. Description of selection procedures for each nomination area 3. Student ratings of congruence of their personal objectives relative to those of their program area	1. Semantic differential given to both faculty and students 2. Measure of creative personality given to students (Torrance's What Kind of Person Are You) 3. Students took Cattell's Sixteen Personality Factor Questionnaire 4. Objective biodata form for students 5. Faculty responded to measure of classroom management philosophy (Pupil Control Ideology)	1. Audio tapes of instructional sessions. These data subjected to Ober interaction analysis 2. Administration of Classroom Activities Questionnaire which allowed students to rate nature of instructional experience (Based on Taxonomy of Educational Objectives)	1. Student ratings of self vs. program contribution to extent of mastery of program objectives 2. Extensive follow-up survey of previous 10 years' worth of participants 3. Post-semantic differential

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
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.	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 vail 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.

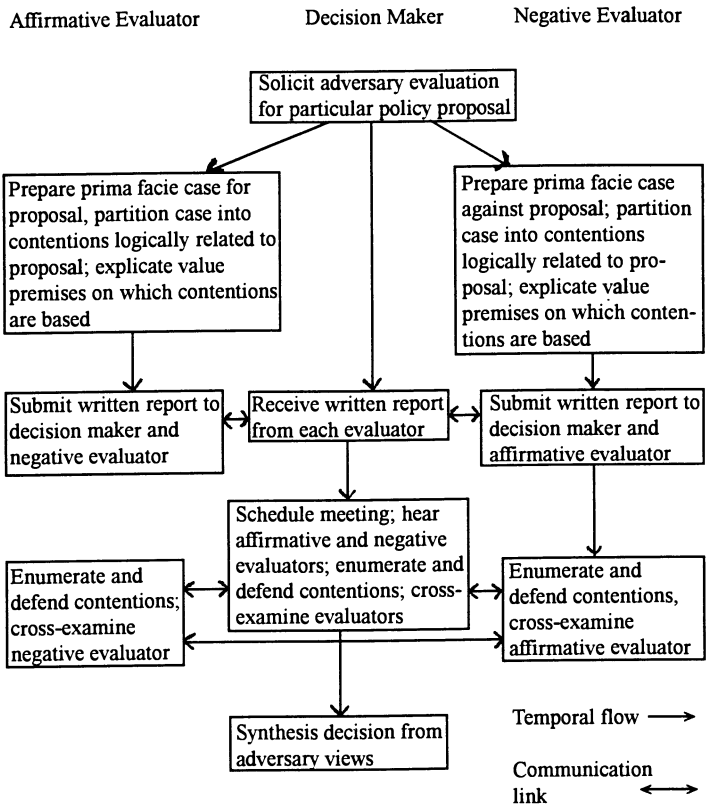


Figure 4-1 Representation of Adversary Evaluation Metaphor Reprinted by Permission of Copyright Holder: Center for Research on Evaluation, Standards and Student Testing. University of California at Los Angeles. Source: Kourilsky, 1973.

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
1. Variety of data may be introduced	1. Need for equally able and motivated adversaries
2. Tends to force higher quality of evidence	2. Expensive method
3. Can operate to diminish unwitting bias	3. Heavy burden on arbiter/judge/decision maker
4. Variety of points of view and opinions presented	4. No control over decision maker having hidden agenda
5. Opportunity to examine opposing data and views	5. Not all propositions amenable to adversarial approach
6. Tends to force assumptions to surface	6. Judicial model can generate false confidence
7. Applicable to situations having complex outcomes	7. Possibility of extremism
8. Facilitates communication	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.

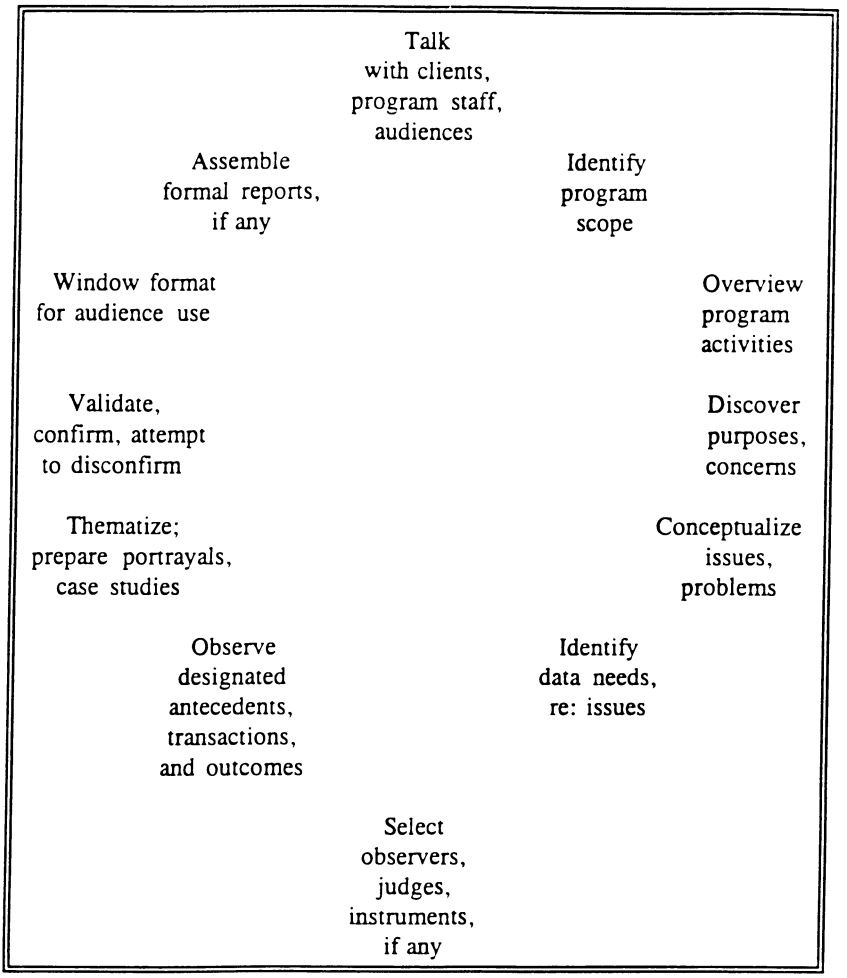


Figure 4-2 Events in Responsive Evaluations (Stake, 1983; reprinted by permission of author)

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
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	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.

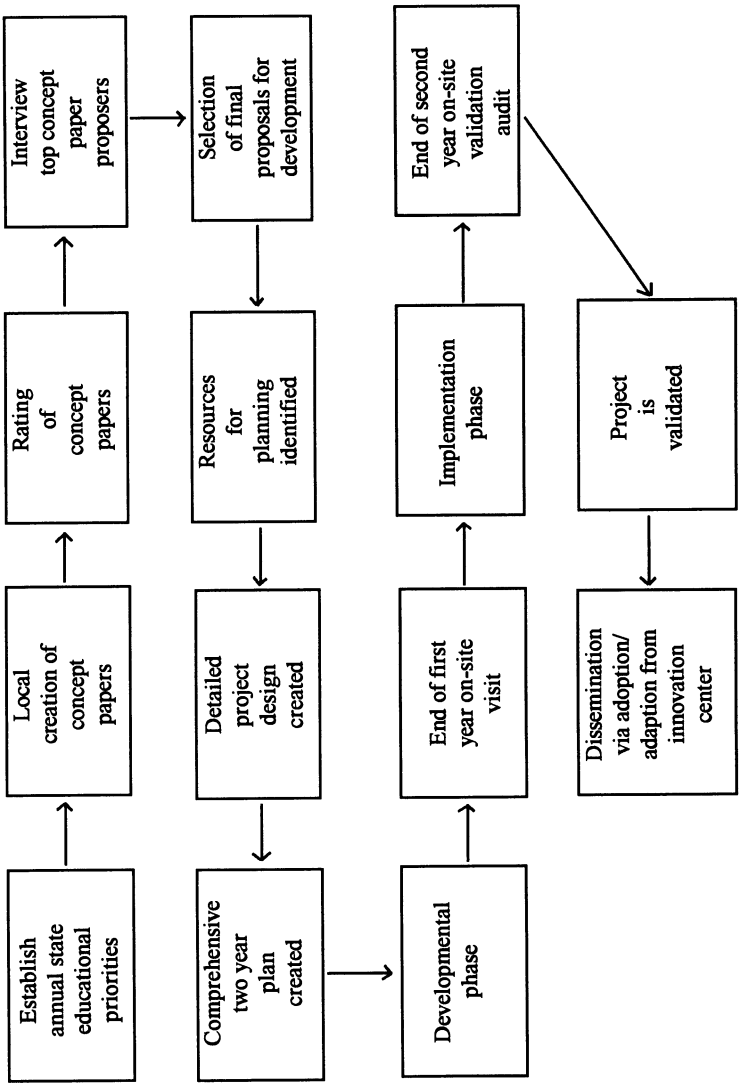


Figure 4-3. Life History for Innovative Project Development and Validation

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
1. Provide independent assessment relative to developer.	1. Evaluation accomplished separate from consumer/ practitioner.
2. Is cost-effective relative to consumer.	2. Requires high degree of expertise.
3. Helps establish standards for product quality.	3. May require considerable resources.
4. Sensitizes consumer to producer hype and dangers of anecdotal advertising.	4. Possible bias of evaluator.
5. Decreases likelihood that "untested" program will be foisted onto the consumer.	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 "waysof 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 connoierseurship 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.

5

QUANTITATIVELY ORIENTED DATA COLLECTION DESIGNS

Decision making requires relevant data. The problem facing the program and project evaluator is how best (meaning efficiently and effectively) to gather that data. The data must be gathered in a systematic fashion and in such a way as to allow the project impact be seen in as clearly defined form as possible.

It is frequently helpful to formalize a somewhat complex process such as project evaluation into a conceptual paradigm, flow chart, or other schematic. Following is an example of such a schematic.

	Fall 1994			Spring 1995						
E (Bulldawg Elementary School)	O_1	O_7	X	O_1	O_2	O_3	O_4	O_5	O_6	O_7
C (Contrast School)	O_1	O_7	X	O_1	O_2	O_3	O_4			O_7

Where:

- X = Relevant Instructional Program
- O_1 = Wide Range Achievement Test - Revised (WRAT-R)
- O_2 = Student Survey of Feelings About School
- O_3 = Parent Survey of Perceptions of School Effectiveness
- O_4 = Student Attendance Data
- O_5 = Student Discipline Referral Data
- O_6 = Teacher Sick and Leave Day Data
- O_7 = Elementary Reading Attitude Survey

This schematic represents a typical evaluation design for a school-based innovation project. In this case the innovation was a K--3 continuous progress program in four classrooms. Students at each of the levels were randomly assigned to one of four classrooms. A variety of activities was mounted to maximize the impact of cross-age grouping. Extensive staff development was completed. Curriculum materials were revised or written. Reporting procedures had to be recast in a narrative form. A variety of stakeholders were involved in this evaluation. Their involvement required the

collecting of data from a variety of sources. A comprehensive plan had to be designed. What are the factors that must be considered in selecting or creating a design?

FACTORS AFFECTING EVALUATION DESIGN DECISIONS

The decision to select a specific design, and hence, a specific control or contrast group, involves the weighing of various factors that may impinge upon the project or program due to the specific circumstances surrounding the evaluation and the program. In general, two major influences appear appropriate. First, considerations involving the evaluation design itself must be addressed. Second, practical and political considerations must be assessed.

The usual evaluation design options involve the use of randomization techniques, matching, or the identification of some externally or internally equivalent group. Sometimes fate intervenes to prohibit or at least inhibit the use of any of these three approaches.

What are some of the forces which operate to reduce the likelihood of being able to assign subjects randomly to treatment and control (term used interchangeably with contrast) groups, matched subjects, or to establish equivalent groups?

Scope of the Treatment

The scope of the treatment may prevent the use of an optimal evaluation design. The project administrator and evaluator may be faced with a situation where the only politically expedient (or possible) course of action is to assign all students to the treatment group (complete or total coverage). In this case, the only choice with regard to establishing a contrast group is to search for another school or system where matching or establishing a similar group may be possible. Project/program costs and staff preparation may also prove to be additional barriers to selecting a design and a contrast group. For example, where considerable release time for teachers or additional training is required, project administrators may have difficulty in locating (either internally or externally) a sufficient number of participants.

Project/Program Purpose

The purpose or purposes of the treatment may affect design considerations. Many programs have as their primary objective the solution of a local problem or a specific set of problems. Although in a general sense one might argue that some projects are developing or testing theory, most agencies require documentation of a need in order for a project to be funded. Should the project administrator be faced with serious deficiencies in student accomplishment or staff performance, the number of potential participants

may greatly increase. The perceived worth (or potential for success) of the treatment may create a condition similar to a bandwagon effect. Systems, classrooms, or schools don't all share the same problem in need of solution.

Concerns of Parents

Parent support or parent antagonism for projects developed to impact upon students may affect the decision as to which students (if any) are to participate in the project/program. If the potential for solution of a significant problem is high, all parents may want their children to participate. If the perceived potential for harm is high, no parents may allow their children to participate.

Extent of Treatment

Finally, the extent to which experimental manipulation will occur has a direct bearing on acceptance or participation in the project/program. The project in which only minor changes in routine occur has a greater probability of acceptance than the project in which major changes in routine occurs. In fact, projects calling for major changes in routine may generate sufficient reaction to alter or halt treatment.

From the parent point of view or the teacher point of view, two questions perhaps summarize the dilemma faced by project administrators in the selection of an evaluation design and hence a contrast group: "Who wants to participate as part of the contrast group in a highly successful project?" or "Who wants to participate in the treatment group of a project that is either a flop or is perceived as potentially harmful to the participants?"

ELEMENTS OF A DATA COLLECTION DESIGN

The three major components of a data collection design are included in the foregoing example of the continuous progress program. They are consideration of (1) application of a "treatment" during a particular time frame, (2) the collecting of data from referent groups, and (3) the specification of the data collecting devices. All three of these elements will be dictated by the nature of the problem being investigated and evaluation questions asked. The design simply specifies *what* data are to be gathered from *whom* and *when*. The general nature of the design awaits creation of a more detailed data management plan (see Chapter 8).

The Use of Contrast Groups

Although there surely are relevant questions surrounding external validity and generalizability and the use of contrast or control groups, the big problem is with internal validity. The basic question is whether we can describe in sufficient detail plausible explanations of the hoped-for differences between groups. So many factors can influence the so-called equivalence of

groups. One need only study the classic "threat list" of Campbell and Stanley (1966) to appreciate that fact (See Table 5-1 later in this chapter). But even randomization is not going to control *all* relevant sources of group contamination. The appeal of randomization technique probably derives from its antecedents in traditional experimental research. Because of that "halo," the technique perhaps has received more accolades than it deserves. The technique can't control all relevant influences. In many instances we in fact want to include those so-called contaminating variables so that their unique interaction becomes part of the "treatment." These "influencing" variables should be free to exert their impact in a naturally occurring environment. The demand characteristics of the evaluation (e.g., student expectations of an improved self-concept) may or may not equate across groups. It is proposed that the term *contrast group* rather than *control group* be used in educational evaluation studies. This term simply refers to an existing or to-be-generated data set against which our "experimental" results are to be contrasted. It is usually the case that in most educational evaluation situations we do not have the luxury of having very extensive control of subjects, or in some cases treatments for that matter. The use of the term *contrast group* would, therefore, be more descriptive of the true state of affairs and in addition tend to remove evaluations from the domain of the traditional experimental paradigm by recasting the nature and focus of the contrast.

Although many evaluation designs are available that do not require the use of contrast or control groups (Cook & Campbell, 1979), most federal and state funding agencies require that such groups be part of the overall data collection and analysis design. The demand for contrast or control groups perhaps reflects an effort on the part of the "money leaders" to force more scientific rigor into the evaluation effort, thereby hopefully generating a more definitive answer to the problem addressed. It also may be perceived that evaluation designs with control groups are more credible and give the appearance of greater validity. Sometimes a norm-referenced external data base might be used, but in contemporary evaluation practice there is a definite affinity for classical experimental designs.

Horan (1980) has suggested that historically we have considered control groups to have received "everything but" the experimental treatment. It may be truer to say that in far too many instances "anything but" might be a better descriptor. But it is agreed that evaluation involves some kind of comparison. That benchmark might come from data generated from a contrast group in the design or be derived from an extant source (e.g., test manual statistics.) *A major criterion for almost any good evaluation design is the use of independent contrast data.* The comparison may be to some like-type group without the prescribed treatment or it may be some extant data base such as a set of national or state norms. The concern is to design our evaluations so

that, within practical limits, rival hypotheses may be ruled less plausible. The realistic evaluator is less likely than the "brass instruments" researcher or the experimental design obsessive to be concerned with causal inference.

Once a project administrator and evaluator have assessed the circumstances and have determined the evaluation design and the contrast group, other questions bearing upon the effective use of the contrast group must be addressed. One such question involves payoffs for both the participants and the decision makers in the contrast group: "What will we get from this experience?" If the contrast group receives no benefits in regard to program, professional development, or other kinds of rewards, a reluctance to participate can probably be anticipated. Project administrators must also decide what kinds of information will be presented to the participants and decision makers of the contrast group. In general, it would appear that the contrast group should receive all of the feedback from all measurements taken in the same time and manner as the treatment group. Finally, the issue of competition must be considered. Where the contrast group resides outside of the school or district, old rivalries may stir up a competitive attitude. Beware of the John Henry effect where the control outperforms the experimental. For example, a project involving students in two high schools (one constituting the control) where athletic competition has been keen in the immediate past may be affected by transfer of the competitiveness to the objectives of the project. One obvious method for avoiding this situation is to select a control school where there is no history of keen competition with the treatment school. Other possibilities to avoid the influence of competition include selective statistical analysis (e.g., ANCOVA), use of project data, and other information.

Categories of Designs

Three general categories of designs will be considered here: (1) experimental, (2) quasi-experimental, and (3) nonexperimental. These three classes of design differ in the degree of control over the treatment that they allow. We are attempting to isolate and measure the impact of our program or project. We want, in essence, to hold constant as many as possible and feasible extraneous factors and influences that might "contaminate" our results. Pedhazur and Schmelkin (1991) note that there are four major methods of exerting control in the design of studies.

First and foremost is *randomization*. Randomization as used here refers to the process of selecting or assigning whatever the sampling unit and ultimate analysis unit is (e.g., individual student, teacher, classroom, school) to a condition (e.g., competing treatments, a treatment and a control) so that each unit will have an equally likely chance of being in each of the conditions.

Chance will determine placement. Tables of random numbers or computer programs can be used effectively to accomplish randomization. Although less efficient and perhaps not useful with extremely large data sets, such manual methods as flipping coins, rolling dice, or drawing numbers from a hat can be used. An approximation of random selection can be accomplished by randomly entering a list of names or identification numbers and then taking every *n*th name as needed. The intent is to "equate" groups so that everyone begins on the same footing and that any potential factors that might influence the outcome measures, independent of the treatment, are controlled or at least confounded (i.e., don't have a systematic effect). There are some evaluators who don't believe in randomization. They say that rare events can and do happen. Yes, they do, but only rarely!

Doing project evaluations in the real world usually does not allow for the luxury of employing complete randomization. In the foregoing example of the continuous progress program, although the students were randomly assigned to classrooms, the contrast school was not randomly selected along with the experimental school. There are a limited number of *statistical controls* available that will help us make adjustments for the lack of equivalence between the two schools. A very powerful technique is analysis of covariance (ANCOVA). This procedure allows post intervention scores or an outcome measure to be adjusted for initial differences between an experimental and contrast (control) group. The adjusting variable (covariate) is usually a premeasure equivalent or similar to the post measures, but any variable(s) thought or known to be correlated (statistically and conceptually) with the dependent or outcome measure could be used. One of the important corollary benefits of using ANCOVA is greater power and precision in the analyses. Partial correlation is another technique useful in holding constant control variables. We might, for example, be interested in the relation between scores on the Graduate Record Examination and graduate school grade point average, holding constant or controlling for age correlationally. Another approach would be simply to run the correlations separately for different age groups.

The actual selection of the treatment(s) in the programs and projects represents another method of control. The choice of intervention in intensity and duration can be controlled or *manipulated* thereby allowing for an assessment of its impact.

Finally the independent and extraneous variables can be controlled by *including* or *excluding* them from influence. Variables that might be hypothesized to be related to the outcome measures or treatment can be controlled by selection. Variables such as sex, age, race, or socioeconomic class can be controlled by limiting a study to a particular group (e.g., females) or the evaluation could be conducted using separate but intact groups (e.g.,

all females versus all males). The variable of gender would thereby be held constant. Using this procedure may, of course, limit the generalizability of the results. The technique is particularly useful when the anticipated influencing variable is categorical.

When one thinks of the myriad of variables that can influence the results of any evaluation, randomization must be considered as an effective control mechanism, particularly for large samples. Failing that, do the best you can with the other methods, but always be cautious in interpreting your results.

THE VALIDITY OF DATA COLLECTION DESIGNS

The literature of classical experimental research is replete with caveats to the investigator about all the factors that can mess up the results. Campbell and Stanley (1963), Cook and Campbell (1979), and more recently Campbell (1986) have helped several generations of investigators understand threats to design validity. The original set aggregated by Campbell and Stanley (1963) included internal and external validity that reflected on the control of the treatment and generalizability of the results, respectively. Cook and Campbell (1979) added statistical conclusion and construct validity to the list. These related to the inferences from statistical tests and the treatment-outcome measure match, respectively. Campbell (1986) has changed slightly the focus and interpretation of internal and external validity. Internal validity has been renamed Local Molar Causal Validity. Translated, that means that there is greater emphasis on controlling the extraneous complex interacting factors that influence implementation of the project at the local level. There is also greater concern now for the theoretical relationship between the treatment and the outcome measures. The external validity concept has been recast as Proximal Similarity. The renaming of this concept is an attempt to capture the uniqueness of treatment-site interaction. Selecting a representative sample for the evaluation may not be as important as describing the conceptual and actual interaction of treatments, measures, populations, settings, and times. Exportability will then be to those environments where there is greatest similarity. Documentation of the experimental and contrast environments is, therefore, an essential element in the design process.

Because of the familiarity of the evaluation community with the original labels of internal and external validity, we will continue to use them here.

What are the threats to design validity and how can they be controlled?

Table 5-1 contains a summary of nine significant factors that can distort (either positive or negative) the evaluation of a program or project.

Any one of the influences described in Table 5-1 can be further confounded by interactions with any of the other influences. Interactions with *selection* in particular can be particularly detrimental to design validity.

TABLE 5-1 Summary of Threats to Internal Validity of Data Collection Designs

Category	Description	Example
History	Events related to outcome of study occur during implementation.	Local outbreak of AIDS occurs during conduct of AIDS awareness program in high school.
Maturation	Naturally occurring uncontrolled changes in subjects that are related to outcome.	Elementary school physical education program shows increased skill development although it could be simply due to aging.
Testing	Repeated data collection may result in increased scores. Operation of practice or memory.	Short duration of attitude toward drug program in middle school requires pre- and post-measurements to be gathered only weeks apart.
Statistical Regression	A real phenomenon where post-treatment scores of those at extremes move toward "average."	At-risk preschoolers selected because of low scores on screening tests show significant gains after one year of intervention.

Category	Description	Example
Instrumentation	Change in instrumentation over course of study. Changes in calibration or scoring accuracy.	Lack of comparability in two forms of high school chemistry test used to assess impact of lab-oriented curriculum.
Mortality	Attrition of study subjects occur at higher rate for experimental group or contrast group.	During three-year project aimed at enhancing English skills of Hispanics finds that more of the contrast than the experimental group has left the area.
Selection	Differential or self-selection biases groups.	Volunteer schools in the contrast group tended to come from low socioeconomic areas in a study of the impact of a self-esteem building program.
Diffusion/imitation of Treatments	A competing treatment to that voluntarily adopted by the experimental group is adopted by the contrast group.	The contrast teachers also attend the staff development sessions on cooperative learning methods meant for the experimental group.

Category	Description	Example
Compensatory Rivalry/resentful Demoralization	Differential effort as a result of real or perceived differential treatment.	The contrast school, not having received a computer lab, tries harder to do a better job teaching elementary math.

Validity is the control of the treatment effect. Take, for example, the possible interaction of selection and history. Due to a defect in the selection or assignment process, for example, more upper socioeconomic students got into the experimental program. A measure of progress in a language arts program might be enhanced artificially, as another example, because of greater availability of academic support mechanisms; books, computers, and so on. It is in fact these interactions, particularly of the treatment with (1) selection, (2) history, and (3) the setting in which the evaluation takes place, that contribute to decreased generalizability (for external validity) of the results. Lack of control, one of the extraneous factors, or a high degree of uniqueness of the group(s) or subject(s) contribute to the difficulty in replicating the results.

One can see how important it is to monitor the treatment as it is being implemented. Lack of fidelity in application of the innovative treatment can totally confound the results. One should in fact *evaluate* the implementation of the treatment.

Concern for Unintended Effects

Another design issue relates to the problem of unintended program effects. It is here where perhaps using a goal-free approach makes a great deal of sense. An evaluator might say, "Don't bias my data gathering by telling me what you expect; let me see what happened for myself." Wolf (1984) has likened the search for unintended outcomes to looking for a black cat in a dark room on a moonless night. It is almost that difficult but also important. We can be both happily surprised or depressed with unintended outcomes. The present chapter began with a description of a continuous progress nongraded program at the elementary school level. Among the unexpected negative outcomes of this project was the finding that kindergarten discipline referrals were greatly increased relative to previous years. This was attributed to the fact that teachers in the experimental classrooms had to deal with four age groups (5,6,7, and 8). The first six months of a kindergartner's school life can be traumatic. The continuous-progress teacher had to deal with *all*

children and perhaps had less opportunity to deal with the kindergartners' special problems. On the positive side was the finding that students saw themselves as special and experienced a concomitant increase in self-esteem. How did these unexpected effects become apparent? Primarily through observation, *ex post facto* examination of discipline referral data, and focus group interviews with students (see Chapter 6). Observation of the in-progress program is a particularly useful method of data gathering related to implementation.

What implications do these factors have for the actual design of an evaluation? In the following section nine data collection designs will be presented and the various advantages and disadvantages discussed.

DATA COLLECTION DESIGNS

It was noted in the previous section that "control" was the key to a good design. If one cannot or does not want to randomize, then other methods might be employed. In the language of experimental psychology the application of randomization procedures should result, for example, in the creation of equivalent groups. One will receive our treatment (or experience the innovative program or project) and the other will act as a reference point, benchmark or comparison group against which data from the experimental group can be contrasted. Rarely can we randomize and get a *control* group after the fact. As was noted earlier, the alternative term *contrast* group is suggested. Every effort will be made to make the groups comparable. Perhaps data from records and files could be used. In the illustrative data collection design presented at the beginning of this chapter, it was found that the school means on the state criterion referenced tests in reading and mathematics were within three points of each other and that the percent of students on free or reduced lunch was 63% for Bulldawg Elementary and 58% for the contrast school. The judgment of assumed comparability was made.

In doing educational evaluation there never exists a setting where a "no treatment" condition exists. Everybody gets something, some more or less than others. Is it better to give than to receive? There is always a "traditional treatment" going on. It may not be systematic or continuous, but it exists. One of the tasks that an evaluator must complete then is to describe *both* the experimental and contrast treatment. Comparing the applications of these two treatments is what it's all about.

The traditional design symbology will be employed here: X = a treatment and O = an observation, measurement, or data collection event. A convention will be used here, however, where observations with the same subscript will mean the same or equivalent measurement e.g., parallel test forms, no matter when they are taken. Consider:

$$O_1 \; X \; O_1$$

In this case the same measurement was used on a pre-treatment/post-treatment basis.

One final introductory comment is that we are here specifying differences between experimental and quasi-experimental designs on the basis of a failure to apply randomization procedures in the quasi-experimental situation. In addition, quasi-experimental designs are differentiated from nonexperimental (sometimes referred to as pre-experimental) designs because the later do not reflect any randomization or manipulation of a treatment variable.

Experimental Designs

Our first experimental design is the ever popular and usually effective *Pre-test--Post-test Contrast Group Design.*

R	Group 1	O_1	X	O_1
R	Group 2	O_1		O_1

Randomization has been accomplished (R). Multiple observations either pre or post could be made if so desired, and they usually are. Because of randomization the major threats to internal validity have been controlled. We will obviously only complete our analyses on subjects (students) who were present for the entire study. Using randomization controls for regression and selection effects and the pre-test allows for examination of the effect of mortality. In addition, presence of a contrast group controls for history, testing, and instrumentation. Finally, the combination of randomization and the presence of a contrast group control for maturation. Interpretation of results is reasonably straightforward. The design could obviously be expanded to include several different treatment groups. Extending the basic two-group design to multiple groups and measurements might yield a configuration such as the following:

R	Group 1	O_1	X_A	O_1	O_2	O_1
R	Group 2	O_1	X_B	O_1	O_2	O_1
R	Group 3	O_1		O_1	O_2	O_1

Such an extension would allow us to examine competing treatments (and perhaps conduct cost analyses; see Chapter 8) and do follow-up studies (the third O_1). In addition, the introduction of an observation (O_2) which the

evaluator felt was a measure of a relevant outcome could be accomplished after the treatment, thereby avoiding any chance of testing X treatment interaction or sensitization (reactivity.)

A possible weakness of this design is the presence of potential interaction between pre-test and treatment. Subjects might be "sensitized" to the treatment simply having taken a pre-test. An attitude toward drugs inventory might cause students to think about their knowledge and feelings regarding this topic even before an educational program was completed. In that sense the pre-test becomes part of the treatment. Students might seek more information on their own or converse at length with their peers about their reaction to the problem.

A useful design to control for the treatment of pre-testing interaction is the *Post-Test--Only Contrast Group Design*, represented as follows:

R	X	0_1
R		0_1

There is no pretest included in the design. Again, multiple treatments and observations of a number of different outcome (dependent) variables could be gathered. The design does not allow for assessing the effect of "mortality" because it lacks a pre-test. If the size of the sample is large and the duration of the study is relatively short, then mortality will probably not be a factor. Remember that one of the meanings of control rests on an evaluator's ability to assess the effect on uncontrollable extraneous influences even if they can't actually manipulate the variable.

If it is important for the evaluator in fact to gauge the amount of gain or change over the duration of the study on a measure that poses a potential pre-test--interaction threat, then perhaps the "mother" of all designs, the *Solomon Four Group Design*, could be used.

It is represented as follows:

R	Group 1	0_1	X	0_1
R	Group 2	0_1		0_1
R	Group 3		X	0_1
R	Group 4			0_1

Groups are constituted by random assignment to one of four separate units. Two groups are pre-tested, and one of them receives the treatment. They are

both post-tested. What you have, of course, is our old friend the Pre-test--Post-test Contrast Group Design. One of the remaining two groups is post-tested and one of these receives the treatment. What we have here is another old friend, the Post-test--Only Contrast Group Design. Contrasting Groups 1 and 3 in the above diagram will allow us to assess the impact of the pre-test-treatment interaction (and mortality) if it was generated. The Solomon Four Group Design enjoys all the advantages of the Pre-test--Post-test Contrast Group Design and the Post-test--Only Contrast Group Design. Drawbacks of the design are that it (1) requires a lot of units (e.g., students, classrooms), and (2) is not very practical in public school settings.

Controlling for Reactivity with Retrospective Pre-testing

Often it is not possible to identify an acceptable control or contrast group. In addition, when sensitive treatments are involved (e.g., attitudes toward drug use) pretesting may generate a so-called pre-test effect which reacts with the dependent measure. To accommodate these difficulties, Campbell and Stanley (1966) have suggested the use of *retrospective pre-testing*. Such a procedure allows the treated group to act as its own control, a particularly useful approach when self-report dependent measures are involved (Howard *et al.*, 1979). An allied problem is the phenomenon of "response-shift bias." Assume for a moment that you are going to be a participant in a workshop on problem-solving skills. The pre-test contains an item like the following: "I am a good problem solver." You strongly agree with the statement and so respond. After getting into the workshop you find that you really aren't a very effective problem solver. At the end of the intervention you are confident about the skills you have developed and again, but for a different reason, strongly agree with the statement, "I am a good problem solver." Obviously, a "no-difference" conclusion would be reached when evaluating the workshop. One method of "finding" some relevant contrast data involves, as the title suggests, actually gathering *ex post facto* pre-test data. One could, for example, have our workshop participant fill out an end-of-workshop questionnaire, providing a summative evaluation of its effects and values. The participant would then be asked to respond to the questionnaire as s/he would have if s/he had taken it prior to the experience. (Often it is not physically or operationally possible to gather pre-test data. A few years ago the author was involved in evaluating the State of Georgia Governor's Honors Program (GHP) for the academically and artistically talented. This enrichment experience for rising high school juniors and seniors lasted for eight weeks in a college campus setting. The lack of availability of anything remotely resembling a contrast group was evident. Several post-experience measures were gathered which basically served the same purpose as pre-testing. Students, for example, were asked to contrast

their summer experience with the regular school treatments. Illustrative are the following questions:

Which holds the student more responsible for work?

In which do students try out their ideas more?

Which provides greater opportunity for close contact with teachers?

Possible answers were: GHP; Regular school; No difference.

Retrospective testing has been used primarily with affective measures, but some researchers have used the technique successfully in the cognitive domain (Rippey, Geller, & King, 1978). In our GHP evaluation students were asked, for example, to make retrospective judgments about the extent to which the program contributed to their mastery of selected instructional objectives.

So much for experimental design. What do I do if I cannot randomly assign units to conditions?

Quasi-Experimental Designs

It was noted that the public school project and program evaluator usually do not have the opportunity to apply randomization procedures to evaluation studies. A very frequently employed design that approximates within certain parameters the *Pre-test--Post-test Contrast Group Design* is the *Non-equivalent Contrast Group Design*. The only difference between this design and the former is that randomization procedures have **not** been used.

Group 1	O_1	X	O_1
Group 2	O_1		O_1

The dotted line between the groups indicates lack of randomization. Two or more groups might be employed and, again, multiple measures possibly used. The lack of randomization allows for the influence of sources of invalidity not present with the pre-test--post-test contrast group design--namely, regression, and possible interaction between selection and variables such as maturation, history, and testing. Since the groups are not equivalent, a frequently used statistical procedure is analysis of covariance. In an effort to help ensure the closest similarity between the experimental and contrast group(s), matching procedures are sometimes used. One method of "constructing" a contrast group has been suggested by Payne and Brown (1982).

Constructing Matched Groups

Although it is not held in the highest regard by all quantitative methodologists, the use of matching procedures can provide meaningful contrast data useful in assessing evaluation data. An argument against matching is that for every variable on which individuals or groups are matched there may be many others of equal or greater importance. Despite this potential shortcoming, matching can be a valuable design technique. The following method, the *Aggregate Rank Similarity Method*, was first described by Brown (1980) and involves the matching of an experimental classroom, school, or school system against a population of possible contrast groups. A more elaborate system has been described by Sherwood, Morris, & Sherwood (1975). One distinct advantage of the matching procedure is that the evaluator has control of the variables, and in many cases data on the more important ones are readily available. Matching can take place either within or outside a district. Let's look at an example to illustrate the procedure.

An evaluator is interested in identifying a school system to use as a contrast system while evaluating a new K--12 science curriculum that involves integrating both career education and environmental concerns. Before a list of matching variables is generated, a sample of potential contrast groups is identified which are geographically contiguous to the "experimental" system. A list of independent (matching) variables is then assembled. Criteria for inclusion in the list could be (1) justification for the relevance of the matching variable found in the research literature, and (2) availability of on-site data. Table 5-2 contains three such variables. The raw data for each variable for each potential contrast system are subtracted from the values of the data for the experimental system, and the *absolute* values entered. Next these absolute values are ranked from smallest to largest for each variable. The ranks are summed across the variables and the system with the smallest aggregate sum

TABLE 5-2 Illustration of Aggregate Rank Similarity Method for Matching Systems

	MATCHING VARIABLES									
	% of Pupils on Free Lunch			Average Daily Attendance			Per Pupil Expenditure			Sum of Ranks
SYSTEM	Raw Data	d*	Rank	Raw Data	d	Rank	Raw Data	d	Rank	
EXPERIMENTAL	76			6284			4395			
Contrast 1	62	14	2	4387	1897	3	5031	636	2	7
Contrast 2	60	16	3	5934	350	1	3977	418	1	5
Contrast 3	98	22	4	3847	2437	4	3828	567	4	12
Contrast 4	74	02	1	5166	1118	2	5127	732	3	6

*Note: Absolute deviation of contrast system data from experimental data.

is selected as the contrast system. In the case of the data in Table 5-2 it would be System 2. A great variety of independent (matching) variables might be identified; for example, scores on standardized tests, pupil/teacher ratios, data from observation instruments, or credentialling/certification data. Selection is limited only by the creativity of the evaluator and project administrator. In addition, the selected independent matching variables might be differentially weighted.

The *Interrupted Time Series Design* is another valuable quasi-experimental design. It is particularly useful when large amounts of comparable archival data are available. Impact is assessed by examining the change in measurements after the introduction of an innovation or treatment. The basic design such as this

$$0_1 \quad 0_1 \quad 0_1 \quad X \quad 0_1 \quad 0_1 \quad 0_1$$

can be augmented with an equivalent or quasi-equivalent contrast group. When used with a contrast group it is referred to as a multiple time series design. Used in the multiple form it might look like this:

Group 1	0_1	0_1	0_1	X	0_1	0_1	0_1
Group 2	0_1	0_1	0_1		0_1	0_1	0_1

The interrupted time series design controls for the history and instrumentation threats to internal validity, which were not controlled in the single form of the design. This design can be used effectively when the collection of repeated measures (e.g., test scores, attendance data, disciplinary referrals) is an ongoing and naturally occurring activity. It is particularly useful when the entire population must receive the treatment (i.e., complete coverage with the intervention).

A final quasi-experimental design is the *Institutional Cycle Design*. This design (a variation on the counterbalanced design) is again useful when all subjects must receive the treatment. A school, for example, wants all elementary students to experience a new environmental AIDS awareness unit. We could take the entire elementary student body (probably using the classroom as the unit) and assign half of them to the first implementation of the environmental unit. They would experience the unit (X) and then be post-tested. The second group would take the environmental post-test as a pre-test; then they would experience the unit and be post-tested.

The data collection design would look as follows:

Group 1	X	0_1	(Y)	0_2			
Group 2		0_1	X	0_1	0_2	(Y)	0_2

We have two measures of program impact. Group 1 Post versus Group 2 Pre, and Group 2 Post versus Pre. The design could be jazzed up by adding another treatment (Y) to each group. The design is an interesting combination of cross-sectional and longitudinal approaches. The design does, however, suffer from a failure to control for three problems: maturation, selection, and possible multiple treatment interactions.

Data Analysis from Nonequivalent Contrast Group Designs

It is not possible to address this enormous analysis topic here with the space limitations at hand. Suffice it to say that there are many technical issues involved in evaluating data from nonequivalent contrast group designs, although it may be as Lord (1967) has said: "With the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups." There do exist, however, a number of useful analysis procedures that can be applied to data generated from quasi-experimental designs (Reichardt, 1979). Such techniques as analysis of covariance, value-added analysis, regression-discontinuity, and selection-modeling are illustrative valuable data analysis procedures. In addition, the reader is alerted to the volumes noted in the Suggested Readings section at the end of this chapter for current information about analysis procedures appropriate for most of the frequently used data collection designs: Freed (1991), Keppel (1991), and Pedhazur and Schmelkin (1991).

An Illustrative Nonequivalent Contrast Group Evaluation Study

It was a dark, rainy, thundering, overcast, dreary, stormy day when the aspiring evaluator was called into the department head's office. As chairperson of the reading education department, she had been contacted by a major instructional materials developer and solicited to serve as a field-testing site for a new set of six computer-assisted instruction modules aimed at teaching the teaching of reading techniques to prospective teachers. There were eight sections of the relevant course available next quarter. Enrollment averaged about 20--25 students per class. The dean and department head requested that she take on this project (for no extra compensation, as is usually the case) citing the benefits of visibility for the department and college,

possible publications, and presentations at professional meetings. The evaluator met with the developer of the computer-assisted instruction modules, department head, and dean, and evolved the following modest evaluation questions:

1. Will the new CAI reading education modules result in increased knowledge about techniques for the teaching of reading?
2. Will the new CAI reading education modules result in greater learning than that resulting from using the current traditional techniques?
3. Will the new CAI reading education modules result in positive attitudes toward computer-assisted instruction about the teaching of reading?

There are virtually an infinite number of possible designs that could have been created to answer the three evaluation questions. What follows is one of many possible approaches. Among considerations involved in creating this design and the ultimate report were the:

1. Desire to utilize already existing classes.
2. Need to gather contrast data from CAI and non-CAI groups.
3. Relatively short period of time in which to conduct the study.
4. Lack of funds to develop elaborate instrumentation.
5. Inability to control assignment of students to classes at registration due to scheduling needs.

Design

Since the evaluation questions required both *absolute* information (questions 1 and 3) and *relative data* (question 2) a nonequivalent contrast group design was used. Four classes were designated as CAI and four non-CAI. Naturally occurring enrollment determined which class a student attended.

In addition, periodic monitoring of how well instructors were using the new materials was undertaken to maintain fidelity of implementation. An in-service program had been held by the materials developer on the use of the CAI materials with the four CAI instructors.

The design could be symbolized as follows:

CAI Classes		0_1	0_2	X_c	0_1	0_2	0_3	0_4
Non-CAI Classes		0_1	0_2	X_n	0_1	0_2	0_3	

Where
- 0_1 = Pre and post measures of knowledge of teaching of reading techniques
- 0_2 = Pre and post measures of attitude toward computer assisted instruction
- 0_3 = Immediate post-test on individual modular units
- 0_4 = Interviews with selected CAI students
- X = Treatment (Interventions, Instruction, Project Program) C = CAI, N = Non-CAI (Traditional)

It should be noted that although we were unable to randomly assign students to classes (and thereby treatments), the only way of looking at growth is by getting some kind of change data over time, and therefore we had to use pre and post measures and a contrast group.

Data-Gathering Instruments

The Reading Techniques Knowledge Inventory (RTKI) was a 75-item, 5-alternative, multiple-choice test designed specifically for the project. It had a pilot-test-calculated Kuder-Richardson Formula 20 internal consistency reliability of .73. The inventory contained content questions as well as problem application exercises. The items were scored right or wrong (and yes or no).

The attitude scale employed was a 17-item inventory using a 5-point Likert rating scale (Strongly Agree. . . Strongly Disagree). Following is a sample item:

Computers are one of the best ways to teach.

This instrument, Attitude Toward Computerized Instruction (ATCI), had a reported test-retest reliability of .87 over a four-week period and had been shown in other studies to be moderately related to actual classroom performance.

Each of the six curricular modules had a 25-item summative test covering only the material of that unit. They were administered to individual students when they had concluded each unit. The items were scored right or wrong. The CAI modules dealt with the same content in the same sequence as in the traditional classes.

Data Collection and Storage

Modern optical scanning technology allows for the collection and processing of very large amounts of data. The data were scanned directly into a micro computer. Item booklets with standard optical scan sheets were sent to each instructor. Data were gathered from all participants using the same directions.

Data Sources

Students in the two groups (CAI and non-CAI) were instructed with their relevant respective materials for one 11-week quarter. Part of the first week was given over to organizational and orientation matters and pre-testing, and part of the last week was devoted to post-instruction data collection. The CAI materials involved six modules spread over a 10-week period. Each CAI lesson required about two and a half hours of working time. Students could work at their own pace. The CAI material was supportive of ongoing instruction and represented approximately 50% of the total instructional time.

Decision Rule/Standard Setting

The use of a decision rule is a way of incorporating the concept of standard/criteria setting (see Chapter 3). Several approaches could be taken. One could specify the percentages of specific instructional objectives that need to be mastered, either by individuals or groups, as a means of rendering a decision about effectiveness, or an overall mean score difference could have been specified. Another approach, the one taken in the present evaluation, 1s to use a statistical model to help make the judgments about effectiveness. A combination of descriptive data together with inferential statistical methods was used.

Data Analysis

The data collection design suggested the following kinds of analyses.

In order to answer Evaluation Question 1, the pre and post scores were contrasted for each of the module unit tests and for the RTKI. A correlated t-test was used (a test of differences between means for a single group of students).

Evaluation Question 2 required application of analysis of covariance on the pre--post modular and RTKI scores across the two groups (CAI vs. non-CAI classes). Use of this particular procedure allowed potential differences between the two groups before the new program was introduced to be adjusted or "equalized."

The final evaluation question dealing with differences in attitude between groups was assessed through application of a t-test of differences between correlated means for the same students.

Additional analyses could have been specified concerning differences between the effectiveness of the program for males as opposed to females, or the effect of amount of familiarity with computers on achievement and attitude.

Results

Table 5-3 contains a summary of the means and standard deviations of the RTKI scores. The class was used as the unit of analysis. This was done because of the potential unique interaction between instructor and student. The percent score was obtained by dividing the means by the number of dichotomous items (75).

TABLE 5-3 Summary of Pre-Test, Post-Test, and Mean Score Differences for Reading Techniques Knowledge Inventory CAI and Non-CAI Groups

		Pre-Test			Post-Test			
	n	Mean	%	SD	Mean	%	SD	Mean Gain
CAI	4	51.32	68	8.88	64.73	86	7.33	13.41*
Non-CAI	4	49.98	67	10.78	54.34	72	6.07	4.36
Differences		1.34			10.38*			9.05*

*This difference significantly greater than zero, $p < .05$. (Some analysts prefer simply to report the actual p-values).

The following interpretations appear justified from these data:

1. There were no initial (pre-test) differences between the CAI and non-CAI classes.
2. There was a meaningful knowledge gain for the CAI group but not the non-CAI group.
3. There were meaningful differences between the two groups at the end of the quarter (post-test).
4. The gains were significantly greater for the CAI than the non-CAI classes.

It appears that the CAI materials had a positive influence on knowledge acquisition. Whatever the reason--ability to self-pace, opportunity for review, or periodic within module testing--the CAI delivery system brought about enhanced learning. Note also that the variability of scores became less at the end of instruction. Such a phenomenon might be interpreted as reflecting positively on the reliability of program implementation--in other words, students were more alike at the end of the instructional experience than they were at the beginning.

Table 5-4 contains a summary of the total scores and subscores (by objective) for Module One. These are presented here as an illustration of the kinds of analyses carried out for each module. The subscores are tied to five general objectives, each of which was measured by five items. The data are presented simply to illustrate the kind of information that can be gathered. Such data can be used formatively to improve the instructional materials.

Following are the instructional objectives for the first module:

1. Associate different instructional practices with three conceptual frameworks (models) of the reading process.
2. Apply phonics generalization to decode nonsense words.
3. Identify the purposes of Language Experience Activity.
4. Apply syllabication rules.
5. Recognize the characteristics of a particular Language Experience Activity.

It can be seen that the module is generally working in the hoped-for way, with the total scores being higher for the CAI group. Two of the subparts of the module are perhaps in need of attention--namely, subparts 2 and 4. Referring back to the objectives we can see that objectives 2 and 4 tend to be a bit more technical and complex than the others. This is a situation where the formative approach to evaluation will help us *improve* the curricular materials. Performance on the items related to objective 2 suggest that the content, materials, or instructional approach are not effective based on the low level of achievement. The comment also could be made about objective 4 where students are not performing well against an absolute criterion. In any event, improvements are needed.

TABLE 5-4 Summary of Means and Standard Deviations of Total and Subscores on Module One Test

Objective	Group	Mean	SD
One			
	CAI	3.14	1.11
	Non-CAI	2.38	.93
Two			
	CAI	2.67	1.24
	Non-CAI	4.01	1.86
Three			
	CAI	4.37	.88
	Non-CAI	3.72	1.32
Four			
	CAI	2.87	1.77
	Non-CAI	1.65	.56
Five			
	CAI	3.99	.74
	Non-CAI	1.44	.66
Total			
	CAI	17.01	5.36
	Non-CAI	13.40	4.98

The attitude data are summarized in Table 5-5. The rating scale means for this 17-item instrument can range from 17 to 85. The attitude data again favor the CAI approach. There is significant gain over time as well as a differential gain across groups. The absolute level end-of-treatment attitude is pretty positive at the conclusion of the quarter.

In summary, what do our data tell us about our 10-week instructional intervention? It appears that it works cognitively and that attitudes also have been positively influenced. The data also suggest that improvements can yet be made in the CAI materials and the instructional approaches used in selected modules. Other changes could be suggested by item analysis of the data from all instruments used in the study.

Education is concerned with the realization and utilization of human resources. Measurement and evaluation can significantly aid in their realization and utilization by providing reliable and valid data on where we have been, where we are, where are we headed, and how much we have accomplished.

TABLE 5-5 Summary of Pre-Test and Post-Test Means and Standard Deviations for Scores on the Attitude Toward the Computer-Assisted Instruction Instrument

		Pre-Test		Post-Test		
	n	Mean	SD	Mean	SD	Mean Gain
CAI	4	49.34	8.31	62.30	7.70	12.96*
Non-CAI	4	48.01	7.11	53.41	6.69	5.40
Differences		1.33		8.89*		8.56*

*Significantly different from zero, $p < .05$.

Nonexperimental Designs

This general class of designs is sometimes referred to as pre-experimental designs. They do not allow for the manipulation of the treatment or randomization. The designs are nevertheless helpful at the formative level in field testing materials or procedures. The first nonexperimental design is the *One-Shot Case Study*:

$$X \quad O_1$$

The evaluator simply describes outcomes apparently resulting from the application of a fixed treatment. The term *design* is used advisedly here as none of the major threats to validity are controlled. A well-done case study, if implemented by an accomplished qualitative evaluator (see Chapter 6), can nevertheless yield invaluable information about the impact of a program or project. Among the potential benefits of the case study method is the generation of hypotheses or questions to be answered in future studies. The evaluator must be extremely careful about drawing inferences from this approach. Its framework is exploratory at best. See books by Merriam (1988) and Yin (1984) for comprehensive treatments of the case study method.

A second nonexperimental design is the *One Group Pre-test--Post-test Design*:

$$O_1 \quad X \quad O_1$$

Its use involves pre- and post-testing a single group which has received a particular treatment. Rossi and Freeman (1989) refer to this design as a *reflexive* control design since the treatment group serves as its own control. The single group time-series design can also be called a reflexive design.

Uncontrolled factors in this design include history, maturation, testing, instrumentation, and selection interactions with a variety of other factors. If the time between pre- and post-observations is lengthy, these threats have the potential for significant impact. The inability to assess potential pre-test-treatment reactivity is a serious drawback of this design. As with the case study method, use of this design might be helpful during the early stages of product or project development. Again, be warned about drawing meaningful inferences from this design since by its very nature it results in rival hypotheses in most cases being more tenable than usual.

The *Static Group Comparison* is a slightly improved nonexperimental design. A case-study-like design is supplemented with some contrast data.

Group 1	X	0_1
Contrast Data		0_1

We might compare, for example, the general equivalency diploma (GED) scores of a group of Hispanic adults who had been working on a computer tutorial program preparing them to take the high school equivalency exam with a mean score for recent high school graduates. The normative test group provides the contrast group data. Another set of contrast or reference data could be derived from what Rossi and Freeman (1989) call a *shadow* control. Judgments from experts or program participants are gathered to serve as benchmarks for interpreting impact data. We might ask a group of experts to fill out a teacher evaluation form in a manner that would describe an "effective multicultural teacher." This profile could be used as a reference point for evaluating the impact on individual teachers of a staff development program. Still another kind of contrast group is sometimes referred to as a *generic* control. Extant data bases, such as the normative GED scores referred to earlier, can be used as comparisons against the outcomes of a particular intervention. State, local, or national indices are available through a variety of public and private agencies and publications.

In concluding this chapter, several observations need emphasis:

- All effective evaluation designs require the collection of contrast, benchmark, or comparison data.
- Evaluative designs evolve. The relationship between problem/question and method of seeking an answer is inseparable.

- Considerations of cost, nature of questions, nature of administrative and political constraints, and receptivity of decision maker(s) will influence evaluability (the viability of even doing the evaluation at all).
- The key to an effective evaluation design is the isolation and measurement of the impact of the treatment, intervention, or innovation.
- Don't let sampling procedures or statistical methods dictate the evaluation design.
- Generalizability is nice, but internal validity is essential.
- If approached in a systematic way, using accepted guidelines, intelligence, and common sense, any evaluation design can yield valuable information.
- Although technical aspects of the evaluation method are important, all will fail if the problem is not properly conceptualized.
- Anticipating the things that can go wrong and planning for their rectification or control is half the battle in the war for truth.

COGITATIONS

1. What are the major considerations that differentiate true experimental, quasi-experimental, and nonexperimental designs?
2. What design factors should be most important to the (a) evaluator, and (b) project director?
3. What designs would be best to use when the entire population of targets must be covered or served?
4. What are the advantages and disadvantages of using reflexive, constructed (matched), generic, or shadow contrast groups?
5. What is "experimental" about an experimental design? What is "quasi" in a quasi-experimental design? What is "pre" about a pre- or nonexperimental design?
6. Are the factors that influence design creation the same as those that effect design selection?
7. Are some threats to internal validity more important than others? Why?
8. What are the advantages and disadvantages of retrospective pre-testing as an approach to controlling testing by treatment interaction?

9. Under what conditions can nonexperimental studies be valuable?
10. Under what conditions would unintended effects be acceptable?

SUGGESTED READINGS

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally. The modern classic treatise.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation design and analysis issues for field settings*. Chicago: Rand McNally. What do you do when you can't randomize?

Freed, M.N. (Ed.) (1991). *Handbook of statistical procedures and their computer applications to education and the behavioral sciences*. New York: American Council on Education/Macmillan. The introduction to the major microcomputer software statistics packages (e.g., SAS, SPSS-X, SYSTAT, MINITAB) is particularly informative.

Keppel, G. (1991). *Design and analysis (A researcher's handbook)*. (3rd ed.) Englewood Cliffs, NJ: Prentice Hall. Very comprehensive, but not for the fainthearted.

Mohr, L.B. (1992). *Impact analysis for program evaluation*. Newbury Park, CA: Sage. Twelve designs in all their glory. Food for thought and action.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design and analysis (An integrated approach)*. Hillsdale, NJ: Lawrence Erlbaum. Eight hundred nineteen pages and 24 chapters of all one could possibly want to know, but the authors really communicate.

Popham, W.J. (1993). *Educational evaluation*. (3rd ed.) Boston: Allyn and Bacon. Chapter 10 contains a realistic overview of a variety of usable designs and hints on their implementation.

Taylor Fitz-Gibbon, C., & Morris, L.L. (1987). *How to design a program evaluation*. Newbury Park, CA: Sage. Easy to read yet comprehensive decision-trees help readers find their way in the forest of designs.

Trochim, W.M.K. (Ed.) (1986). *Advances in quasi-experimental design and analysis* (New Directions for Program Evaluation, No. 31). San Francisco: Jossey-Bass. Six important papers representing current conceptions and issues.

6

QUALITATIVE AND ETHNOGRAPHIC EVALUATION

Mary Jo McGee-Brown
University of Georgia

> *We do a lot of looking: we look through lenses, telescopes, television tubes. . .Our looking is perfected every day--but we see less and less.*
>
> F. Franck (1973, p. 3)

Rist (1980), in his article "Blitzkrieg Ethnography: On the Transformation of a Method into a Movement," expresses a number of concerns about the decline in quality of the conceptualization, process, and products of ethnographic research in education simply because the approach was "in vogue." Rist claims that "The term 'ethnographer' is now being used to describe researchers who neither studied nor were trained in the method." Rist raises other concerns such as researchers conducting hit-and-run research rather than designing and conducting longitudinal studies; researchers having simple description as the end in itself rather than exploration of the underlying cultural framework and deep meanings of participants; multiple-researcher multisite research focusing on breadth rather than single-ethnographer, single-site designs focusing on in-depth understanding; and entrance into a research site with preformulated research problems and concepts resulting in a predetermined approach to data collection and analysis rather than allowing them to emerge from extensive time and interaction with participants at the site. Rist predicted that as the number of untrained researchers and evaluators employing this method grows, the rationale for using the qualitative approach will be undercut, the resulting reports will be of poor quality, and disenchantment with the qualitative approach will be inevitable.

We are at a similar place in evaluation today. It is common for funders to want to know "what is happening out there" when they provide support for programs. That question requires a qualitative approach to data collection. It is generally assumed that untrained persons cannot conduct quantitative data collection, statistical analysis, and data interpretation. On the other hand, many have the misconception that "anyone can do qualitative research and evaluation" because the methods include observation and interviewing. While it is admirable that evaluators are looking to expand their tools, we are moving toward a disaster if qualitative evaluators are not systematically trained in the philosophical and theoretical assumptions underlying the approach, field strategies, design issues, data collection methods, and data analysis and

interpretation approaches. The author knows of one situation in a state where project evaluators are trained in a two-day workshop, an hour of which focuses on qualitative evaluation. The coordinator then feels that all workshop participants (most of whom have a quantitative background) participating in that workshop were qualified to design and implement qualitative components in their evaluations. Nothing could be further from the truth!

A RATIONALE FOR INTERPRETIVE INQUIRY

The underlying assumption of qualitative evaluation is that the perspectives and actions of all participants or stakeholders in a program are important. There are four primary reasons for selecting a qualitative approach in evaluation. They are to:

- *Discover* the meanings that the innovation, program or project has for persons across levels within at the site(s).
- *Observe* the effects of the innovation or change on behavior, actions, and interactions for all persons at the site(s).
- *Document* the process in the natural setting without manipulating any variables.
- *Assess* cultural changes that are a direct or indirect result of the program as well as determine the effects of the larger cultural context on changes associated with the program.

Qualitative inquiry is an umbrella term that includes many different research designs. The term *interpretive inquiry* is preferred (see Figure 6-1), because understandings from all qualitative methods of data collection by nature include multiple levels of interpretation. As Erickson (1986) notes, interpretive inquiry is more inclusive than qualitative, and it avoids the connotation that quantification is not used in interpretive research.

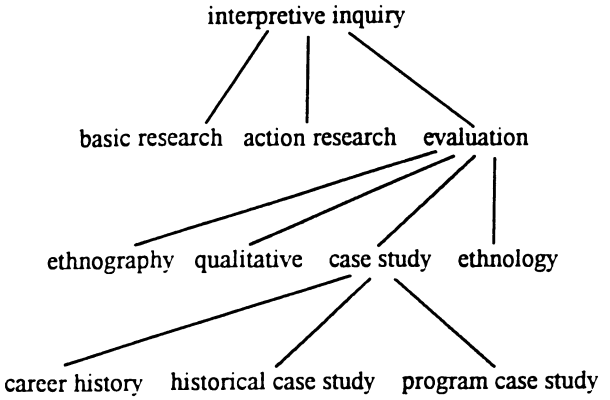


Figure 6-1 Interpretive Evaluation Designs

The key features of each of the interpretive evaluation designs are that (1) they are longitudinal, (2) there is a fieldwork component in which the evaluator collects data in context at the site(s) while experiencing the program first-hand with participants, (3) data are generally narrative in form (but numeric data are also collected), (4) there is a focus on finding participants' meanings for the program in that context, and data are analyzed inductively. The interpretive evaluator is seeking participants' reactions to and evaluations of a program and reasons for those evaluative perspectives. Qualitative evaluators use the interpretive inquiry mode.

There are multiple levels of interpretation of social reality. The first level is the interpretation of the lived experience by all those who live it. The second level of interpretation is a result of the pieces of the lived experience that remain in the memory of a participant. The third level of interpretation is what is selected out of those memories to share with the evaluator at any given point of data collection. Each participants' interpretations are filtered through different lenses which are constructed of all previous knowledge, experiences, and beliefs of that individual as a result of being a part of a particular culture at some given point in time. The most important caution given to qualitative or ethnographic evaluator is, therefore, that "There may be a correspondence between a life as lived, a life as experienced, and a life as told, but the anthropologist should never assume the correspondence nor fail to make the distinction" (Bruner, 1984).

Figure 6-2 illustrates the social "reality" that is constructed by individuals who share experiences.

Geertz (1973) suggests that all ethnography involves only second and third order interpretation, claiming that only "natives" can make first order interpretations of their culture. While the "natives" in a context, or program participants, can make first order interpretations, the participant observer functioning as evaluator tries to approximate that level of interpretation by assuming a role in the social organization so that s/he experiences the program first-hand along with participants from the site. The evaluator brings her/his own subjectivity to the situation, and from that subjectivity constructs meaning about the program. By interacting with participants, they construct meaning together. The interpretive evaluator builds a second order interpretation through careful systematic observations, ongoing conversational and informal structured interviews, and ongoing participant documentation based on a commitment to the importance of revealing what is "actually happening" in the program and a total trust of the evaluator with their data. Longitudinal designs where the evaluator interacts closely with participants

allows the evaluator to understand the circumstances in which participants would tell, share, or reveal something other than what they understand to be the "real" lived experience in the program. When this "deception for survival" (Brown, 1991) occurs in the evaluation context, understanding *why* this is happening is critical to accurate interpretation of *what* is happening. Usually an understanding of the larger cultural context and power differences provides the insights for interpreting the difference between lived experiences and told experiences.

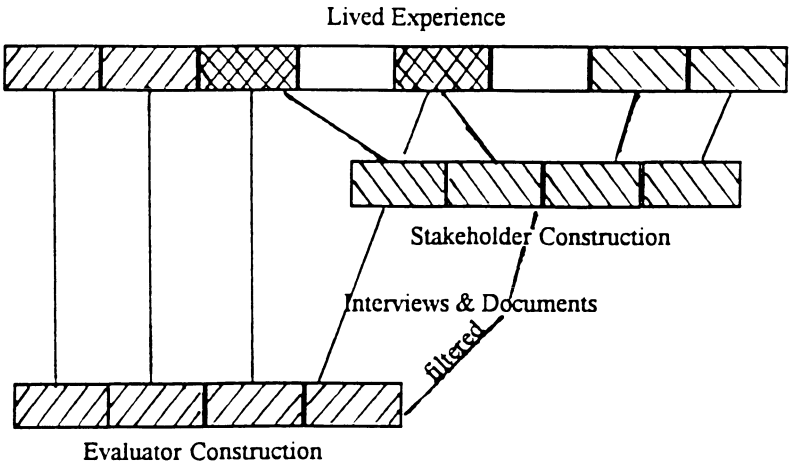


Figure 6-2 Individual Constructions of Social Reality

Note in Figure 6-2 that only parts of the lived experience are remembered by the stakeholders and used in their retrospective construction of what the experience was like. The same process occurs with the evaluator who functions as participant observer. The evaluator's construction is further impacted by the information collected through interviews with stakeholders and documents about a program. Interview data are filtered further through the types of questions the evaluator poses, the selection process by the stakeholder of what to share with the evaluator, and the evaluator selection of data to include as examples in the final report. Note also that parts of the lived experience have been totally forgotten by all participants.

An ethnographic evaluator functions as a participant observer at the evaluation site and simply "lives"with the participants as much as possible to both experience and interpret the innovation or program and systematically observe the effects of it in context as a basis of a cultural analysis. A qualitative evaluator uses the tools (participant observation, interviewing, and document analysis) of ethnography but does not necessarily stay on site continuously nor conduct a cultural analysis of the program (Fetterman, 1984).

QUALITATIVE DESIGN ISSUES

You might ask yourself, "How do I begin to create a good qualitative evaluation design?" The immediate answer to this question is to begin by being open to finding out *how* a program is affecting all persons involved and *why* it affects different levels of persons differently. These two broad goals for the evaluation can guide all data collection and analysis as you revisit them periodically. While you begin holistically, the outcomes of your inquiry will dictate the narrowing of the focus based on participants' perspectives of what is important rather than you as the evaluator defining the focus prior to entering the site. Structured flexibility and openness are key to a good qualitative evaluation design.

Unit of Analysis Problems

In designing qualitative evaluations, the first thing to consider is identification of the unit(s) of analysis. Patton (1990, p. 168) asserts that "The key issue in selecting and making decisions about the appropriate unit of analysis is to decide what it is you want to be able to say something about at the end of the study." Determine what information policy makers need about a program and from whom they need it. Following this, a careful examination of the social organization will lead to identification of the sampling process within the selected site(s) which would address the unit of analysis.

As can be seen from the example in Exhibit 6-1, many evaluation decisions are based on the determination of the unit of analysis. The sampling process, the sample, data collection methods, and the number of evaluators needed on the team are all affected by the identification of the unit of analysis. The evaluator, not the stakeholders, will determine the unit of analysis, but it cannot be done without gaining a clear understanding of what the stakeholders want from the evaluation.

It is important to note that the evaluation can have more than one unit of analysis. The evaluator may want to be able to say something about the overall effects of a program as well as particular effects in subunits of the organization. The two are not mutually exclusive.

Triangulation

In designing naturalistic evaluation, the strategy of triangulation is important. Triangulation is defined (Denzin, 1970, p. 297) as the "combination of methodologies in the study of the same phenomena." Denzin (1970) identifies four types of triangulation: (1) investigator (multiple evaluators investigating the same program); (2) data source (use of as many data sources as possible to understand events being analyzed); (3) data collection methods

Exhibit 6-1. Unit of Analysis--An Educational Conference Evaluation

I was asked by members of the conference committee to design and conduct an evaluation of the lived experience of participants at the 1993 American Educational Research Association Annual Meeting (AERA). This is the annual meeting of a large organization of educational researchers and evaluators with diverse content and research interests.

Prior to making any design decisions, I asked the committee chairperson why they wanted the study done, how they would use the data, and whether they wanted data from persons in all levels and divisions of the organization. I learned that what it was they wanted me to be able to say something about when I finished was what different people's experiences of the conference were on a daily basis so that they could consider making changes which would enhance the experience for all participants in future years.

Based on that understanding, I knew that I needed an ethnographic evaluation design to be able to hear different persons' feelings and reflections as well as observe their behaviors and interactions of the lived experience within the larger culture of the organization on a daily basis. I made a decision to select 5 interviewers and target about 15 participants to obtain in-depth information and thick description (the *meanings* that different events at the conference had for them) rather than to use a larger sample and different data collection techniques for increased breadth of understanding. I also carefully sampled persons to get diversity in research approach (qualitative or quantitative), geographic location, gender, academic statue, and years of membership in AERA so that I would be able to say something about the lived experiences of different levels of persons in the organization.

(within-method and across-method); and (4) theoretical (approaching data with multiple theoretical perspectives and hypotheses). Multiple triangulation is the use of all forms of triangulation. The two most frequently used types of triangulation in evaluation are data source and data collection methods. Data source triangulation includes data collection from different levels of persons, different times, and different places at the site. Data collection methods triangulation in evaluation includes the use of different forms of the same approach as well as different techniques of data collection such as interviews, observations, open-ended questionnaires, surveys, and so on. Denzin (1970) asserts that use of multiple methods of data collection reduces

threats to validity in that weaknesses in one method are offset by strength of another. Multisite investigations are another way evaluators can triangulate the design.

Evaluators believe that triangulation will result in corroborative data across sources, methods, or sites. Triangulation is commonly perceived as a strategy for enhancing validity of research findings. Researchers (Caracelli & Greene, 1993) assert that triangulation seeks convergence, corroboration, and correspondence of results across different methods. Miles and Huberman (1984, p. 234) assert that "triangulation is supposed to support a finding by showing that independent measures of it agree with or, at least, don't contradict it." Mathison (1988, p. 13) notes that historically it is seen as "a strategy that will aid in the elimination of bias and allow the dismissal of plausible rival explanations such that a truthful proposition about some social phenomenon can be made." Mathison (1988, p. 17) argues, "More realistically, we end up with data that occasionally converge, but frequently are inconsistent and even contradictory." This understanding of the result of triangulation places the burden on the evaluator of collecting data which explain *why* data are different or contradictory from different data sources about the same social phenomenon. When evaluators use across-method data collection triangulation as described in Exhibit 6-2, it is important to note comments made by participants about the different data collection methods which might provide insights for valid interpretation of data and the ability of the evaluators to explain differences or conflicts in data from the same data sources and contexts.

Mixed-Method Designs

Mixed-method designs have been defined as those that include at least one quantitative method and one qualitative method where neither type of method is inherently linked to a particular inquiry paradigm or philosophy (Caracelli & Greene, 1993). A mixed-method design using investigator triangulation, where the evaluation team consists of both qualitative and quantitative evaluators committed to their inquiry paradigm and philosophy, is a particularly strong design, however. Evaluators bring extensive training, expertise, and experience in their particular paradigm and data collection approach to inform different aspects of the evaluation. This approach addresses concerns raised by Guba and Lincoln (1988) that internal consistency of each paradigm would be violated by mixing different inquiry approaches. It can be argued (Brown, 1992) that different and important understandings can emerge by triangulating qualitative and quantitative evaluation methods using investigators who are strong in each approach.

Exhibit 6-2. Triangulation: Generating Understanding from Data Obtained from Different Methods

As a qualitative evaluator, I worked on an evaluation team with a quantitative evaluator to investigate the impact of a climate improvement program in an elementary school. At the end of the second year of the program, we asked all teachers in the school to meet in the media center to provide information on two instruments. The quantitative evaluator administered a standardized Likert-scale response survey. I administered an open-ended questionnaire with non-leading questions I had written to allow teachers to write about the meaning and impact of the climate improvement program on them, their students, and the school. Questions on the open-ended questionnaire included things such as the following: 1) Describe ways the climate improvement program has impacted you and your students; 2) What were the most positive aspects of the project for you personally and professionally? 3) What were the most negative aspects of the project for you personally and professionally? We got different results from the same group of people about the same phenomenon using across-method triangulation. Comments made by some of the teachers as they left the media center provided the "why" for us. Each of them that reflected on the process indicated that the responses they wrote for the open-ended questionnaire "really told it like it is" and the responses on the survey did not because it "didn't ask the right questions" and "didn't provide enough responses to select from."

In mixed-method designs such as the one described in Exhibit 6-3, data from each paradigm are analyzed independently. Quantitative data are numerical and are analyzed statistically. Qualitative data are generally narrative (but sometimes numerical for descriptive purposes) and are analyzed using a strategy like constant comparative analysis or phenomenological analysis which allow for emerging categories and relations among categories to be generated from participant data. There is no need to use strategies to artificially numerically code and transform rich narrative data into numerical form for analysis that would be comparable to quantitative data analysis. That undermines the basic reasons for conducting rigorous in-depth qualitative evaluation. Analyzed and interpreted data from each approach are examined and compared by all evaluators to generate a broader understanding of the impact of the program being evaluated.

Exhibit 6-3. Mixed-Method Evaluation of an Innovative Preschool Program for At-risk Children

I was the qualitative evaluator for an innovative preschool program for at-risk children where triangulation of investigators was a planned part of the evaluation design. Statistical comparisons of pre- and post-implementation data from a variety of standardized instruments clearly demonstrated the positive impact of the program on understandings and skills development of the preschool children. From a different paradigm, the qualitative paradigm, findings from observations, interviews, and videotape analysis revealed the positive impact of the program on the teaching--learning situation, parent involvement, student--parent interactions, community involvement, and on-location learning at sites other than the classroom. Evaluation reflecting philosophical underpinnings and data collection methods of either paradigm alone could not have resulted in the rich diversity of understanding that resulted from the mixed-method investigator triangulation design.

Ethical Considerations

All evaluators must consider the ethical implications of their work. Qualitative evaluators must exercise extreme caution in detailing every aspect of the social system and social hierarchy of an organization prior to gaining entry in order to avoid as many ethical blunders as possible. Qualitative evaluators, and particularly ethnographic evaluators, will spend a great deal of time in the site interacting with and observing persons. Every conversation for the ethnographic evaluator is data collection. Because the evaluator is an outsider, persons at different levels begin to reveal aspects of the life and culture at the site which may bias the evaluator in data collection and interpretation.

More important, however, the information itself may present ethical dilemmas where the evaluator has to decide whether to reveal information to stakeholders relative to anticipated injury, social loss to participants who provide the information, or other potential changes in the social structure or organization that would not occur if s/he kept the information confidential. If a participant asks the evaluator not to use information and the evaluator agrees, then whatever the participant shares with the evaluator, even if it is critical for an accurate evaluation, must be excluded from fieldnotes and the final report. If the evaluator indicates that participant information will remain confidential, then fieldnotes, interview transcripts, and the final evaluation report must reflect that promise. Sometimes (see Exhibit 6-4) evaluators

inadvertently break the promise of anonymity by the way data are reported. This can be avoided by carefully masking sources and using member checks before submitting reports.

Exhibit 6-4. Ethical Decisions--Confidentiality Agreement Broken

In conducting individual interviews for the second year of an ongoing project, I began the interview with the first site manager in my usual manner stating, "I will be interviewing the manager of each site, but the information you share about the impact of the project at your site will remain totally confidential in that I will not use your name in my notes and neither you nor your site will be recognizable in the final evaluation report." The first interviewee sat quietly, answered questions politely, and left. I was puzzled in that the rapport I usually quickly establish in interview situations never materialized.

The second manager, after hearing my promise, leaned back in his chair laughing loudly and said, "Yea"? Well that's just what the last interviewer said last year and when the boss got the report, we were labeled Site #1, Site #2, Site #3, and so on, and he and everyone else knew by the clear descriptions exactly who had said what and what was going on at each site. I'm not going to tell you anything that isn't common knowledge around here and none of the rest are either."

I had a very difficult time convincing those managers that I would not write the evaluation report in the same way. I was finally able to gain the trust of all of the participants, but at great cost in time in each interview as I had to explain what I had learned about the last year's evaluation and how I would function differently. Then I assured each manager that I would write the evaluation report, send a copy to each of them to edit as they felt it needed to be to protect individual identities, and only after that, send the final evaluation report to their boss. They agreed to that process. I received no suggestions for revisions of the final evaluation, but only approvals to send it forward.

Being in the natural setting and "living the program" with the participants, the qualitative evaluator will be privy to conversations about controversial situations in the site, many of which do not have anything to do with the project being evaluated. The best advice in such situations is to listen and respond normally as one would in any conversation, but do not act on anything shared in informal conversational situations unless it relates to the project. Personal information about participants or other non-project events should not be recorded or repeated by the evaluator. The evaluator's job does

not include righting all wrongs or instituting changes that appear necessary. The goal of qualitative evaluation is simply to determine the impact of the program on participants at the site by understanding the meanings it has to them; why it has those meanings; and how it affects behavior, actions, and interactions in that context.

Participants should be interviewed in the least threatening circumstances and locations at the site. Teachers, for example, at one school forewarned the evaluator that they would casually walk away if a particular female colleague came into the area because they did not want her to think they were "telling [the evaluator] what is really happening in the project" and make the project director and principal angry with them. Recorded data (fieldnotes or tapes) should be destroyed as soon as data have been used to generate the evaluation report. Because of the personal nature of data collection in participant observation evaluation, participants feel particularly betrayed when they reap negative and unexpected impacts of an evaluation effort after providing the evaluator with ongoing information for an extended time.

Participants at a site should be given an opportunity to hear about the evaluation design and their roles in it initially and then be provided with an opportunity to give written consent to participate. Qualitative evaluators should not share any participant's perspectives with others at the site. Fieldnotes and interview transcripts should remain confidential. Analyzed and interpreted data, on the other hand, can and should be shared with participants as member checks (Guba & Lincoln, 1989) to determine whether the final understandings make sense to them and make sure that the evaluator "got it right."

DATA COLLECTION METHODS

A variety of data collection methods are associated with qualitative evaluation designs. These include participant observation, observation, interviews, open-ended questionnaires, and document collection. Characteristics shared by those methods are that they are unobtrusive, inductive, labor and time intensive, and generally result in narrative data. The goal of using these types of data collection methods is to generate data from the perspectives of the participants in programs being evaluated.

Participant Observation

Participant observation is the most common data collection method in qualitative evaluation. The role assumed by the evaluator falls on a continuum (Gold, 1958) from total researcher to total participant. Most qualitative evaluators assume a role more toward the total researcher end of the continuum. This means that the evaluator does not assume an active role within the social group in the ongoing program. Taking the role of total

evaluator requires a period of negotiating a trust relationship with participants at the program site. Gaining entry, establishing a trust relationship, negotiating reciprocity, finding an unobtrusive niche, and meeting persons at different levels in the program are important steps for successful qualitative evaluation.

A participant observer participates in and observes as much of the social interaction relative to the program in the natural setting as possible. Observation is unstructured, holistic, and constant in the setting. Participant actions, interactions, and responses to programs guide observations. A "verbal photograph" (what is happening here) of the social actions and interaction can be recorded in fieldnotes, audio tapes, or videotapes. The advantage of fieldnotes recorded by hand or computer is that the lived experience does not have to be "lived again" as it does when the evaluator watches or listens to tapes of events. Each entry in a field notebook should be contextualized, have the date and time, and include a brief description of the event and persons involved in it. The disadvantage of recording events in fieldnotes is that interactions are missed when the evaluator is writing. An advantage of videotapes of events is that the event can be observed multiple times to obtain different types of information (verbal interactions, nonverbal interactions, and context clues). In addition, participants can watch segments of videotapes with the evaluator and provide interpretations of interactions from the insider's perspective.

Fieldnotes often include observer comments and reactions to things observed. Fieldnotes are for the exclusive use of the evaluator and are not shared with participants. Daily ongoing systematic analysis of fieldnotes provides a guide for further observation and interviewing needs throughout the evaluation. The ongoing analysis allows the evaluator to identify phenomena and events that are clearly understood, missing, and incomplete. Changes in observation strategies are based on these understandings.

Interviewing

The purpose of interviewing in qualitative evaluation is to find out the meaning of the program to participants. Interview formats can vary on a continuum from highly structured evaluator directed question and response guides to informal conversations whose focus and direction are directed by participants. Participant observation always includes conversational interviews because of the level of interaction between the evaluator and participants. The selection of interview format is determined by the type of information that is desired, the amount of time available to the evaluator to collect data, and the level of comparability of findings that is desired. The less structured interview formats require more time and are less comparable, but they allow participants to discuss issues and concerns that are of utmost importance to

them. Evaluators must carefully consider the tradeoffs when selecting interview formats.

Interviews can be with individuals or groups of participants. Each approach has advantages, and both can be included in a single evaluation design. Focus group interviewing is a form of qualitative data collection in which the evaluator functions as discussion facilitator for a small group of participants and relies on interaction within the group to provide insights about topics proposed by the evaluator. Krueger (1988) argues that focus group interviews can provide vital information on the impact of programs on participants. Morgan (1988) explores the advantages and disadvantages of focus group interviews.

Advantages of focus groups are:

- They are relatively easy to conduct.
- They require less time than multiple individual interviews.
- They provide the opportunity to collect data from group interaction.
- They provide an opportunity for group discussion opinion formation of researcher-generated topics.

Weaknesses of focus group interviews are:

- They are not conducted in the naturalistic setting.
- It is impossible to discern individuals' perspectives.
- The degree to which the presence of the evaluator and other participants affects responses of any individual cannot be determined.
- Comparison of data across focus groups is difficult because group interaction determines the direction or focus of discussion.
- Fewer questions can be asked because more interviewees are involved.

In addition to the format, the wording and sequencing of questions affect interviewee responses. Interview questions in qualitative evaluation should be singular, clearly worded, nonleading, and open-ended (Patton, 1990, pp. 277-368).

One of the most important goals of interviewing in evaluation is to find out *why* different individuals or different levels of persons construct different meaning about a program. In other words, in order to be able to say something about the reasons for different or conflicting findings about a program, data need to be generated that account for those differences in perspective or meaning. One of the most *ineffective* question formats is to ask "Why?" after other questions. Role playing and simulation questions or

questions asking for descriptions or examples are superior techniques for finding out why participants function the way they do or have the perspectives they share about a program.

Interviewing as many participants as possible in different contexts and across the times throughout the program will provide an understanding of evolving perspectives. Key informants are special people in the social context with whom the evaluator spends more time than with other participants. The key informant provides insights and insider interpretations that the evaluator may not be able to access as an outsider to the group and situation. Key informants are selected because they may be particularly well informed about the program, may be available to the evaluator, may have played a key role in helping the evaluator gain access to the site, or other characteristics that make her/him special and different from other participants. Selection of a key informant who is peripheral in the social structure or who is viewed negatively by some or all of the program participants can be detrimental to the evaluation process.

Questionnaires

Carefully constructed, open-ended questionnaires serve the same purpose as interviews in that they help the evaluator can "get inside the head" of participants to find out their perspectives of the program (see Exhibit 6-5.) Questionnaires require less time to administer than interviews so comparable data can be collected from many more participants. (See Chapter 7 for additional discussions of questionnaire and opinionnaire design and use.) Cautions in composing questions for an open-ended questionnaire include avoiding leading questions, writing clear questions, providing enough space for responses, and carefully arranging questions so that a response to one will not affect the response to subsequent questions. Questionnaires should be as concise as possible to ensure the greatest number and highest quality of responses. Pose only those questions whose responses will allow you to be able to say something about the phenomena you want to say something about as defined by your selected unit of analysis.

Some issues need to be addressed when using open-ended questionnaires.

1. The first issue is whether questionnaires should be administered in the program context or mailed to participants. Mailed questionnaires usually result in a relatively low response. On the other hand, there may be time or human presence constraints in the context that could affect the way participants respond to a questionnaire.

2. The second issue is whether questionnaires should be confidential or anonymous. If comparison of data by individual participant with data from other methods is essential, then participants would be required to supply their name or some different form of identification which would allow them to remain anonymous. When participants are required to put their names on questionnaires, the content of responses may be affected if participants feel that they or their job security is threatened by persons at different levels within the organization if they respond honestly.

3. The third issue is whether questionnaire data will be analyzed by person across questions or by question across persons. If data are compared by question across persons, the identity of individuals is generally easily masked. When data are analyzed by individual or subgroups of individuals, identities of respondents can often be identified by persons within the context.

Decisions about each of these issues must be made after consideration of each evaluation situation and context to determine which approaches will result in the most valid information and the highest response rate.

DATA ANALYSIS

Analysis of qualitative data is an ongoing cyclical process that consists of synthesizing information across data sources and data collection methods. The analysis process is generative in that hypotheses are not tested but generated from participant data. There are different approaches to qualitative data analysis, and each addresses different evaluator needs relative to the types of data collected and evaluation questions asked.

Most qualitative data are in narrative form, but often qualitative evaluators have numerical data that are presented as descriptive statistics. Examples of numerical data in qualitative evaluations might be proportions of different categories of persons participating in a program, proportions of time participants are engaged in specific activities on a daily basis, group sizes, and other such frequency counts. Stakeholders want to know how representative claims in evaluation reports are, but this is not an argument that supports the transformation of rich narrative qualitative data into simple frequency counts.

Qualitative Data Analysis Strategies

Qualitative data analysis is inductive. Evaluators engaged in interpretive inquiry generally begin with rich data from a variety of data sources and methods to determine "what is in the data" rather than beginning with a theory or hypothesis to test. Identification of categories or themes in the data is followed by establishing relations among categories and then seeking further

evidence to support categories and relationships in the field setting. In some cases, qualitative evaluators will, however, begin with a theory to test and use analytic induction to analyze data.

Exhibit 6-5. Example of an Open-ended Questionnaire

TEACHER QUESTIONNAIRE ON
STAFF DEVELOPMENT EXPERIENCE

Instructions: Please take the time to read carefully and respond honestly to each question. We sincerely want to know about your experience in the Project 2061 staff development workshop and value your suggestions for improvements. Thank you.

1. Describe any ways that you think differently about the way children learn or the way children learn science as a result of experiences or articles given to you in this staff development.
2. Describe ways that you will teach science differently next year because of things you read, learned, or experienced during this Project 2061 staff development.
3. How would you describe what science is to a group of students in your school?
4. What was the most positive aspect of the Project 2061 science workshop for you? What about it was positive for you?
5. What was the most negative aspect of the Project 2061 science workshop for you? What about it was negative for you?
6. If you were going to conduct a similar Project 2061 science workshop for teachers, what changes would make to improve teachers' experiences?

Phenomenological Analysis. The goal of phenomenological analysis (Hycner, 1985) of narrative data, particularly interview data, is to understand the phenomenon or program in its own right and not from the perspective of the researcher. The evaluator brackets or suspends her/his own meanings and interpretations as much as possible and allows meaning to emerge from the data (interviews, questionnaires, open-ended surveys, and documents) that have been generated by participants. The evaluator as analyst delineates units of meaning in participant data relevant to the evaluation questions and established relations among units generated in different data sources and data

collection methods. The clusters of units of meaning are themes that address the evaluation questions.

Content Analysis. Content analysis is a well known method for analyzing documents and written communication. Documents are often produced at a program site without guidance from the evaluator and for different reasons other than evaluation. Documents, however, can be a good source of information about program implementation and interpretation by participants. Content analysis is defined by Holsti (1969, p. 14) as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages." Guba and Lincoln (1981) make a case for qualitative content analysis, explaining that frequency counts are not necessarily associated with the importance of assertions in documents. Qualitative content analysis includes generation of categories from the data which are relevant to the purposes of the evaluation. Evaluator-generated rules for categorization, demonstration of representativeness of categories, relations among categories, and definitions of categories from participant perspectives are important outcomes of content analysis.

Analytic Induction. Qualitative evaluators beginning with a theory to test about a program in a particular setting would probably analyze data using analytic induction. Rather than starting with holistic observation and interviewing, cases would be selected using specific criteria in the setting to test the theory. As data saturation (finding no new or different cases) occurs, the evaluator stops collecting data and presents the evidence to support the theory.

Constant Comparative Analysis. Constant comparative analysis (Glaser & Strauss, 1967; Strauss, 1987) is an approach to analysis that results in grounded theory. Analysis is ongoing throughout data collection. As data are displayed and reduced into categories of meaning and relations among categories, hypotheses are proposed to account for social meaning and interaction that emerge in the data. Through theoretical sampling, the evaluator is guided in data to collect, data sources to approach, and data collection methods to use. A process of writing theoretical and methodological memos, or notes about ongoing insights, informs the interrelated data collection and analysis process.

Presentation of Evaluation Data

The key to presenting qualitative or ethnographic evaluation findings effectively is the idea of contextualization or interpretation of behavior, interactions, and constructed meanings within the context or culture in which they were collected. Historically, qualitative evaluators have not been concerned about generalizability of findings, but rather are concerned about presenting an accurate and holistic description of the immediate and larger

contexts in which findings are generated. The strength of qualitative evaluation is the generation of in-depth understanding of the process, social interaction, participants' perspectives, and meanings constructed by participants in programs being evaluated.

Multisite qualitative evaluation designs are a way to enhance generalizability of findings (Herriott & Firestone, 1983). Multisite evaluations of the same program in dissimilar contexts provides greater generalizability than single site designs or multisite designs where contexts are similar (Kennedy, 1979; Sinacore & Turpin, 1991). Firestone (1993) discusses issues related to generalizability of qualitative research and presents an approach using Boolean algebra to compare large numbers of cases systematically. Generalizability, however, is not the goal of qualitative research and evaluation. Generalizability of qualitative findings depends on the degree to which other situations, programs, and participants are similar to those described in the evaluation report. Thus, the primary responsibility for generalizability rests with the evaluator as the program, context, social structure, and participants are clearly, systematically, and holistically described relative to program related findings.

Qualitative evaluation findings may be shared throughout an evaluation to guide changes in programs, presented in a final report at the end of the piloting of a program or in both ways. In a multiphase or multiyear program in which an evaluation report must be present at the end of each segment, or when the goal is formative evaluation, it is important to remain cognizant of the possible negative impact of interim reports and responses to those reports by stakeholders on participants at all levels of the organization.

Many stakeholders in programs have constructed misunderstandings about the nature of naturalistic and interpretive inquiry without having any formal training in it. Although they know neither the types of questions which are best answered using naturalistic evaluation nor the methodology of interpretive inquiry, they form very strong beliefs about the nature of findings and value of this approach to knowing. It is critical to keep this in mind for the duration of a qualitative evaluation, knowing that there are many ways in which the evaluation process or stakeholders' responses to it might negatively affect one or more participants (see Exhibit 6-6.)

The specific and important role that interpretive inquiry plays in evaluation was considered in this chapter. Interpretive inquiry in evaluation can be in the form of ethnographic evaluation, qualitative evaluation, or naturalistic evaluation. These approaches to evaluation are conducted in the natural setting with an emphasis on understanding the program impact from the perspectives of all levels of persons in the organization. When the goal of evaluation is to find out the impact of a program on social interaction, the

Exhibit 6-6. Demoralized Participants--A High School Vocational Education Project

At the end of the first year of a four-year high school vocational education project to enhance basic skills of vocational students by having vocational and academic teachers plan together, I was required to present a qualitative evaluation report to the project director, school principal, county superintendent, participating teachers, and project validation onsite team members. My report was based on a vast amount of very rich data from teachers and students and indicated a number of very positive outcomes of the project for teachers and students in the school. The quantitative evaluator indicated that there were not sufficient test data to make an evaluation that early. After both reports were presented, the superintendent asserted that the qualitative data could be "manipulated to show anything," and the chairperson of the onsite team proclaimed that while the qualitative data were interesting, that the "jury would remain out on the effect of the project until the hard data [test scores] were in."

Teachers had cooperated totally with self-reflection procedures required in the qualitative design. They kept weekly logs of events and experiences, they agreed to multiple interviews with me, they allowed me to interview students and administer open-ended questionnaires to students, and they provided me with all documentation relative to the project. It had been time consuming, but they had felt it was valuable as they began to use their own data to make decisions about needed changes to make the project more effective.

At the end of the reporting meeting I left with a large group of project teachers. All teacher comments went something like, "Forget it! We won't do any more logs or anything. We thought we were doing good things, but they don't value what we've done at all. It simply goes back to test scores again--are they significantly better or not! They don't even care if kids are getting a better education and we are doing a better job. They don't care what we think or what we do." Needless to say, qualitative data collection the last three years of the project was very difficult. Teachers understood the need for the process for themselves and the good of the project, but they were very discouraged by the responses of the administrators and outside validation team members to the positive information that had been presented.

construction of social meaning, individuals in a social setting, or to find out what is actually being done at a site that is piloting or implementing a program, an interpretive or qualitative paradigm provides the epistemological and methodological basis for providing understanding.

Evaluators trained in the use of qualitative data collection and analysis methods are the best persons to design and conduct qualitative evaluation. Schooling in the philosophical and theoretical underpinnings of qualitative inquiry allows the evaluator to make informed decisions throughout the evaluation concerning ethical issues in the field, the role of subjectivity, evaluator roles, and use and interpretation of multiple perspectives and triangulation. Qualitative and ethnographic evaluation is fieldwork based and includes data collection methods of observation, interviewing, and document collection. Generally, questions and hypotheses are generated from data collected about the program or phenomenon being evaluated. In some cases, however, evaluation questions and data categories are established prior to fieldwork and data are collected specifically to test those questions.

Qualitative evaluation data are analyzed in an ongoing and interactive manner with data collection so that findings can guide hypothesis generation and testing. The data collection method and evaluation questions serve as a guide to selection of an appropriate data analysis strategy. Systematic and rigorous qualitative data collection with ongoing data analysis and interpretation can provide formative and summative participant-based information for enhanced stakeholders' decision making.

COGITATIONS

1. Which types of evaluation goals would best be addressed by using qualitative evaluation? Ethnographic evaluation?
2. What are the characteristics of the qualitative paradigm that inform qualitative and ethnographic evaluation?
3. How does the concept of multiple perspectives of reality support Mathison's assertion that triangulation often results in convergence, inconsistency, and contradiction?
4. What are the strengths and weaknesses of different data collection methods in qualitative evaluation, and what are the best ways to offset those weaknesses?
5. What issues must be addressed when designing qualitative evaluation?
6. What are different approaches to analyzing qualitative evaluation data? What are the advantages and disadvantages of each approach?
7. What meaning does generalizability have in qualitative evaluation?

SUGGESTED READINGS

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

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

Krueger, R.A. (1988). *Focus groups: A practical guide for applied research*. Newbury Park, CA: Sage.

Patton, M.Q. (1990). *Qualitative evaluation and research methods*. (2nd ed.) Newbury Park, CA: Sage.

Strauss, A.L. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.

7

CREATING AND SELECTING INSTRUMENTATION

Individuals charged with the development or selection of measuring instruments like to chide statisticians and evaluators, saying that they would be nothing if it were not for the data provided by the measurer. We are all indebted to someone. But it is true that we cannot really judge the impact or effectiveness of a program or project without some data. Data are broadly defined as anything ranging from verbal descriptions of playground behavior to scores on a standardized achievement test. Here we are considering data as coming from a variety of sources, sometimes using paper and pencil or behavioral/performance measures or perhaps using the human observer as an instrument. But no matter what the source, there are some standard criteria that all data should meet.

All that can be done in this single brief chapter is to overview some measurement issues, concepts, and resources that evaluators and project directors might consider when faced with the task of documenting the impact of a program. Check out the chapter references and end-of-chapter Suggested Readings on how to do it.

CHARACTERISTICS OF HIGH--QUALITY DATA

Because of their intimate association we will discuss data and measurements as one. Measurements yield data. If the measures are good, the data will be good. Data become "informative" only after they get communicated. But what constitutes a "good" measurement? Following are six desirable characteristics to be sought in a measuring instrument whether it is to be used to collect data on achievement, aptitude, or attitudes.

1. Relevance. Relevance is the correspondence between the data and the intent or objective in gathering it. It might be the match between a test item and a behavioral objective, or the match between a series of planned observations and projected teacher-student or student-student interactions. In the measurement sense, relevance is the primary contributor to validity or the degree to which the measurement is a true and accurate reflection of the variable of interest.

2. Balance. Any measurement needs a framework or plan for its development. The extent to which the developed measure corresponds to the ideal measure reflects balance. In developing an achievement test, for example, a blueprint in the form of a table of specifications is created whereby content and outcomes are summarized in a 2 x 2 table. Entries in the cells reflect the proportion of instructional time devoted

to those objectives. The test is then built according to those proportions, resulting in a balanced measure. Any multidimensional instrument can benefit from the application of this concept.

3. Efficiency. Basically we are looking for the greatest number of meaningful responses per unit of time. Gathering data costs time and money, so we want to conserve our resources. A balance between time available to collect the data, cost, requirements for scoring and summarization, and relevance should be sought.

4. Objectivity. Do experts agree on the interpretation of the data? With regard to paper-and-pencil exams, for example, do different scorers or raters come up with the same results? If behavioral observations are involved, will different observers "see" the same thing? Along the same line, given a series of extensive field notes from participant observers, will different ethnographers evolve the same interpretation? Objectivity, then, is a characteristic of the "scoring" or the assignment of meaning to the data, rather than being a description of the method of data collection.

5. Reliability. Reliability is a complex characteristic but generally involves the idea of consistency of measurement. Consistency of measurement might be judged in terms of time, items, scorers, examinees, examiners, or accuracy of classifications. It also has important applications when dealing with qualitative data. These might relate to such activities as interobserver agreement and consistency in drawing inferences from observational data or written descriptions.

6. Fairness. The criterion of fairness relates to a wide range of data characteristics ranging from freedom from bias (gender, ethnic, or racial) to the administration of a test in a manner that allows all students, for example, an equal chance to demonstrate their knowledges or skill. Everybody should play by the same rules, and the rules should be the same for everybody.

These criteria, then, represent characteristics that need to be kept in mind when developing or selecting measuring devices. The six categories are obviously not mutually exclusive: for example, an irrelevant item would not be fair. The criteria represent targets for our psychometric arrows.

What kinds of measures are available where we can use these criteria?

TYPES OF MEASURES

A traditional list of measurement devices would include a tremendous variety of approaches. The list might contain descriptions of multiple-choice

test items, rating scales, checklists, observation schedules, physical tasks, projection devices, and manipulative performances. The use of standardized paper-and-pencil tests of cognitive performance, self-report affective measures, and observation methods make up the majority of approaches available to the evaluator. A nontraditional scheme of classifying tests has been developed by Campbell (1957) who classifies measures according to these three dichotomies:

Voluntary	vs.	Objective
Respondent understands that any response is acceptable. He or she is not aware of any external classification scheme or criterion for evaluating responses. Self-description is encouraged.		There is an implicit frame of reference--correct or incorrect. The concepts of accuracy and error are implied.
Indirect	**vs.**	**Direct**
Interpretation of responses rests on either involving categories established after the fact or those different from those that might be expected by the respondent.		Respondent and data collectors are in general agreement regarding the purpose of gathering data.
Free-Response	**vs.**	**Structured**
Responses are from open-ended questions in unstructured format. Individual imposes own organization upon task.		Specific kinds of responses are required relative. Form, activity, and so on are regulated.

If combinations of these categories are made using one from each of the three categories,

Category 1	1	=	Voluntary
	2	=	Objective
Category 2	3	=	Indirect
	4	=	Direct
Category 3	5	=	Free-response
	6	=	Structured,

the following codes result. Following are some illustrative measures.

Code	Illustration
1-3-5	Open-ended questions requiring content analyses: *What do you admire most in people? What is the most embarrassing thing you can think of?* Some traditional projective devices such as the Thematic Apperception Test would also belong here.
1-3-6	Certain kinds of preference measures such as art preference or judgment. Q-sorts, and structured projective tests.
1-4-5	Autobiographies and sentence completion measures. (Respondents are aware that they are revealing their attitudes.) Structured survey forms as used in public opinion polls, questionnaires, and opinionnaires. Also interest inventories such as the Kuder or Strong.
2-3-5	Certain kinds of projective devices fall in this category; for example, making up stories on the basis of limited verbal cues or pictorial stimuli (e.g., photographs).
2-4-5	The traditional essay exam would be a good example.
2-4-6	Again, traditional achievement tests, aptitude measures, and intelligence tests would fit this category.
2-3-6	The open-ended task of 2-3-5 could be made structured by providing the response categories for the respondent.

The one dimension that this scheme does not address is the actual collection method: namely, does respondent provide data as in a self-report situation or is a second party involved? For example, a 1-3-5 could be a questionnaire responded to by an individual or an interview by a second party. Analogously a 1-3-6 could be observation of a subject in a structure simulation of a job interview. The point here is not so much to create an elaborate code system, but to cause us to think about the variety of measures that need to be created or identified, and what the characteristics of those measures might be. The system also helps us think about the many possible ways that can be used to collect data. Appendix A contains an extremely valuable collection of possible criterion measures that could be used to evaluate school programs (Metfessel & Michael, 1967). After deciding on the variable of interest, an evaluator might find a way to measure it in that list. On the other hand, just reviewing Appendix A might stimulate ideas about outcome variables that need to be addressed in developing a new program or project.

New programs and projects frequently require the collection of survey data from a variety of stakeholders and audiences. The use of well-constructed opinionnaires can greatly facilitate the efficient collection of data related to goals and objectives, needs, effectiveness of inservice programs, and general perceived program impact.

OPINIONNAIRE AND FREE-RESPONSE METHODS

The questionnaire survey is a frequently used polling method to gather opinion and attitude data. The term *opinionnaire*, as opposed to questionnaire, is actually used more frequently since it suggests an emphasis on feelings rather than facts. The use of a well-constructed opinionnaire tends to systematize the data-gathering process and to help insure that the relevant questions are asked and that all important aspects of the problem are surveyed.

The opinionnaire method, either open-ended or closed (structured), is frequently maligned. But as is often the case, it is the user that should be castigated for improper use, not the method itself. Opinionnaires, if properly constructed and analyzed, can provide very valuable information about cognitive and affective variables. They are, or can be, efficient with regard to time for construction and administration to large or small groups of respondents. They are also relatively inexpensive. They do require carefully crafted questions. The unstructured free-response opinionnaire will require large amounts of time for content analyses of the responses. A great deal of subjectivity may be involved in interpreting responses. Respondents may "wander around" in answering the questions, so evaluators should be prepared to separate the wheat from the chaff. Unfortunately, opinionnaires are often haphazardly constructed, without proper concern paid to the phrasing of questions, or the means of summarizing or analyzing the data. Pilot testing is frequently not conducted.

Six criteria for a "good" opinionnaire are as follows:

- Brevity.
- Inclusion of items of sufficient interest and "face appeal" to attract the attention of the respondent and cause him or her to become involved in the task.
- Provision for eliciting sufficient depth of response in order to avoid superficial replies.
- Wording of questions to be neither too suggestive nor too unstimulating.
- Phrasing of questions in such a way as to allay suspicion about hidden purposes and not to embarrass or threaten the respondent.
- Phrasing of questions so that they are not too narrow in scope, allowing the respondent reasonable latitude in his or her responses.

Opinionnaires are generally of two types: the "closed" or precategorized type and the "open" or free-response type. Rating scales are also frequently associated with the structured opinionnaires. It is recommended that the open-ended form of opinionnaire be adopted for most uses unless a very large number of respondents is involved. The use of such free-response questions allows the evaluator to cover a wide variety of topics in an efficient manner. Analysis of the responses to free-response questions can, however, be quite time consuming and difficult. In preparing opinionnaires, some general cautions should be observed (Payne, 1951):

- Spell out in advance the objectives, purposes, and specifications for the instrument. This task should be undertaken *before* questions are written.
- Try to limit the length of the questionnaire (e.g., 10 questions). If the respondent becomes impatient to finish, he or she is likely not to consider his/her answers carefully.
- Make sure respondents understand the purpose of the opinionnaire and are convinced of the importance of responding completely and candidly.
- If possible, use a sequence of questions. Green (1970) illustrates the advantages of this approach with a series of questions that could be used to stimulate attitudinal responses toward labor unions in a unit of a social studies course.
  a. How have labor-management relations been affected by unions?
  b. How have working conditions been affected by unions?
  c. What means, if any, should be used to control unions?
  d. What effects have unions had on the general economy of the country?
- Make sure respondents are motivated to answer questions thoughtfully.
- Control the administration of the opinionnaire so as to prevent respondents from talking with one another about the questions before answering them.
- Urge respondents to express their own thoughts, not the responses they think the evaluator, teacher, principal or project director wants.
- Be sure the directions are clear, definite, and complete.
- Urge respondents to ask about questions that are unclear to them.
- If possible, try out the opinionnaire with a couple of respondents to identify and clear up ambiguous questions, difficult terms, or unclear meanings.

Content Analysis

An evaluator will ordinarily undertake a content analysis of the responses to opinionnaire questions. Content analysis is a systematic, objective, and sometimes quantitative examination of free-response material. In addition to examining opinionnaire responses, content analyses of textbooks, television broadcasts, essays, records of interpersonal interactions, plays, stories, dramas, newspaper articles, speeches, or propaganda materials may be undertaken.

Several steps are involved in completing a content analysis.

1. *Identify the units for the purpose of recording results.* The specification of units, which requires great care, may be undertaken before beginning the analysis if the analyst knows what to expect or after a sample of the responses has been examined. A unit is usually a single sentence, although any brief phrase that summarizes an idea, concept, feeling, or word will suffice.

2. *Identify the categories into which the units will be placed.* For example, the unit might be a sentence and the category a type of sentence: for example, declarative or interrogative.

3. *Analyze all the content (or a representative sample) relevant to the problem.* A given piece of material could be sampled for a given document (log, diary, etc.), or samples could be taken.

4. *Seek to attain a high degree of objectivity.* The analyst may want to finish an analysis or to put it aside and redo it (or a portion of it) later to check agreement of results. A comparison of the work of two analysts working independently could serve as another check on objectivity.

5. *Quantify the results, if at all possible.* The use of simple summary indices such as frequency counts and percentages can be very helpful.

6. *Include a sufficiently large number of samples to insure reasonable reliability.* The larger the sample of material(s) analyzed, in general, the greater the reliability.

An Illustrative Content Analysis

In an effort to evaluate the impact on an eight-week summer enrichment program for academically and artistically talented students, the author asked several questions such as the following on a participant follow-up opinionnaire:

1. What contribution, if any, did the program make toward your developing a positive attitude toward learning?
2. How suitable were the instructional methods?
3. To what degree did the program influence your desire to attend college?
4. What do you feel were the most beneficial dimensions of the program?

A content analysis of the last question yielded the following results (with a sample of 50 subjects):

		Frequency	Percent
a.	Contact with individuals with both different and similar interests.	34	68%
b.	Freedom for independent and in-depth study.	12	24%
c.	The high quality of teachers.	9	18%
d.	The availability of cultural events, films, speakers, and the like.	8	16%
e.	Freedom to broaden interests.	5	10%

Not only were relevant dimensions of the program identified but a ranking of the importance of these dimensions also became possible. The fact that this information came from the participants themselves helps insure the validity of the responses. If precategorized responses had been used, data might have been biased.

OBSERVATION METHODS

Many program and project outcomes cannot be measured with paper-and-pencil devices. An invaluable alternative is observation. This alternative is, of course, a major tool of the anthropological evaluator. Observation results may be summarized as verbal descriptions or frequencies and percentages of occurrence of emerging sets of categories. The treatment of the "observation" method in the following section focuses on program outcomes less than on documenting project process or implementation.

Many of the myriad possible program outcomes in our schools and classrooms can be assessed with formal paper-and-pencil devices. This is particularly true of learning outcomes that involve knowledge, verbal and

thinking skill development, comprehension, and problem solving. In addition, it is becoming increasingly apparent that many affective outcomes can be assessed with paper-and-pencil measures, providing data useful for project director, evaluator, and teacher. But this is not the whole story. It is difficult at best to assess proficiency in many skill areas and in situations in which personal-social development is emphasized. The best approach to assessing behavioral changes is direct observation of those behaviors. We need to know not only whether a student *knows* what to do, but also whether he or she *can* do it, and finally we would like to know if he or she *will* do it. The assessment of behaviors and outcomes in lifelike and realistic (as opposed to the classroom environment) situations can supply us with some of the most valid data for decision making.

Advantages of Observational Methods

The use of observational methods has been slighted in many program evaluations and school assessment situations, probably because of the difficulty of developing and applying the techniques. There are, however, many advantages to the observation method (Evertson & Green, 1986; Medley, 1982).

- Observational and qualitative evaluation methods allow us to gather data, particularly about social-emotional-personal adjustment, in valid and reliable and precise ways not possible with more traditional methods.
- Observational methods allow us to examine an individual's ability to apply information in lifelike situations.
- Because of the similarity between the testing situations and the setting in which the skills and knowledge are likely to be used, we find that observational measures tend to have higher predictive validity than do many other methods of predicting successful job performance.
- Observational methods are easily adapted to a variety of settings, tasks, and kinds of individuals, at all age and educational levels.
- Observational data can serve as an invaluable supplement to achievement and ability data available from other sources.
- Observation can provide both qualitative and quantitative data.
- Observation methods allow us to document individuals interacting with individuals or the environment.
- The use of data from a variety of sources results in a more reliable overall assessment.

- Observations taking place in natural settings should enhance, for example, their integration into the total instructional program and allow the instructor to use observation as part of the teaching process.
- The use of observational data to record *and* analyze project-related meetings for purposes both of documentation and clarification should not be overlooked.

Disadvantages and Difficulties in Using Observational Methods

Observing individuals is a difficult task. Many factors influence what an observer perceives and how his or her observations are reported. Training and experience are the prime contributors to the development of effective observational skills. There are a number of pitfalls that both experienced and inexperienced observers need to avoid (Prescott, 1957, p. 100; Cronbach, 1963a).

- *Faulty knowledge*. Armed with misinformation and mistaken ideas about human development and behavior, an observer can distort records and the resulting interpretations.
- *Uncritical acceptance of data*. Failure to distinguish between fact and opinion and the acceptance of rumors can lead to distortion of facts.
- *Failure to prespecify objectives*. Obviously, if we don't know what we are looking for, we may observe that which is irrelevant.
- *Conclusion leaping*. Drawing inferences from a single incident and failing to consider contradictory data can lead to faulty conclusions.
- *Failure to consider situational modifiers*. Behaviors result from many influences, and observations must take context into account. A single behavior may have two or more antecedents.
- *Making false inferences from unreliable data*. The tendency to generalize from too limited a sampling of behaviors, and to make judgments on the basis of a few incidents, is a common pitfall.
- *Failure to distinguish behaviors*. In most modern classrooms, for example, many activities take place simultaneously. It is difficult to distinguish relevant from irrelevant behavior.
- *Failure to recognize personal expectations*. Observers must realize that their observations will be screened by their own expectations, preferences, biases, and psychological needs.

- *Failure to record observations accurately.* Observations should be recorded when they occur or immediately afterward. Otherwise, selective forgetting may operate to reduce the validity of the report. There is a tendency to forget things that conflict with our own beliefs and expectations more readily than those that coincide with them.
- *Excessive certainty.* Inferences from observations should be considered tentative and hypothetical until corroborative evidence is obtained.
- *Oversimplification.* One should guard against assigning a single cause to a single behavior; behavior has multiple determinants.
- *Emotional thinking.* We tend to give disproportionate weight to incidents that have had a disturbing effect on us.
- *Substitution fallacy.* There is a tendency to substitute an observed behavior for a desired objective--for example, substituting teacher behavior (process) for the criterion of pupil performance. Observing shop work or physical education, one may tend to substitute "*how* students do something" for the quality of the product. This pitfall suggests the danger of giving such variables as "student-teacher interaction" or "group participation" the status of ultimate criteria.

Applications of Observational Data

Despite the pitfalls to be avoided in collecting and using observational data, such data can make a number of very valuable contributions (Galton, 1987). Observational data may be used to evaluate:

- Group responsibility.
- Group participation.
- Attitudes toward subject matter.
- Individual student interaction with the group.
- Individual student and teacher interaction.
- Teacher and class interaction.
- Individual student achievement.
- Class achievement.
- Unanticipated but related outcomes.
- Individual students in light of instructional hypotheses.
- Teaching techniques.
- Personal and academic problem areas.

Observational methods are particularly useful in studying manipulative and psychomotor skills. In addition, opportunities to gather data in naturally occurring or contrived (simulated) situations are limited only by evaluator creativity. Interpersonal relationships can be observed and objectively summarized. Observation is a means of monitoring important outcomes, particularly those dealing with application skills, without encroaching on instructional time or disrupting the class. The presence of an outside observer may, however, inhibit the "naturalness" of the situation.

Using Rating Scales To Record Observational Data

Rating scales are frequently used to record the results of observations. They can be easily applied in collecting self-report data. The three scales most frequently used in educational settings are numerical, graphic, and checklist. These types of scales are efficient with respect both to the amount of time required to complete them and to the number of individuals who can be rated. Moreover, they do not require sophisticated raters and are relatively easy to construct. On the other hand, rating scales are all too often based on undifferentiated gross impressions and can be susceptible to conscious or unintentional distortion. The use of "anchored points" involving descriptions, behaviors, products, illustrations, or samples can enhance the usefulness of rating scales (Berk, 1986b; Borman, 1986).

Numerical Scales

Numerical scales generally take the form of a sequence of defined numbers. The definitions of the numbers might be in terms of degree of favorableness, frequency, pleasantness, or agreement with a statement. Color or odor, for example, might be rated as:

5	=	Most pleasant
4	=	Moderately pleasant
3	=	Neutral
2	=	Moderately unpleasant
1	=	Most unpleasant

Guilford (1954) cautions against using negative numbers or defining the end categories so extremely that no one will select them. It is probably a good idea to create more categories than one actually intends to use so as to maximize discrimination. One might, for example, use a scale like the following to gather the data:

4 = Almost Always
3 = Usually
2 = Sometimes
1 = Very Seldom

Then combine categories 1 and 2, and 3 and 4 for analysis purposes. Research seems to indicate that, depending upon the nature of the task and the sophistication of the rater, from 7 to about 20 categories at the outside may be used. With checklists, as few as two categories (e.g., present-absent) can be used reliably.

The verbal definitions of the ratings or numbers on a numerical scale may lead to semantic confusion. This problem is well illustrated by a study by Simpson (1944), who asked a population of high-school and college students to indicate what certain terms connoted for them. For example, does the term *often* mean: 65 times in 100 (or even less) or 85 times in 100? The research data revealed that it was the overlap in the range of the middle 50% for adjacent terms that really introduced ambiguity into the measurements. The terms *frequently* had an average of 73% and *generally* had an average of 78%, but the range of the middle 50% of respondents was 40%--80% and 63%--85%, respectively. Such differences undoubtedly serve to lower both the validity and the reliability of ratings. One possible method of overcoming the problem of variable definitions is to specify the frequency to be assigned to each rating term. The following scheme might be used:

5 = Almost always (86% to 100% of the time)
4 = Generally (66% to 85% of the time)
3 = Frequently (36% to 65% of the time)
2 = Sometimes (16% to 35% of the time)
1 = Rarely (0% to 15% of the time)

Although this procedure allows for some latitude in interpretation, it provides raters with a common frame of reference.

Graphic Scales

Another popular rating format is the graphic scale, which is ordinarily a straight line--sometimes vertical but usually horizontal--adorned with various verbal cues to the rater. Graphic rating scales are simple and easy to administer and can be intrinsically interesting, although scoring can be time consuming relative to other methods. Guilford (1954) presents the following example of a graphic scale:

Is the student a slow or quick thinker?

Extremely Slow	Sluggish Plodding	Thinks with Ordinary Speed	Agile-Minded	Exceedingly Rapid

The rater is free to place a checkmark anywhere along the scale. In scoring we might assign numerical weights; for example, the midpoint between Agile-Minded and Exceedingly Rapid might be weighted 5, and so on. Another scoring approach involves the use of a ruler to measure the distance between one of the end categories and the checkmark. The inference of increased precision using such a procedure is probably not justified. One should avoid using extremely long lines, which tend to produce a clustering of ratings. It is probably also a good idea to determine the location of the "high" or "good"end of the scale randomly; for example, it might be on the right for one characteristic and on the left for another.

The form of rating scales may range from simple to fairly structured. Increased structuring better defines the task for the rater. The optimal number of rating points is probably seven to nine. Generally, one should provide more rating categories than one intends to measure to help spread ratings out, get more discrimination, and avoid "clumping"around a particular value. It is not unusual to select an odd number of points so that the "average" will have a central position on the scale, yielding maximum discrimination. An even number of points can sometimes be used profitably, particularly if they are collapsed for scoring. Some raters find it easier to make judgments when numerical point or graphic scales are converted to verbal scales; however, in doing so, an additional interpretation problem is introduced: the possibility that raters will read different meanings into the verbal cues. In any event, rating scales used with care and intelligence can yield very meaningful data. They are particularly adaptable to assessing products and performances.

Checklists

Another popular method of recording the results of observations is the checklist. Even though the observer is checking categories, checklists are still considered ratings since we sum the number of or frequency with which behaviors occurred or characteristics were checked or noted. Checklists can be used by relatively naive raters and tend to make complex judgments unnecessary. It is imperative, however, that the categories be as clear and

precise as possible. The developer of a checklist would be well advised to use behavioral terms if at all possible. Checklists may be used to assess:

- Which instructional objectives or skills have been met or mastered.
- Student interests, hobbies, problems, preferred reading matter, radio or television programs, and the like.
- Student behavior in a variety of settings, especially behavior problems in elementary school.
- Conformity to prescribed sequences of steps in task performance.
- Student products.

An individual's "score" on the checklist may simply be the number of items checked or not checked, or the respondent may be asked to identify a standard or set of criteria for an acceptable product or performance. If some elements of the checklist are more important than others from an evaluation standpoint, differential weights might be applied. A range of three to five possible values would probably suffice.

There is a tendency on the part of some raters to use too many or too few items in a checklist. This response set can be combatted by requiring a fixed number of responses from the rater. Whether or not to use this technique will, of course, depend on the nature and intended use of the checklist.

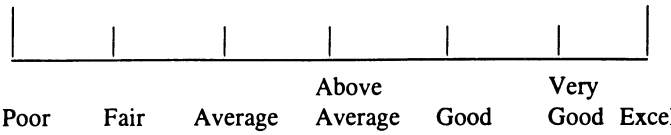
Rating Errors

Despite the many advantages of using rating scales, there are several kinds of errors associated with their application. Among these errors are:

Ambiguity. The tendency to have different raters interpret rating terms in different ways.

Leniency. The tendency to rate or evaluate favorably those whom raters known well higher than they should. This kind of "generosity" can be offset somewhat by adjusting the scale to include a greater proportion of more positive points. For example:

Physical Health



This kind of error is greatest, as one might expect when the rater must face the ratee with the results.

Central Tendency. The reluctance to give extreme ratings. Sometimes raters are reticent to take extreme positions, either positive or negative. This tends to cause ratings to clump up in middle of scale. Sometimes this can be counteracted by spacing descriptive phrases physically further apart on the scale.

Halo. A gross undifferentiated rating on a specific trait or behavior that is biased because it is based on an overall or total general attitude. "Michael Jordan is a great basketball player; therefore, everything he does on the court is excellent--pass, shoot, play defense, etc." Halo can be either positive or negative and can generalize across domains: for example, obnoxious personality = intellectual deficits.

Logical. The tendency to give similar ratings to traits that seem to be logically related in the mind of the rater. If raters view "self-confidence" and "aggressiveness" to be part of the same personality dimension, then they might rate an individual similarly just on that basis rather than the behavior being observed.

Contrast. Some raters will evaluate or describe ratees in a direction opposite of themselves. "I am an extremely well-organized person; therefore, no one can be as organized as I am."

Proximity. Nearness in time or location on a rating form. Traits to be rated on the same page tend to correlate higher than if they were rated on different pages.

Probably the best approach to reducing rating errors is through training and practice. Intelligence is only very modestly correlated with rating reliability. The dictum, "know thyself," however, is good preparation for an observer/rater.

As noted, rating scales can be adapted to measuring a variety of variables: cognitive, performance, or affective. With regard to affective program outcome variables (or process variables), the use of rating scales is only one approach.

APPROACHES TO THE ASSESSMENT OF AFFECTIVE VARIABLES

A larger number of innovative school programs and projects are now concerned with such variables as motivation, self-esteem/concept, climate, morale, and attitudes. There are a variety of approaches to measuring these types of variables.

Cattell, Heist, and Stewart (1950), after extensive review of the literature and personal research, have identified numerous methods that can be applied in the assessment of attitudes and sentiments, or, as they refer to them,

"dynamic traits." Some of these methods are more useful for evaluation purposes than others, but any of the techniques can be adapted for use in a variety of ways. Selections from their list and some additional methods follow:

1. *Money.* The amount of money an individual actually or hypothetically (in a simulated situation) spends on certain activities or courses of action is a direct reflection of his or her attitude and interest. In elementary school, simulation exercises involving purchases can be very revealing.

2. *Time.* The amount of time an individual devotes to certain activities is, to some extent, a reflection of attitude toward them. A survey of time spent by students in various activities can be most revealing of their relative interests.

3. *Verbal expressions.* A host of assessment methods use verbal expressions of attitudes. The Thurstone, Likert, semantic, differential, and opinionnaire methods are illustrative (Remmers & Silance, 1934).

4. *Measures of attention/distraction.* Records of the length of time an individual attends to a stimulus, or a ranking of stimuli (e.g., pictorial) according to responsiveness to them, could profitably be used as measures of attitudes. Failure to respond to certain stimuli is also meaningful behavior. We know only too well about the attention span of students, particularly very young ones. If we can capture that attention and hold it, we at least have a chance to teach.

5. *Fund of information.* The amount or type of information an individual possesses about a certain topic, object, individual, or issue is to some extent a reflection of his or her attitude. There can be interaction between the cognitive and affective domains.

6. *Speed of decision (reaction time).* It may be that decisions are made more quickly about questions on which the subject has the strongest convictions.

7. *Written expressions (personal documents).* Analysis of such documents as biographies, diaries, records, letters, autobiographies, journals, and compositions can be very revealing of an individual's attitudes. Student autobiographies are both revealing of important facets of an individual's life and representative of an opportunity to practice writing skills.

8. *Sociometric measures.* Analysis of friendship choices, social distances, preferences, and the general social structure of a classroom, for example, can be very informative about attitudes.

9. *Misperception/apperception methods.* Provided with ambiguous stimuli, an individual may be tempted to perceive them in accordance with his or her own interests, attitudes, and wishes. A great many projective techniques (e.g., ink blots) have been based on this assumption.

10. *Activity level methods.* There are a number of measures of the individual's general excitement level in response to a stimulus, among them (a) fluency (amount written), (b) speed of reading, and (c) work endurance.

11. *Observations.* The use of standardized reports systematically gathered by trained recorders operating within the limits of an explicitly stated frame of reference has provided extremely valuable data on attitudes per se and on the impact of these attitudes within the individual.

12. *Specific performances and behaviors.* An individual's behavior can illustrate his or her attitudes and their influences. Some argue that behavioral measures are by far the most valid. The indirect methods we commonly use, however, can provide valid data if reasonable precautions are taken and stringent criteria are employed during the developmental stages. Webb et al. (1981) have written an extremely valuable (and intellectually entertaining) reference work with examples of unobtrusive behavioral measures and observational methods.

13. *Physiological measures.* The use of autonomic and metabolic measures can provide useful data in controlled situations. Psychogalvanic response, pulse rate, muscle tension and pressure, and metabolic rate are some of the procedures employed.

14. *Memory measures.* Instructing an individual to learn given material, varying the controversial nature of the content, introducing an unrelated activity to distract the subject, and then asking him or her to recall all or part of the original material is one approach to the use of memory as an instrument of attitude assessment. The selective operation of memory in reminiscence, dreaming, or fantasy may be also analyzed.

15. *Simulations.* Contrived structured or unstructured activities can be used to stimulate and simulate affective responses. The use of role playing, for example, is useful both as an assessment as well as an

instructional technique. Gamelike activities provide particularly good opportunities to observe students under a variety of conditions, particularly with regard to interpersonal relations.

Writing Items for Self-Report Affective Measures

General guidelines and criteria are crucial to the development of statements for affective measures. Obviously, the statements themselves are of critical importance. All the sophisticated analytic techniques in the world will not overcome an inferior item that does not communicate. Edwards (1957a, 1957b) has provided a list of informal criteria for development and editing activities. *Avoid* statements that

- Refer to the past or future rather than to the present.
- Are factual or capable of being interpreted as factual.
- May be interpreted in more than one way.
- Are irrelevant to the psychological object under consideration.
- Are likely to be endorsed by almost everyone or by almost no one.
- Do not reflect the entire range of the affectivity.
- Use language that is complex, ambiguous, obtuse, or indirect.
- Are too long (more than 20 words).
- Contain more than one complete thought.
- Contain universals such as *all*, *always*, *none*, and *never* because they often introduce ambiguity.
- Contain ambiguous words such as *only*, *just*, *merely*, and others of similar nature.
- Are formed with compound or complex sentences.
- Use words that may not be understood by those who are to be given the completed scale (readability).
- Use double negatives.

Most of these suggestions are common sense and are based on the need to communicate. Some of the suggestions are in common with the suggestions for writing cognitive test questions, particularly true-false items.

Corey's Simplified Attitude Scale Construction Technique

Corey (1943) has described a relatively efficient method for constructing an attitude scale. The test development process itself can serve as a learning experience. Its steps are as follows:

1. *Collect a pool of statements*. Each student, for example, might be asked to write three or four statements representing various attitudes toward cheating. Illustrative statements might be:

   Cheating is as bad as stealing.

If a test isn't fair, cheating is all right.

I won't copy, but I often let someone else look at my paper.

A little cheating on daily tests doesn't hurt.

2. *Select the best statements.* Using the criteria for constructing attitude statements described in the previous section, about 50 items might be culled from the initial pool of 100 or 150 statements. Duplicates are eliminated, as are statements that are obviously ambiguous to the teacher or students. The students, for example, might be asked to indicate all those statements on the master list that represent opinions favoring cheating (with a plus sign) and those representing negative opinions about cheating (with a minus sign). An agreement criterion of 80% is suggested; a show of hands is an efficient way to gather these data.

3. *Administer the inventory.* The following directions might be used:

   Directions. This is not a test in the sense that any particular statement is right or wrong. All these sentences represent opinions that some people hold about cheating on tests. Indicate whether you agree or disagree with the statements by putting a plus sign before all those with which you agree and a minus sign before those with which you disagree. If you are uncertain, use a question mark. After you have gone through the entire list, go back and draw a circle around the plus signs next to the statements with which you agree very strongly, and a circle around the minus signs if you disagree very strongly.

The inventory may be duplicated and distributed or administered orally. Discussion should be discouraged. Anonymous administration is preferable.

4. *Score the inventory.* Scoring may be accomplished by either teacher or student. The first step involves identifying those statements that were judged by the entire group (in Step 2) as favoring classroom cheating. Next, the following score values are applied: a plus sign with a circle receives five points, a plus sign alone four points, a question mark three points, a minus sign two points, and a minus sign with a circle one point. Thus, when a person disagrees very strongly with a statement that favors classroom cheating, s/he earns one point; if s/he agrees very strongly with the same statement, s/he gets five points.

   Those statements that express opposition to cheating are scored in the opposite fashion: a plus sign with a circle receives one point,

a plus sign alone two points, a question mark three points, a minus sign four points, and a minus sign with a circle five points. In other words, a student who disagrees very strongly with a statement that opposes cheating actually has a very favorable attitude toward such a practice.

If the inventory contains 50 items, the maximum score possible is 250, which indicates a favorable attitude. The minimum score possible is 50, and an indifference score is in the neighborhood of 150.

The field of measurement is changing as required by the changes in education. We are doing different things in our schools. It seems only logical that different approaches to assessment would follow.

AUTHENTIC ASSESSMENT

The last several years has seen the development of a somewhat nontraditional philosophy about educational measurement. Although the basic elements are not new, they are being organized in a new way that hopefully will provide stronger links between instruction, learning, and measurement. This approach has been variously termed as authentic assessment, direct assessment, or performance assessment. What is desired is a more operational definition of what students can do, what skills they possess, and problems they can solve. There is a definite emphasis on higher order thinking skills. The development of this philosophy was in partial response to dissatisfactions expressed about some current testing practices, especially multiple-choice tests. Dissatisfaction has been expressed particularly as regards state-mandated testing programs established for accountability purposes with their predominant emphasis on minimum competencies and basic skills, and norm referenced interpretations. There is a desire to reduce the "inference gap" between assessment and criterion, hopefully yielding better validity.

The notion of performance assessment is not new; we have been doing it for many years. Assessments of writing skills, typing, computer applications, science laboratory skills, foreign language learning, and a variety of physical education, art, and music outcomes all qualify as performance measurement. There appears to be a cry to do more hands-on assessment where actual student behaviors as well as products can be examined.

Five general characteristics of authentic assessment are:

1. *Value beyond assessment.* The task should be meaningful in and of itself, and not derive value from the fact that it is a "test."

2. *Student constructed response*. Having a record of an actual student behavior observed and evaluated or a product evaluated speaks to the desire to bring criterion and assessment closer together.

3. *Realistic focus*. This characteristic relates to the contemporary need to show students that they are involved in "meaningful" (real world) learning that will have an ultimate tangible payoff.

4. *Application of knowledge*. The need to measure problem solving and critical thinking skills represents educational outcomes being reemphasized.

5. *Use of multiple sources*. A variety of approaches will enhance validity and reliability and allow for greater adaptability to individual student differences.

One example of authentic assessment in the elementary grades is "portfolio assessment." A portfolio is a collection of student products intentionally selected to represent a variety of achievements over a specified period of time, usually a school year. The portfolio must include the actual student productions, a statement of why each was included, together with the criteria used in evaluating them. A final item sometimes included is the report of a student self-evaluation of his or her own selected products or the portfolio as a whole. Student progress can therefore be documented and evaluated over time. The opportunity for student creativity--for example, in writing--is significant. In addition, there is a real sense of ownership and investment in assessment using this approach. Of course, good teachers have been doing this kind of thing for a long time. Assessment definitely gets integrated into learning when approached in this manner.

Despite obvious advantages, several disadvantages are evident. The first relates to time. The development and evaluation of a representative authentic assessment, particularly portfolio, are labor intensive. And just from a physical standpoint, only a limited number of objectives can be documented to any depth. If we are focusing on a large number of respondents, then time and cost can be very significant factors indeed.

If authentic assessments truly represent operational definitions of what we want our students to accomplish, then perhaps "teaching to the test" might no longer be considered an academic crime.

The aggregation of portfolio data useful in evaluation at other than the student level is difficult if not impossible. This is due to the very idiosyncratic nature of the portfolios. Creative methods for identifying and synthesizing themes and trends across portfolios are needed.

The harried (and harassed?) project director or evaluator frequently does not have time to engage in instrument development. The following section describes resources that may be consulted for already existing instrumentation or guidelines for locating relevant instrumentation.

LOCATING INFORMATION ABOUT MEASURING DEVICES

It would be impossible to list, let alone critically evaluate, all those tests that might be of interest to a particular project director or evaluator. The noun *test* is used here in its generic sense, namely "a systematic sample of behavior." A potential user needs information bearing on such questions as: (1) What types of tests are available that will yield the kinds of information I am interested in? (2) What do the "experts" say about the tests I am interested in? (3) What research has been undertaken on this test? (4) What statistical data relating to validity and reliability are available for examination? and (5) With what groups may I legitimately use this test? Answers to these and many other relevant questions may be found in one or more of the following resources:

1. Mental Measurements Yearbooks (Conoley & Kramer, 1989).
2. Test reviews in professional journals.
3. Test manuals and specimen sets.
4. Text and reference books on testing.
5. Bibliographies of tests and testing literature.
6. Educational and psychological abstract indexes.
7. Publishers' test catalogs.

Six additional sources should be mentioned.

8. Test Critiques. Seven volumes of in-depth evaluative studies of over 600 psychological, educational, and business tests (Keyser & Sweetland, 1988). Test Corporation of America also publishes compendia of reviews for testing children, young children, adolescents, adults, and older adults.
9. Tests in Print (Mitchell, 1990). Published by the Buros Institute of Mental Measurements, this is a comprehensive listing of commercially available tests. It is cross-referenced to test reviews in the Mental Measurement Yearbooks. This is a very valuable reference.
10. Directory of Selected National Testing Programs (1987). A detailed listing of over 200 academic selection and certification tests.

11. Educational Testing Service Test Collection Catalogs (1989). A four-volume guide to thousands of tests in such areas as school and reading readiness, screening tests, vocational areas (clerical, mechanical), learning-disabled, cognitive ability and style, creative and divergent thinking, and intelligence.

12. Index to Tests Used in Educational Dissertations (Fabiano, 1989). An index to over 40,000 tests used in educational dissertations covering 1938--1980. It is cross-referenced to the Thesaurus of ERIC Descriptors and users can identify tests (and the population with which they were used) in such areas as achievement, personality, aptitude, physical fitness, and vocations.

13. Directory of Unpublished Experimental Measures (Goldman, Saunders, & Busch, 1974--1982). A reference list of tests identified from journal articles and organized into 23 content categories.

Any competent librarian can assist the reader in accessing other sources of information such as Education Index, Dissertation Abstracts International, Psychological Abstracts, Educational Resources Information Center, Resources in Education, and Current Index to Journals in Education. Computer searches of these and other databases (DIALOG) can greatly facilitate information gathering.

Of the 13 resources listed above, the first three are probably the most immediately informative. These three sources will be discussed in turn, highlighting the types of information that each will provide.

The Mental Measurements Yearbooks

Probably the most useful sources of evaluative information about commercial tests are the Mental Measurements Yearbooks (Conoley & Kramer, 1989). Originated by the late Dr. Oscar K. Buros, they are now the provence of the Buros Institute of Mental Measurements at the University of Nebraska. Up-to-date and comprehensive bibliographies, test reviews, and book reviews are published in the Yearbooks, 10 of which have been published to date. Buros' goal was to develop in the potential user and publisher a critical attitude toward tests and testing, to facilitate communication, and in general to bring about a significant increase in the quality of published tests. Specifically, Buros wanted the Yearbooks "(a) to provide information about tests published as separates throughout the English-speaking world; (b) to present frankly critical test reviews written by testing and subject specialists representing various viewpoints; (c) to provide extensive bibliographies of verified references on the construction, use, and validity of specific tests; (d) to make readily available the critical portions of test reviews appearing in professional journals; and (e) to present fairly exhaustive listings

of new and revised books on testing, along with evaluative excerpts from representative reviews which these books receive in professional journals." The Yearbooks have made a significant and lasting contribution toward these ends.

Some sense of the extensiveness of the Yearbooks can be gained from a brief look at the contents of the 1,014-page Tenth Yearbook (Conoley & Kramer, 1989). The Tenth contains a bibliography of 396 commercially available tests and 569 critical reviews by measurement experts. In addition, there are 1,153 references to the professional literature with an additional 727 from the reviewers.

Due to the need for the most current information available there also exists an easily computer-searchable database for the MMY. It is based on the MMY classification schemes, and a user can access the MMYD with a variety of algorithms to isolate tests for specific variables, populations, price, publication date, and so forth. Between Yearbooks the Buros Institute publishes a softback MMY Supplement with the most recent test reviews.

Test Reviews in Journals

Despite the fact that such authoritative comprehensive sources as the Yearbooks are available, it is often difficult to locate recent data on either new or old tests. Research data or questions related to reliability, validity, and usability, and occasional test reviews are periodically carried in the following journals: the Journal of Educational Measurement, Applied Measurement in Education, Measurement and Evaluation in Guidance, Applied Psychological Measurement, and the Journal of Psychological Assessment. In addition an excellent source of validity studies is the quarterly publication Educational and Psychological Measurement. Of the periodicals listed, E & PM is probably the best single source.

Test Manuals and Specimen Sets

After preliminary decisions have narrowed the field, a potential user should probably obtain specimen sets from publishers. Such a set usually contains a copy of the test questions, scoring key, answer sheet(s), examiner's manual, and occasionally a technical manual. The sets, available at a nominal cost, should be ordered on official school or institution letterhead stationery, because most publishers attempt to insure that their materials are distributed to qualified individuals only in order to maintain security. If there is any question about the qualifications required for the purchase of a particular test, one should consult the publisher's catalog. Following are some excerpts from the 1989 Catalog for Tests, Products and Services for Education for the Psychological Corporation with regard to qualifications for test purchases.

The tests listed in this catalog are carefully developed assessment devices that require specialized training to ensure their appropriate professional use. Eligibility to purchase these tests is therefore restricted to individuals with specific training and experience in a relevant area of assessment. These standards are consistent with the 1985 Standards for Educational and Psychological Testing and with the professional and ethical standards of a variety of professional organizations.

Tests are categorized as being:

Level A: Purchase orders will be filled promptly. Registration is not required. [Author's note: An illustrative Level A instrument would be an occupational interest inventory called the Self-Directed Search.]

Level B: These tests are available to firms having a staff member who has completed an advanced level course in testing from an accredited college or university, or equivalent training under the direction of a qualified supervisor or consultant. Registration is required. [Author's note: Examples of tests at this level would be the Watson-Galser Critical Thinking Appraisal and the Metropolitan Achievement Tests (Sixth Edition).]

Level C: These tests are available only to firms for use under the supervision of qualified professionals, defined as persons with at least a master's degree in psychology or a related discipline and appropriate training in the field of personnel testing. The qualified person may be either a staff member or a consultant. Registration is required. [Author's note: An example of a Level C test would be the Wechsler Intelligence Scale for Children--Revised.]

Once a user has been granted access to a particular test, additional guidelines must be adhered to, relating to test security. Following in an excerpt of such test security precautions.

Purchaser agrees to comply with these basic principles of test security:

1. Test taker must not receive test answers before beginning the test.
2. Test questions are not to be reproduced or paraphrased in any way by a school, college, or any organization or person.
3. Access to test materials is limited to persons with a responsible, professional interest who will safeguard their use.
4. Test materials and scores are to be released only to persons qualified to interpret and use them properly.

5. If a test taker or the parent of a child who has taken a test wants to examine responses or results, the parent or test taker is permitted to read a copy of the test and the test answers of the test taker in the presence of a representative of the school, college, or institution that administered the test.

The test manual is the most informative and readily accessible source of information about a specific test. Directions for administering and scoring the test, brief statistical information about validity, reliability, and norms, a description of the test's development, and suggestions for interpreting and using the test results constitute the usual content of the manual. The reviewer should remember, however, that the publisher has a vested interest, and all tests should be evaluated critically. Most Level B and C tests also have technical manuals available that contain extensive data on the development of the tests.

Well, this has been quite a chapter. Lots of ideas, sources, approaches, methods, and ideas hopefully worth considering. Evaluators and project/program directors need reliable valid data upon which to base decisions. We want to be sure that those data are the best available!

COGITATIONS

1. Are some of the criteria for high-quality data more important than others? Which ones and why?
2. What dimensions do observational methods contribute to measurement and evaluation?
3. What are the difficulties and problems in using observational and opinionnaire methods?
4. What kinds of errors are there in using rating scales and what can be done to control them?
5. What are the pros and cons of developing your own evaluation instruments versus selecting an existing instrument?
6. Are cognitive, performance, or affective outcomes more important to the evaluator? Why?
7. Are affective outcomes legitimate educational goals? Why or why not?
8. What are the major sources of information about already existing instrumentation?

SUGGESTED READINGS

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development. An informative seven-chapter paperback introduction to performance assessment.

Moos, R. H. (1979). *Evaluating educational environments: Procedures, measures, findings, and policy implications*. San Francisco: Jossey-Bass. The climate is a very important moderator of the teaching-learning-testing process.

Nowakowski, J., Bunda, M. A., Warking, R., Bernacki, G., & Harington, P. (1985). *A handbook of educational variables (A guide to evaluation)*. Boston: Kluwer-Nijhoff. An extensive type of thesaurus setting forth in an organized and systematic way variables that might be appropriately addressed in a wide range of educational programs and services.

Payne, D. A. (1992). *Measuring and evaluating educational outcomes*. New York: Merrill/Macmillan. A representative instrument development text.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice*, 6(3), 33-42. An instructional module demonstrating how planned systematic observations can be used to measure communication skills. Generalized step-by-step guidelines are presented. (See also Stiggins, R. J. *Evaluating students by classroom observation: Watching student effort* Washington, D.C.: National Education Association, 1986).

Webb, E. J., et al. (1981). *Unobtrusive measures: Nonreactive research in the social sciences*. (2nd ed) Chicago: Rand McNally. This fascinating collection of methods is presented in a most appealing and understandable form.

Compendia of instruments or references to sources can be very time-saving for the evaluator. Following is a set of references that should prove helpful to the evaluator in search of instrumentation. They would be supplemental to the *Mental Measurements Yearbooks* and other sources mentioned in the chapter.

Bonjean, C.M., Hill, R.J., & McLemore, S.D. (1967). *Sociological measurement: An inventory of scales and indices*. San Francisco: Chandler.

Chun, K., Cobb, S., & French, R.P., Jr. (1975). *Measures for psychological assessment* (A guide to 3,000 original sources and their applications). Ann Arbor: Survey Research Center, University of Michigan.

Johnson, O.G., & Bommarito, J.W. (1971). *Tests and measurements in child development: A handbook.* San Francisco: Jossey-Bass.

Lake, D.G., Miles, M.B., & Earle, R.B. (1973). *Measuring human behavior: Tools for the assessment of social functioning.* New York: Teachers College Press, Columbia University.

Pfeiffer, J.W., & Heslin, R. (1973). *Instrumentation in human relation training.* Iowa City: University Associates.

Robinson, J.P., Shaver, P.R., & Wrightsman, L.S. (Eds.) (1991). *Measures of personality and social psychological attitudes.* San Diego: Academic Press.

Shaw, M.E., & Wright, J.M. (Eds.) (1967). *Scales for the measurement of attitudes.* New York: McGraw-Hill.

Walker, D.K. (1973). *Socioemotional measures for preschool and kindergarten children: A handbook.* San Francisco: Jossey-Bass.

8

MANAGING EVALUATIONS

After reviewing the literature of thousands of notable quotables, George R. Allen (1979, p. viii) has revealed that an agile administrator (or mindless manager):

- Avoids small errors as s/he sweeps on to a grand fallacy.
- Realizes that you can't fool all of the people all of the time, but is satisfied with a majority.
- Will not believe what s/he sees, because s/he is too busy speculating about what s/he does not see.
- Will always claim that s/he has a right to argue about issues which s/he does not understand.
- Has the facility to exaggerate the importance of some pieces of data and to completely overlook the deficiency of others.
- Is a person who takes ideas and facts that we understand and makes them sound confusing.

Anyone who has faced deadlines, an information-hungry public, or personnel schedules can appreciate how easy it is to fall into the traps illuminated so cleverly in the above quotations. If you manage, be ever alert. If you manage an evaluation program, be more so!

Internal and External Evaluators

The evaluator may be internal or external to the project. There are some pros and cons of having an evaluation conducted by either an internal or external evaluator (Huberty, 1988). Following is a contrasting of these two positions along a number of relevant dimensions. The comparisons are very general at best, and are here presented simply to highlight factors to be kept in mind by the project director/manager.

It is impossible to make a definite statement about which way to go. Experts in the field disagree. Budget will obviously play an important role in determining whether to "go external," as will the nature of evaluation questions and estimated duration of the project. A logical resolution to the internal-external problem is to use both, since they can be quite complementary.

Dimension	Internal Evaluator	External Evaluator
Objectivity (Bias control)	Can be good	May be better--fresh perspective
Credibility	Can be good	May be higher
Expense	Constant	May be higher
Knowledge of Project	Very high	Can be quite high, but takes effort
Loyalty to Project	Very high	Distance can cause problems
Sensitivity to Political Pressure	Very high	Less intense
Expertise	Depends on training and experience	May be considerable
Sources Reveal Information	May be quite low	Comfort level may be higher

The intimate involvement of the *internal* evaluator with the day-to-day operation can blend well with the expertise of the *external* evaluator. The skills and knowledge base required for evaluating the project may be enhanced by the use of an outside consultant.

THE ROLE OF THE EVALUATOR

It is probably not a good idea for the evaluator also to serve as the project director. In and above the potential conflict of interest problems, it would be too much work in a project of significant proportions. In addition to the supervision of the entire evaluation program, the evaluator can assist the others in completing a variety of tasks. Evaluators are first of all team members, and by cooperating and sharing information (and, yes, problems also) they can make for an efficient an effective evaluation. The evaluator might also serve partial roles as auditor, advisor, communicator, researcher,

and even historian. Documenting the life history of a project or program evaluation can provide a rich data base for examining factors ultimately related to success or failure. Translation and communication of findings to relevant audiences such as school board, colleagues, taxpayers, central administrators, teachers, parents, and students is one of the extremely important activities likely to significantly influence utilization of results.

Obviously, the evaluator will play many different unique roles, depending on the specific requirements of the evaluative task at hand. A great variety of competencies and skills need to be possessed or developed, and vast quantities of knowledge need to be digested and entered on memory drums. An insight into the varieties of evaluators one might encounter is suggested by the following brief survey reported by Niehaus (1968). After declaring that evaluators range from the "knee-jerk conservatives" to the "wild-eyed liberals," he describes the different kinds he has observed.

> There is the myopic nit picker who seems to have an anxiety compulsion to try to measure the differences between the tickle and itch. There is the cautious creeper who is terrified at the thought of any type of innovation. There is the free swinger who arrives at his evaluation through some weird mixture of ESP and dianoetics and whose ignorance is bolstered by emotion. There is the anxiety evaluator: the worrier, who lives under a perpetual state of existential threat and who feels that if what he evaluates does not coincide with his preconceived and doctrinaire attitudes, all is lost. There is the belaborer of the obvious who after a sizable expenditure of time and effort comes up with a ponderous announcement of something which has been obvious all along--something like the man who suggested, upon first viewing the Grand Canyon, "Something must have happened here." There is also the circumstantial evaluator who uses a hundred words to do the work of one. He gets his observations wound up into such a cocoon that no one can figure out just what he is trying to communicate.

In a more serious vein, it must be accepted that a well-trained, sensitive, effective, and competent evaluator must be both a scientist and a human relations expert. Certain technical skills and knowledges must be mastered. In addition, a great part of the evaluator's time will be given over to working with individuals and groups to plan, implement, and communicate the results of his or her evaluative effort. The role of the evaluator, if viewed objectively and honestly, is an enormous one. To describe its dimensions is an almost

impossible task. It is, therefore, not without some trepidation that the following list of competencies is suggested.

The competent evaluator should be able to:

- Specify information needs for program planning and evaluation.
- Develop a plan for evaluating specified questions.
- Locate, read, and integrate relevant research, measurement, and evaluation literature.
- Specify evaluation objectives and data base requirements in appropriate form(s).
- Critically evaluate a given evaluative research design.
- Relate theoretical evaluation models and real-life requirements.
- Relate input, transaction, and outcome variables.
- Demonstrate appropriate interpersonal relationship skills in working with evaluation team and program staff.
- Differentiate advantages and disadvantages of cross-sectional and longitudinal studies.
- Conduct systems, functions, and task analyses.
- Design an effective measurement-management system.
- Describe evaluation design and analysis requirements in computer programmer or data-processing terms.
- Specify criteria for selection or development of evaluation instruments.
- Apply appropriate data-gathering procedures.
- Apply appropriate data-analysis procedures.
- Make a cost-benefit analysis of a given program or project.
- Use evaluation information to make decisions about programs or projects.
- Administer the activities of an evaluation unit.
- Design a system of data presentation that describes format, responsibility, procedures, recipients, and schedule.

- Redesign and refine evaluation systems based on data implications of previous cycle.
- Create reports that communicate.

This list is obviously not exhaustive. It does reflect, however, certain emphases dictated by real time and experience factors, and it is intended to suggest how and what the evaluator must actually do in a real-life situation to function effectively (and survive).

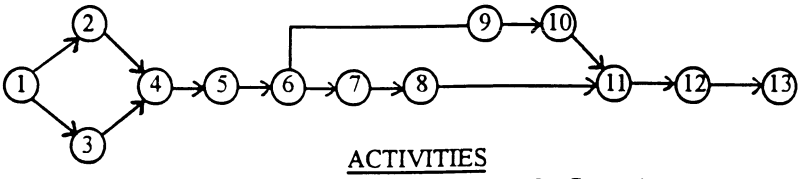
The reader can appreciate the variety of tasks and responsibilities. The foregoing list pleads for shared responsibility and labor. An effective evaluator/manager will call upon expertise no matter where it exists, inside or outside the project.

PLANNING CONSIDERATIONS

Planning must precede management. A plan might be likened to a road map or musical score. The evaluator then might be conceived of as navigator or conductor. Beware of wrong turns and dissonance.

There are two management tools that can significantly facilitate the organization and supervision of an evaluation project: (1) Program Evaluation and Review Technique (PERT) charts, and (2) data collection management charts.

The concept of PERT charting was introduced by the U.S. Defense Department to help with military planning (Cook, 1966). A PERT chart depicts the flow of activities (sometimes called events) in implementing a particular project. It is a helpful graphic device for visualizing the relationships among tasks. In addition, timelines/periods can be added to assist in the allocation of personnel or other resources. Following is a generic PERT chart that might be followed in the creation of an item pool from which multiple test forms are extracted for a statewide assessment program.



ACTIVITIES

1. Start item development process
2. Specify item objectives
3. Review item construction techniques
4. Select development techniques
5. Create test blueprint
6. Write item specifications
7. Train item writers
8. Create item pool
9. Write administrative directions
10. Design field test
11. Conduct field test
12. Conduct item analysis and revise items
13. Select final pool of items

One must be careful not to let the chart become too complex. A separate set of charts for each major component would be advisable. The charts become more useful when time values can be added. In the above example, the amount of time between Creating the test blueprint (5) and Writing the item specifications (6) is quite short (an estimated two weeks), but the actual writing of the items (8) to (9) will take many months. Distance on the chart in a "real" PERT chart should be proportional to time. One should also consider the consequences to the entire system of not meeting critical deadlines ("milestones"). Assigning dollar values to the events can prove to be a valuable budgeting tool.

Another useful administrative tool is the data management plan (chart). It is a simple notion but one that allows the evaluator to see the entire evaluation design at a glance. The plan is a two-way chart listing the evaluation questions along the Y-axes and the evaluation events along the X-axes. Merriman (1972) has created a management plan using Stufflebeam's Context-Input-Process-Product evaluation model (see Chapter 4). The chart in Figure 8-1 is for a not atypical Title I situation where achievement data revealed that the inner-city math performance was two to three years below grade level. The math department proposed competing alternative solutions including remedial programs, curriculum revisions, and innovative instructional methods. Any one or combination of the four component evaluations could be implemented. The "outcome" evaluation, for example, could focus on traditional mathematics performance, as well as holding power of the school and student average daily attendance.

A less complex data management chart can be found in Figure 8-2. Several differences in this chart are obvious in comparison with the format of Figure 8-1. The categories of Data Source and Responsible Person have been used. These two components are particularly important in data collection and should help the evaluator do a better job of planning for a variety of activities such as the ordering of instrumentation, assigning specific responsibility for gathering and data entry, and scheduling data collection times and locations. Another way in which a chart similar to that in Figure 8-1 can be helpful is with regard to creating reports. The chart captures the entire evaluation event and should allow the evaluator to create a more effective communication of the results.

Component \ Phase	Identification of Information Needs	Decision Rule Criteria	Information System Specifications	Data Collection	Data Organization & Reduction	Data Storage and Retrieval	Data Analysis	Reporting
CONTEXT To depict deficiencies in educational opportunities	Socioeconomic status Current status Norms desired Mastery desired Cost-effectiveness	Significant disparity between status and norms or desired mastery level	Source(s) Type of Information Time Requirements Criticality Sample Requirements Quantity Accessibility	Census data Demographic study Standardized tests Pupil grades Pupil attendance Dropout data Attitude survey Opinionnaire Locally constructed tests	Manual Man-Machine • general programs • special programs	Data Bank Knowledge file Machine Manual	Statistical analysis Content analysis Depth study Case study	Formal Reports Written Tabular Informal Reports Oral-group Oral-one-to-one
INPUT To acquire and assess alternative solution strategies	Available solutions to problem Data on prior trials Relationship to context	Feasibility Sufficiency Validity Viability Barriers Tensions Cost-effectiveness	Source(s) Type of Information Time Requirements Criticality Sample Requirements Quantity Accessibility	Review of literature Interviews: LEA personnel, experts, community leaders, parents, residents Panels, seminars, group meetings Transfer from other information centers Observations of demonstrations	Manual Man-Machine • general programs • special programs	Data Bank Knowledge file Machine Manual	Statistical, cost and case study Comparison of prior outcomes of alternatives Consultants for feasibility, barriers, tensions Force field analysis Educator jury for context, validity	Formal reports Written Tabular Informal reports Oral-group Oral-one-to-one
PROCESS To monitor for • a priori barriers • unanticipated problems • progress	Barriers to success Interactive tensions Problem areas Progress benchmarks	Acceptability Utilization Integration Assimilation	Source(s) Type of Information Time Requirements Criticality Sample Requirements Quantity Accessibility	Logs Observations Interviews Group interviews Group debriefing Other instruments: • Attitude scale • Acceptance scale • Facilitant-restraint scale •Structured questionnaire	Manual Man-Machine • general programs • special programs	Data Bank Knowledge file Machine Manual	Content analysis Statistical analysis	Formal reports Written Tabular Informal reports Oral-group Oral-one-to-one
PRODUCT To measure outcomes in relation to objectives	Project outcomes • achievement level • attitude • mastery • cost-effectiveness	Mastery level desired Achievement level desired Growth desired Attitude desired	Source(s) Type of Information Time Requirements Criticality Sample Requirements Quantity Accessibility	Standardized tests Pupil grades Attitude scale Attendance level Dropout rate	Manual Man-Machine • general programs • special programs	Data Bank Knowledge file Machine Manual	Statistical Analysis • pre-post • experimental-control Population analysis Accounting	Tabular Statistical

Figure 8-1 Evaluation Management Plan Using CIPP Question Categories (Merriman, 1972)

Evaluation Question/ (FOCUS)	Instrument/ Variable/ Procedure	Data Source	Level	Experimental and/or Contrast Group	Collection Point	Responsible Person	Analysis Procedure
ONE (Achievement)	Wide range achievement test - revised	S	K-4	E&C	Fall Spring	Teachers Para-professionals	ANCOVA
TWO (Parent Perception)	Evaluator created survey form	P		E&C	Spring	Teachers Project coordinator	Chi-squared
THREE (Student Attitudes)	Evaluator created survey form	S	1-3	E&C	Spring	Teachers Project coordinator	Chi-squared
	Elementary reading attitude survey	S	1-3	E	Fall Spring	Teachers Project coordinator	t-tests
	Behavioral academic self-esteem	S	K-4	E&C	Spring	Project coordinator	Descriptive & t-tests
FOUR (Teacher Commitment)	Teacher ADA	T		E&C	Records	Project coordinator	Descriptive
FIVE (Student Behavior)	Student ADA Discipline Referrals	S	K-3	E&C	Records	Project coordinator	Descriptive

Data Sources: S = Student P = Parent T = Teacher

Figure 8-2 Data Management Plan for K-3 Continuous Progress Project

Relationships between the project director and the evaluator, and their respective staffs, are essential for a productive evaluation. The interface between the two groups is clearly represented in Figure 8-3 which was taken from a 1986 unpublished paper by Daryl Adams of Mankato (MN) State University. The iterative interaction of the two components is readily apparent. It is the reciprocal or "mutual assistance" nature of an effective working relationship that brings success. There are primary responsibilities for both groups, together working toward common goals. Of particular importance is the assistance that the evaluator can provide the project director. The internal evaluator can provide valuable assistance to the director in managing a project by helping the director to

- Clarify the objectives of the program and the evaluation questions.
- Design a plan for implementing the program or project.
- Identify or select measures which will reliably reflect the impact of the program.
- Clarify responsibilities of key program personnel.

- Create an adequate and feasible budget.
- Report accurately the results of the program.

The keynote of all of the relationships, however, is cooperation.

In managing an evaluation project, it is not just the technical details that need attention. We must be sensitive to many interested audiences and organizations. Demands come from many sources. Evaluation can be very provocative and even threatening. If a target feels threatened, fearful, or defensive, it may strike back. We live in litigious times.

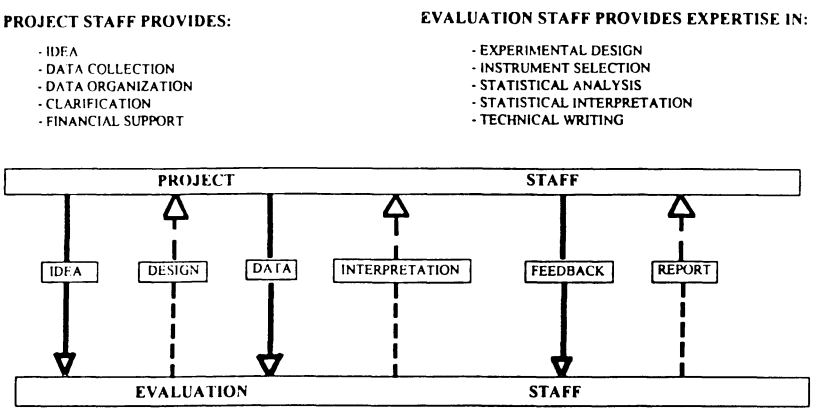


Figure 8-3 Relationships Between Project and Evaluation Staff

LEGAL CONSIDERATIONS

The evaluator must be aware of legal constraints on his or her activities in order to operate in the most efficient and effective manner. The evaluator must feel comfortable in making decisions and not vulnerable when rendering decisions. Thurston, Ory, Mayberry, and Braskamp (1984) have noted that there have been very few specific appellate court decisions related to program or project evaluations or evaluators. This may mean that we are doing such wonderful jobs that nobody wants to take us to court. On the other hand, it may mean that what we are doing is not important enough to become litigious about. As is usually the case, the truth is somewhere in the middle. Evaluation results do not get used to the extent to which they should and

therefore their existence is often overlooked. Thurston, Ory, Mayberry, and Braskamp (1984) have identified four areas that touch on evaluation and evaluators where the law may be of concern.

Defamation. There are four tests for defamation: (1) defamatory language (e.g., alleging criminal guilt, unchastity of a woman, or venereal disease); (2) reference to a specific person or small identifiable group (around 20); (3) publication to a third party (written or overheard); and (4) proof of damage to reputation. Libel is defamation communicated in a permanent form (e.g., written), or slander (spoken word). This is a two-way street, that is, an evaluator can be accused of "fudging" data, and if the allegation cannot be proven, he or she has been defamed. An evaluation report might discuss the incompetent administration of a school, where the school is known, and therefore the principal might be considered defamed. One must, however, be able to prove malice and that the defendant knew of the untruthfulness of the allegation. The best defense against charges of defamation is demonstrable truth.

Contracts. Contracts are important legal instruments that can serve to protect both the evaluator and client. The language should be clear and the responsibilities of each specified. An offer and an exchange for services must be noted and phrases used that reflect on mutual assent and a meeting of the minds. Appendix B contains two sample contracts. The first, Contract for Professional Services, is a very generic form reported by Renzulli (1975). The important items are noted: delivery dates, costs, etc. The document could be used as a front-piece attached to a detailed evaluation plan which specifies responsibilities and tasks. The second example, Contract for Educational Program Audit, is a more detailed form and really helps clarify the assignment of roles, responsibilities, and requirements. Note that there is a "failure to perform on time" clause. Each contract must, of course, be subject to the laws of the state within which it is drawn. One provision that might be considered controversial is that related to dissemination. Quite frequently the evaluator would have more to say about what and how the results are disseminated. With regard to time lines, a projected PERT or GANTT chart would be helpful. A less formal approach is to use a simple one- or two-page letter of agreement.

Malpractice. Following logically from the tort theory of negligence used to adjudicate medical malpractice the major question to be addressed is: Did the evaluation practice deviate significantly from accepted practice and common standards? The importance of adhering to the professional standards outlined in Chapter 2 cannot be underestimated or should not be understated.

Confidentiality of Sources. If confidentiality of sources is not maintained, there may be a breech of contract. The courts have access to sources and source documents. About the only protection an evaluator has is (1) to gather data anonymously, or (2) code the data in such a way that although it can be matched for statistical purposes if necessary the origin cannot be identified, or (3) obtain a waiver from the respondent.

In summary, a legally sensitive evaluator should (1) substantiate conclusions, separating fact and opinions; (2) follow accepted professional practice; (3) have a comprehensive contract; (4) communicate frequently with contractee about issues related to confidentiality; and (5) explicate a detailed evaluation plan cooperatively with contractee.

The assessment of program impact will involve considering an examination not only of changes or differences in outcome (dependent) measures but also of efficiency as reflected in costs.

COST CONSIDERATIONS

The evaluator and his crew of data gatherers will be concerned with two categories of costs. The first and most obvious will relate to costs associated with the actual conduct of the evaluation itself. Mundane expenditures related to salaries and materials purchase will be high on this list. The other category of cost consideration will provide data for decision makers concerning the probable cost-effectiveness and benefits to be derived from implementing the program or project.

Costs Associated with Implementing the Evaluation

A commonly used rule of thumb related to the usual costs associated with the operation of an evaluation program is 10% of the total budget. The range may be from a low of 2% to a high of 20% or more, depending on the number and complexity of the evaluation questions, duration, and the expense of data collection. Many factors contribute to the bottom line evaluation program implementation costs, among them:

Salaries. To this category we must add benefit packages that appear to increase with each change in national government.

Data Collection. A large item in most evaluation projects is the purchase, development, and/or duplication of instrumentation. One could also include the time to train data collectors and their time in the field as theyactually administer instruments. Scoring and recording costs should also be estimated.

Travel/Per diem. Costs for meetings and travel to data collection sites for full-time staff and data collectors as necessary. If the budget allows, presentations at state and national professional meetings would be appropriate both to learn of recent developments and to share results of the home project.

Data Processing. With the advent of computers, data processing has become quite cost-effective. One must still wrestle with data entry. Optical scanning is probably most efficient, but keyboard entry by a whiz also can be quite efficient. Although quite easy to accomplish, data processing and analysis always takes longer than projected.

Printing/Duplicating. Don't forget to allow for the costs of report production, particularly if different reports are projected for multiple audiences. Some instrumentation may have to be copied and duplicated. This can be a big item where survey methods are used.

Office Supplies/Communication. Don't under-phone your work environment. You can never have too many writing pads and pens, and the usual stationery, envelopes, stamps, mailers; the list is endless. Put those computers and printers on your wish list.

Miscellaneous. This catch-all category could include consultants and overhead expenses if relevant.

Planning, perhaps using Program Evaluation and Review Techniques and GANTT charts, and the use of volunteers can help reduce costs.

Evaluating Program Impact Costs

The assessment of the various cost dimensions of an implemented program are much more complex than the budget of the evaluation. At the outset we note that any costs of or funds spent for a program or project represent opportunities or benefits lost. The term *lost* as used here means *spent in lieu of.* In adopting a continuous progress program at Dean Elementary School for grades K--3, one might have opted to spend money on teacher preparation rather than spend it on a site-based management program, a workshop on technology, or the development of student performance assessments. Spending represents a series of choices. Conscious (and hopefully rational) decisions have been made about costs. What cost-related data can decision makers use in assessing program alternatives?

Popham (1993, p. 308) has very succinctly summarized four cost-analysis procedures: cost-feasibility, cost-utility, cost-effectiveness, and cost-benefit.

Cost-feasibility simply considers the cost of implementing a single program or optional programs relative to available funds. We have all faced the problem of affordability, and school systems continually wrestle with budget constraints. If a year-long workshop on teacher mentoring is estimated to cost $17,225 and our entire staff development budget is only $25,000,this workshop may not be an option to consider. We might be able to scale down the workshop or find an adjunct source of funds, but clearly we are involved in establishing at the outset some realistic fiscal parameters.

Cost-utility analysis involves estimates or judgments of the likelihood that each of two or more competing programs will yield outcomes that are useful and of value to the intended target groups. Judgments about utility are gathered about the utility of the programs in arbitrary units; for example, "On a scale from 1 to 10, how much will the Muth Elementary Reading Program enhance our language arts program"? Data would be gathered and aggregated from a number of judges, perhaps a variety of classes of stakeholders such as teachers and administrative personnel. The gross cost of the program is then divided by the mean utility estimates. If the Muth program was an alternative to the Manning program, and both met the cost feasibility criterion, then a decision maker might want to compare utility indices. For example, the following data were gathered:

	Gross Cost	Mean Utility	Utility Index (C/U)
Manning program	$18,725	7.4	$2,530
Muth program	$13,830	5.5	$2,514

We would probably select the Muth program because the utility indices are very close, but the gross cost is considerably less. If the initial gross costs were close but one program clearly had a better utility index, we would go with it.

Examining *cost-effectiveness* requires the evaluator to document the effectiveness (in a criterion sense) of two or more competing programs with a common measure. Turning back to our two reading programs, let's assume that the Campbell Reading Inventory was given on a pre basis to two random samples of fifth grade students (60 students in each group) and the competing programs implemented. A post administration of the Campbell was then conducted. Our measure of effectiveness was the sum of the item gains from pre to post in each of the two 60-student groups. Our cost effectiveness might look like the following:

	Gross Cost	Sum of Item Gains	Cost Effectiveness Index (C/E)
Manning program	$18,725	1320	$14.19
Muth program	$13,830	1484	$ 9.32

We have divided the program costs by the "effectiveness" measure, the sum of the increased number of items answered correctly. In essence we have in the cost-effectiveness index the dollar cost of answering an additional item correctly. Since we get "cheaper" items with the Muth program, we would go with it.

The final cost-analysis method to be considered is *cost-benefit* assessment (sometimes referred to as benefit/cost analysis). This technique allows competing programs having the same or different goals to be compared. The difficult task is the conversion of outcome or benefits to dollar values. The comparison, then, is program dollars to benefit dollars. Benefit dollars can be estimated (the usual procedure) or derived from monitoring over time. Again, with regard to our reading program example let's assume that there was a reasonable way to estimate the dollar value of being able to read better. It would be hypothesized that better readers (or writers, or speakers, or figurers) will make more money or get bigger and better scholarships. The following data might be realized from our two program comparisons.

	Gross Cost	Benefit Value	Cost Benefit (C/B)	Net Benefit
Manning program	$18,725	$16,586	1.13	-$ 2,139
Muth program	$13,830	$27,418	.50	$12,588

It doesn't take a rocket scientist to see that if benefits exceed costs, then we have a winner. In this case, Muth is again in first place.

Doing cost analyses is hard work. They often don't get done because they are not perceived as necessary once the feasibility criterion is met. Comprehensive evaluations will look at *both* impact and cost.

Well, strike up the band, turn on the searchlights, and light the fireworks--that long-awaited moment has arrived: decision-making time! Our efforts to design relevant questions, set standards, and create a framework for collecting and processing data will now be rewarded by having presented the opportunity to answer our most important evaluation questions.

DECISION MAKING

Project and program evaluations are undertaken to reduce uncertainty about effectiveness and efficiency. Data are gathered, sifted, and summarized. Decisions using those data must then be made. These are often complex and sometimes gut-wrenching decisions. There are probably as many theories about how to make evaluation decisions as there are decision makers. The approach to decision making will be influenced very much by the evaluation metaphor being used by the evaluator. There are not only methodological considerations but some very deep-seated philosophical ones as well. The reader is urged to refer back to Chapter 4 and become refreshed with metaphors.

To illustrate elements in the decision making process, the management metaphor delineated by Stufflebeam (1983) will be employed.

Types of Decisions

Stufflebeam et al. (1971) suggest that managers may be called on to make any or all of four basic types of decisions. They are:

Planning:	The primary focus here is on what objectives should be sought. An evaluation of the acceptability of current objectives relative to needs may be undertaken. These are policy decisions, and data are needed to help establish or sustain goals.
Structuring:	Given the objectives identified or confirmed at the planning level, what means are available to meet them? Consideration is made of resources and the advantages or disadvantages of alternative procedures in creating an action plan or design.
Implementing:	Once the objectives have been set and an action plan mounted, evaluation of the implementation must be undertaken. Concern is on how to refine the procedures. These decisions are made almost continuously as the program/ project is monitored and the agreement between intent and design is assessed.
Recycling:	The basic question here relates to goal attainment. Quality control of products or services is a continuing process.

Obviously the nature of decisions made will depend on the nature of the questions asked about a program or project. It will be recalled that in Chapter 3 a developmental project was described which eventuated in the creation of a computer-assisted tutorial program for the teaching of statistical methods. Using that problem setting as a starting point, four kinds of decision questions can be illustrated:

- What are the most important learning needs for students of statistical methods? (Planning)
- Which of three different approaches to meeting student learning needs should be selected and why? (Structuring)
- Is the computer assisted tutorial program operating as intended? (Implementing)
- Has there been an increase in the rate of student achievement of mastery of the learning objectives for statistical methods? Or, Is the program cost-effective? (Recycling)

With the questions in mind, how does one in fact make a decision?

The Decision Making Process

Decision making is selecting among alternatives. Sounds simple enough, right? Yes and no. If systematic procedures have been followed, the actual choice should be routine, but it's all the pre-decision preparation that can be difficult, frustrating, complex, and can drive an evaluator crazy. Stufflebeam et al. (1971) have identified four generic stages in the decision-making process: Awareness, Design, Choice, and Action. Briefly, these stages can be characterized as follows:

Awareness:	The individual(s) with program/project responsibility have recognized, based on a qualitative or quantitative analysis, that a disequilibrium exists and needs are not being met.
Design:	A set of procedures for processing the "decision" is established. This will involve specifying the (1) decision question, (2) decision alternatives, (3) criteria for evaluating the alternatives, (4) decision-rules, and (5) timeline.
Choice:	At this stage, the procedures of the design stage are implemented. Data and standards are contrasted. The criterion data are applied to the decision-rules.

Action: The alternative selected is operationalized by the designated authority.

If the key elements in the process can be identified, they would be the specification of the (1) decision-rule and (2) criteria used in evaluating the alternatives. Ross (1980) has defined decision rules as ". . . interpretive principles used to summarize a large body of information about the value of each alternative in a decision, in order to determine which course of action is most desirable." Returning to our computer-assisted statistics course, we might state two recycling decision rules as follows:

D_1: If the aggregate average mastery of objectives across all winter sections of Statistics 200 is 35% greater than for fall, the program will be continued.

and/or

D_2: If the per student cost for implementing the computer-assisted statistics class is less than $15 per student, the program will be continued.

Where did the standards (35% and $15) come from, you ask? They hopefully came from a systematic standard setting procedure like those described in Chapter 3.

Decision-rules (and standards) are at the heart of the decision-making process. They help us focus, structure, and organize our evaluation study. Targeting the choices should reduce the collection of superfluous data and help reduce bias. Developing the decision-rules in a collaborative atmosphere should also reduce dissonance and conflicts at the end of the evaluation. Finally, the use of decision-rules provides a framework for reporting and communicating the results. There is the danger, however, that the use of decision-rules will so narrow the focus of the evaluation that unintended effects will be missed. One must also be conscious of the fact that projects are living and changing organisms so that questions and therefore decision-rules will change from time to time during the life of a project.

The following quote from Brinkerhoff, Brethower, Hluchj, and Nowakowski (1983) summarizes and highlights the value of effective evaluation management.

> While management of evaluation is important, if it is effective it can go almost unnoticed. Good evaluation management helps rather than hinders audiences, it shortens rather than lengthens the time necessary to run an evaluation, it reduces potentially controversial issues, and it serves well those involved in an evaluation (p. 175).

COGITATIONS

1. How do you respond to the general question, "Is this program/project worth the money it costs?"
2. What are the three most important activities in data collection?
3. What are the pros and cons of the evaluator's also being the project director?
4. Assume you are going to do a cost analysis related to the use of this book. How would you address questions related to *utility* and *benefit*?
5. The answer is: Create a PERT chart. What is the question?
6. Would it be easier from just a plain amount of work standpoint to be an internal or an external evaluator?
7. What and how important are the four major kinds of evaluation decisions?
8. What are the four major generic stages in the decision-making process?
9. What are some critical items that need to be included in an evaluation contract by both evaluator and client?
10. What actions can evaluators take to help protect the professional and legal integrity of their results?

SUGGESTED READINGS

Alkin, M.C.,& Solmon, L.C. (Eds.) (1983). *The costs of evaluation.* Beverly Hills, CA: Sage.

Brinkerhoff, R.O.,Brethower, D.M.,Hluchj, J.T.,& Nowakowski,J.R. (1983). *Program evaluation (A practitioner's guide for trainers and educators).* Boston: Kluwer-Nijhoff. Lots of lists, charts, and graphs to help the reader see how to do it.

Childs, R.A. (1990). *Legal issues in testing.* Washington, D.C.: American Institutes for Research (ERIC Clearinghouse on Tests, Measurement, and Evaluation).

Levin, H.M. (1983). *Cost-effectiveness: A primer.* Beverly Hills, CA: Sage.

Popham, W.J. (1993). *Education evaluation.* (3rd ed.) Boston: Allyn and Bacon. Chapter 14, "Cost Analysis and the Evaluator," contains a very readable analysis of cost assessment techniques.

Thompson, M.S. (1980). *Benefit-cost analysis for program evaluation.* Beverly Hills, CA: Sage.

Thompson, M.S.,& Fortess, E.E. (1980). Cost effectiveness analysis in health program evaluation. *Evaluation Review*, 4(4), 549-567. Seven excellent examples of cost-effectiveness and benefit-cost analyses are presented together with a nine-step methodology.

Wolf, R.M. (1984). *Evaluation in education*. New York: Praeger. Chapter 7, "Program Costs," provides a nontechnical overview of various cost estimation and accounting methods.

9

COMMUNICATING AND USING EVALUATION RESULTS

Illustration of a failure to communicate:

> At a parent-teacher conference the teacher complained to Mr. Bird about the foul language of his children. Mr. Bird decided to correct this behavior. At breakfast he asked his oldest son, "What will you have for breakfast?" The boy replied, "Gimme some of those damn cornflakes." Immediately Mr. Bird smashed the boy on the mouth. The boy's chair tumbled over and the boy rolled up against the wall. The father then turned to his second son and politely inquired, "What would you like for breakfast?" The boy hesitated, then said, "I don't know, but I sure as hell don't want any of those damn cornflakes!" (Yelon & Scott, 1970, p. 5)

Moral: If you want to change behavior, communicate your goals.

Corollary Moral: If you want to have your evaluation results used, be sure to communicate them to the relevant audience.

Throughout the evaluation process, efforts should be made to help insure that the results can be used to help make decisions or simply to obtain a better understanding of the operation of the program or project. In this later sense of the word *use* is intended to mean *illuminate* the program to help see relationships, processes, and outcomes. What we don't want to happen is to have evaluation studies being conducted simply to fulfill a requirement or to be completed for public relations reasons. The "symbolic"use of evaluation is a sham and marks anyone knowingly involved (decision makers or evaluator) as unethical. There is a sense in which a non-decision-oriented use of evaluation data make sense, and that is to persuade an audience of the value of a program. This assumes no intent to deceive. Can evaluators control the use of results? Obviously, no, but they can build safeguards into the entire process.

How do we insure that the evaluation results will be used and in an appropriate way? There is nothing the evaluator can do to guarantee legitimate use of the results, but hopefully best professional practice has been followed throughout the evaluation design and implementation process so that the likelihood of appropriate use of the results has been increased. Staying in touch with the stakeholders is not only smart politically but also assists in laying the foundation for maximizing the utilization of results. *Involving the stakeholders at every stage* should be among the most important tenets in the evaluator's philosophy. Nowhere is that tenet more significant than when it comes time to report and communicate the results of the evaluation.

A RATIONALE FOR REPORTING

Generating relevant data is undoubtedly the most important phase in the evaluation process. The second most critical stage is reporting--or perhaps more correctly, communicating--the results. A wise person once said that data do not become information until they have been communicated, and communication is a two-way process.

Several of the metaevaluation criteria mentioned in Chapter 2 are relevant to the presentation and reporting of evaluation results: timeliness and pervasiveness. Data need to get to the decision makers in time for their consideration, and all the relevant decision makers need to get the data.

Figure 9-1 contains a representation of the relationships between the evaluation itself, the communication process, and utilization. Previous chapters have detailed the evaluation process, touching upon the ways which

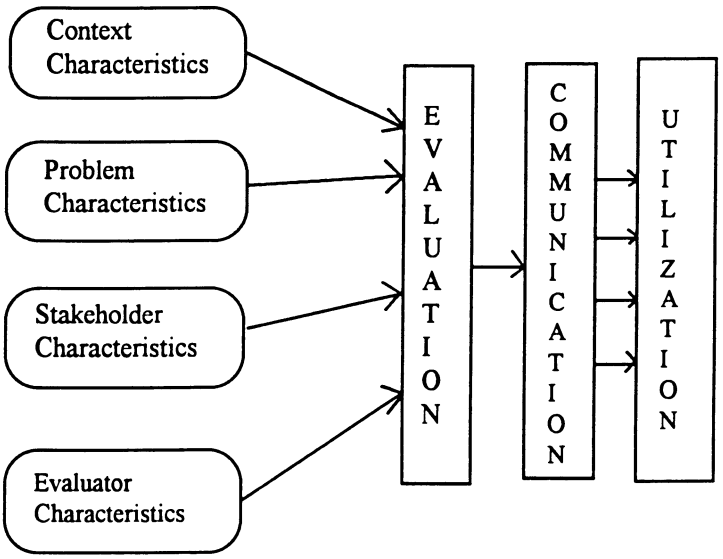


Figure 9-1 Relationships Between Evaluation, Communication, and Utilization Processes (After Johnson, 1993)

the evaluator works with the stakeholder(s), on particular problems, in specific environments, to create a meaningful study. The study is then summarized and reported, and then hopefully used. Research has suggested the importance of several elements in the reporting process.

First one might ask, does having information really influence decision makers? An interesting study by Locatis, Smith, and Blake (1980) confirms the relationship between information and decision making. In an experiment aimed at evaluating the effectiveness of an educational film, these investigators created four "information conditions" and four "treatment conditions." Combinations of the following conditions were made:

	Information		Treatment
Positive:	Two positive film reviews	Control:	Film only--then rate
Conflicting:	One positive/ One negative review	Review:	No film--rate from reviews
Equivocal:	Two neutral reviews	Film-Review:	Rate after film and read reviews
Negative:	Two negative reviews	Review-Film:	Rate after read review and see film

It was found that there was an *information effect*. The investigators demonstrated that when the nature of the information became less negative, there was an increase in the positive evaluation of the films. It was also found that information that was not uniform tended to be associated with lower film quality ratings, and that familiarity moderated the effects. A final and unfortunate finding of the study was that negative information tended to carry more weight than positive information. Perhaps we in education are suffering from the "everything is great" syndrome. It isn't, and therefore contrary views are given disproportionate credence. Full and fair description will help insure balanced interpretation. There's an old adage that states that often it's not what you say, but how you say it that makes a difference.

Two of the crucial elements in any report are style and language. Brown, Braskamp, and Newman (1978) investigated these variables in an interesting simulation study. In the study, an audience of educators was asked to read an evaluation report, which was related to community concerns about testing and grading in the public schools. Specifically, the evaluation report was an advocacy position taken by an external evaluator about the issues of (1) using criterion-referenced standards in classroom assessment, (2) acknowledging the need for more parent conferences, (3) assessing student efforts as well as

performance, and (4) assessing nonacademic achievements. Four reports were developed that varied only in the amounts of jargon and the use of data-based statements: jargon-loaded/objective, jargon-loaded/subjective, jargon-free/objective, and jargon-free/subjective. Educators are probably the worst single group of professionals when it comes to perpetuating a specialized and stylized language. To speak in educationaleze is to obfuscate. Jargon in this study was defined as conveying concepts that were in frequent usage in an educational setting and for which there are more general phrases or words to a professional audience. The audiences were asked to rate the difficulty and technicality of the reports. In addition, they were asked to rate the writer of the report in terms of inferred thoroughness, self-confidence, knowledge about testing and grading, believability, awareness of school needs, logic, practicality, convincingness, and objectivity. The study results did suggest that the style of an evaluation report affects audience perception of the evaluation. The jargon-loaded reports were perceived as significantly more technical and difficult. The data also suggested that the jargon-loaded reports tended to impart a greater air of knowledgeableness and result in a slightly higher likelihood of agreement with the recommendations of the author.

REPORT PREPARATION

The communications paradigm from journalism is applicable in the reporting process. To be effective we need to determine

WHO should say

WHAT to

WHOM and

HOW.

Studies of the message source (WHO) suggest that perceived or actual credibility of the evaluator (e.g., title, credentials, sex) will influence acceptance of results. Content of the message (WHAT) will obviously influence acceptance of results and recommendations. The message must be relevant but also presented in an understandable form (percentages, graphs) and nontechnical language (no jargon). The receiver of the message (WHOM) has characteristics related to acceptance and use of results. Such characteristics as extent and intensity of need for information, intelligence, years of professional experience, position in organization have been shown to be related to receptivity of an evaluation report. Very little research has been conducted on the influence of the nature of the channel (HOW) on report acceptance. Ripley (1985) has shown that receivers who are presented evaluative information via cassette tape or videotape are more likely to agree with an evaluator's recommendations than if the message were in written form. He also found that an advocacy report is more likely to be accepted than an adversarial one.

With this orientation in mind, what are some general guidelines for report preparation?

General Guidelines for Report Preparation

Following are some common-sense things that an evaluator or project director can do that might help in the communication process.

1. *Prepare relevant audiences and where appropriate the public and press for the impending report.* A preliminary written statement or briefing session focused on the aims, interests, and expected outcomes should be provided. Not only does this aid communication but the potential public relations value cannot be underestimated.

2. *Prepare a tabular summary and synopsis of the study results for general release.* Breakdowns accompanied by description of samples, treatments, and factors known or hypothesized to have influenced the results will help convey a clear picture of the conclusions.

3. *Emphasize contingency variables in communicating results.* Ability level of the student body, teacher turnover rate, per-pupil cost factors, socioeconomic base of the school population, class size, pupil-teacher ratios, teacher salaries, and attitudes toward their work, pupil mobility, average daily attendance, average years of teaching experience, educational level of the teachers are just some of the general classes of variables that can impact on a project or program. The presence of competing innovative approaches (e.g., a large number of open classrooms) should also be considered.

4. *Prepare an overall summary of the results.* It is better for the local educational agency to prepare the summary than to allow those less informed about evaluation procedures to make possibly erroneous interpretations or draw unwarranted conclusions. Emphasis should be placed on the broad significance of the results, particularly as they relate to improvement of the teaching-learning process.

Specific Guidelines for Report Preparation

In addition to the forgoing general suggestions, research (Braskamp & Brown, 1980, p. 96) and experience (Patton, 1986, pp. 271-272) have suggested some specific reporting procedures that might substantially help influence decision makers. Among these are suggestions with regard to:

Focus

- Involve stakeholders (and other relevant audiences) in outlining the purposes and format of the report

- Write custom-made reports for specific audiences
- Prepare a mock or simulated report for early discussion
- Solicit interpretations of results from stakeholders
- Link presentation of results to major objectives, issues, and decisions
- Be sensitive to clients' and stakeholders' identities, feelings, and interests

Structure / format

- Begin with a nontechnical executive summary
- Use as many words as necessary but as few as possible
- Use nontechnical terminology and interesting language as much as possible
- Use illustrations, examples, graphs, and figures
- Detail both strengths and weaknesses (limitations) of the evaluation
- Describe evaluation plan and procedures in as great a detail as possible
- Include recommendations for use of results
- Describe criteria used if judgments were made
- Put highly technical information in appendices or a separate report
- Make provisions for minority report and secondary findings if relevant

Presentation

- Pay attention to attractive visual appeal and helpful organization
- Use multimedia where possible, (e.g., slides, videotape)
- Use professionally designed and colorful graphics

Timing

- Make periodic formal and informal reports
- Insure that reports are delivered when appropriate and on time to relevant audiences
- Meet personally with decision makers frequently

It's time to outline the report. Be sure to check the outline with stakeholders, staff, and relevant audiences.

Morris, Taylor Fitz-Gibbon, and Freeman (1987) have presented a useful and generic outline for a final evaluation report. Obviously the final form of the report will depend on (1) the nature of the audience to be addressed, (2) the type of evaluation, and (3) the intended use of the results. Following is an outline of a report based on their suggested seven categories.

I. Executive Summary

(Brief overview of intent, target(s), and results of the evaluation presented in nontechnical terms).

II. Background of Program/Project

A. Origin (Description of context and needs that gave rise to program).

B. Goals of Program/Project (Explicit and implied).

C. Clients Serviced (Demographics, selection process, outline of program/treatment).

D. Characteristics of Program Materials, Activities, and Implementation. (Resources, materials, arrangements, rationale).

E. Staffing Patterns (Tasks, responsibilities, credentials needed, assignments, schedules).

III. Purpose, Intent, and Design of the Evaluation Study

A. Purpose (Intent in doing study and how will results be used).

B. Design (Specification of design type and justification).

C. Instrumentation and Data Management Plan (Outcome measures are described, together with specification of source, responsibilities, and schedule).

D. Monitoring of Program Implementation (Data related to fidelity of implementation and gathered to describe operational program/project process).

IV. Results

A. Description of Implementation Fidelity (Was the promised program delivered?).

B. Summary of Outcome Analyses (Include description of analysis procedures).

V. Discussion

A. Consideration of Internal Validity (Degree of confidence in program-outcome relationships).

B. Value Judgments About Program/Project Results

VI. Costs and Benefits

A. Description of Cost and Benefit Estimation Procedures.

B. Description of Dollar and Nondollar Costs and Benefits (Consideration of start-up costs and implementation costs, as well as what were positive outcomes of program/project).

VII. Conclusions, Recommendations, and Options

The last category of the report raises two controversial questions with which evaluators are frequently faced. They are: (1) Should the evaluator in fact make a value judgment about the program/project? and (2) Should the evaluator make recommendations about the program/project? This author responds in the affirmative to both questions. The evaluator is more than a collector and processor of information. Because of their very intimate association with both the program/project *and* the details of the evaluation, evaluators are probably in the best position to render an overall judgment of the worth of the program (and each of the subcomponents) as well as to suggest changes in the program. These changes may relate to content, process, or application of the results. They may also relate to adjustments in the evaluation design if the study were to be replicated. Whether recommendations are made part of the report is an item that the evaluator and stakeholder (clients) should negotiate.

FACTORS RELATED TO UTILIZATION OF EVALUATION RESULTS

Two kinds of factors influence the use of evaluation results. One is controllable by the evaluator. The basic evaluation itself--focus, design, implementation--will have credibility to the extent that the evaluator had followed sound professional practice. The reporting of the results also can be managed by the evaluator. There are, of course, many sociopolitical factors impinging on the evaluation environment that are outside the control of stakeholders and decision makers. The influence of both these kinds of factors can be seen in the list of problems inhibiting the use of evaluation results compiled by Cox (1977). His list includes:

- Mismatch in roles/styles of stakeholder and evaluator
- Lack of rigor in evaluation design

- Excessive methodological rigor to the detriment of relevance of report
- Lack of utility in recommendations
- Lack of timeliness of report
- External influences (money, politics) having more importance than the data
- Failure to communicate results in a usable way

While the well-designed evaluation may be a thing of beauty and a joy forever, if the results are not used to (1) aid in decision making, (2) support previous decisions or actions, and (3) establish or change attitudes, then a valuable opportunity as well as time, effort, and resources is likely to be lost. Evaluation without utilization is kind of like unrequited love: lots of expectations and anticipations, but no satisfaction or fulfillment.

There are probably three kinds of problems in evaluation utilization, each as bad as the other: underutilization, overutilization (particularly of poorly designed and implemented studies), and nonutilization. Depending on the context, each of these inappropriate practices can have devastating effects. Unscrupulous administrators have been known to use evaluation as a cat's paw to meet their own needs. Davis and Salasin (1975) discussed this problem in what they call "compliance control evaluations," which are usually undertaken by superordinate organizations. Such evaluations are frequently conducted with little required collaborative effort from the staff of project or program being evaluated. Evaluations are easy to critique. A clever administrator can "fuddle" away implementation or evaluation recommendations or can claim that the evaluation supports already implemented changes. Evaluation data, however, should be viewed like the data a driver gets from a speedometer, as an aid in making decisions about progress toward a goal or goals. Superficial or cosmetic utilization is a sham and unprofessional.

The consideration of utilization raises perhaps another valuable role that the evaluator might perform--that of a change agent. In most situations, the utilization of evaluation data is incremental rather than resulting in one or more dramatic changes. Evaluators should really be more than simply providers of information. The evaluator is in a very opportunistic position (in the most positive sense of that phrase) to aid in the application of evaluation results. Through continuing and intimate contact with all segments of the local educational system and community, the evaluator can see and suggest ways to use data.

Attempts to implement utilization may encounter a variety of complicating social, political, psychological, and economic forces that may inhibit application of results. Drawing on the case study method, Alkin, Daillak, and White (1979) have identified eight categories of factors that they found to have had an impact on the utilization of evaluation. These category titles are presented as follows, together with examples of the content of the categories.

Preexisting Evaluation Bounds	*Orientation of the Users*	*Extraorganizational Factors*
School-community conditions	Questions and concerns about the program	Community Influence
Mandated bounds	Expectation for the evaluation	Influence of other governmental agencies
Fiscal constraints	Preferred forms of information	Site-level organizational agencies
Other nonnegotiable requirements		Other information sources
		Teacher and staff views
		Student views
		Costs and rewards
Evaluator's Approach	*Evaluator Credibility*	*Information Content and Reporting*
Use of formal evaluation model	Experience	Substance
Research and analysis considerations (quantitative vs. qualitative)	Training	Format
		Information dialogue
		Delivery method
Choice of role (judgment vs. non-judgment)		
User involvement		
Dealing with mandated evaluation tasks		
Rapport		
Facilitate and stimulate the use of information		
Organization Factors		*Administrator Style*
Interrelationships between site and district		Administration and organizational skills
		Initiative

In the final analysis, utilization may be the key to an effective evaluation system. At times the key is rusty and stubborn, but if we are to unlock the door of truth we must make maximum efforts toward utilization. The discovered truth may not set us free, but it should (1) make us feel better, and (2) make for more effective and efficient realization of human resources.

The interpretation and communication of evaluation results--whether to students, colleagues, parents, or members of the community in general--requires careful preparation and a thorough understanding of evaluation methodology, what influences it, and how the results can be used. It is a very demanding task.

COGITATIONS

1. What elements about an evaluation report make it most credible to you?
2. Assume you have conducted an evaluation of the usefulness of this book as an instructional aid. What can you do to maximize the communication value of your report?
3. At the (a) local and (b) state level, what uncontrollable factors can influence the use of evaluation results?
4. Assume you are going to conduct a study of the utility of computer software aimed at enhancing elementary school students' general problem-solving ability. What steps can you take to help ensure that the results will be used?

SUGGESTED READINGS

American Psychological Association. (1983). *Publication manual of the American Psychological Association.* (3rd ed.) Washington, D.C.: Author.

Braskamp, L.A.,& Brown, R.D. (1980). *Utilization of evaluative information* (New Directions for Program Evaluation, No. 5). San Francisco: Jossey-Bass.

Morris, L.L.,Taylor Fitz-Gibbon, C., & Freeman, M.E. (1987). *How to communicate evaluation findings.* Newbury Park, CA: Sage.

Patton, M.Q. (1986). *Utilization-focused evaluation.* (2nd ed.) Newbury Park, CA: Sage.

Royse, D. (1992). *Program evaluation.* Chicago: Nelson-Hall. See Chapter 9, "Writing the Evaluation Report."

Smith, N.L. (1982). *Communication strategies in evaluation.* Newbury Park, CA: Sage.

Wolcott, H.F. (1990). *Writing up qualitative research.* Newbury Park, CA: Sage.

10

EVALUATING EDUCATIONAL MATERIALS

Educational materials come in all colors and sizes. An evaluator will frequently be called upon to assist a stakeholder, program developer, or director in selecting materials for a particular project. There usually will not be time to do a full-blown systematic comparative study. We must then rely on best professional judgments, usually using accepted absolute standards.

The term *educational materials* is used interchangeably with the term *instructional materials* in this chapter. The terms are probably not identical. The broader term *educational* could include story books and videos that simply present information. Materials considered *instructional* have specific objectives and methodologies built into them which are focused on bringing about behavioral or cognitive change.

It was noted in Chapter 4 that one prevalent evaluation metaphor is that of evaluator as consumer surrogate. An evaluator can serve a very useful function, by doing just that--evaluating materials for particular applications, uses, and situations.

What are some general criteria that need to be addressed in evaluating educational materials, particularly instructional materials?

CRITERIA FOR EVALUATING EDUCATIONAL MATERIALS

One need only briefly peruse the professional journals or wander around the exhibits at national conventions to develop a sense of being overwhelmed by the crush of "new" instructional materials, procedures, and devices. Theoretically publishers and developers should field test their products before they are offered for professional consumption. Unfortunately this does not happen in a majority of the cases. Again we have another opportunity for the evaluator to make a valuable educational contribution. Most devices for evaluating instructional materials must be tailormade for the objectives, uses, and situation. One may wish to focus on a primary or payoff evaluation where data on the impact of the materials are of greatest concern or on secondary evaluation where the attributes or characteristics of the materials themselves are of greater interest. There are probably four major attributes of the secondary type that need to be addressed (Eash, 1972):

> *Objectives*: Are there actual objectives, stated in operational form, and aimed at the use of the materials? In addition, both general and instructional objectives should be available. Hopefully, the objectives flow from some relevant conceptual framework or theory. Are relevant problem-solving and creative skills addressed?

Scope and Sequence: The organization of the material should be such that it follows some conceptually developed pattern, which was based on a task analysis or other relevant research. Is a recommended sequence specified that is responsive to a variety of individual or system needs?

Methodology: Are a variety of approaches and media used? Is the mode that is used based on a rational match of instructional intent and student readiness in the sequence? Is the methodology relatively straightforward, not requiring extensive, complicated preparation?

Student Evaluation: Are procedures provided whereby student progress and achievement can be assessed? Are the procedures available at different levels? Are the evaluation procedures compatible with the objectives and methodology?

To these attributes we might add cost, appearance, and attractiveness to student and teacher, durability, ease of production, and--if special materials are required (e.g., glassware for science activities)--accessibility.

Evaluations of materials can be informal and take place on a small scale. Teachers do it every day when they try out an activity with a student who is having a particular problem. These evaluations might be as extensive as the National Assessment of Educational Progress, the giant federal program aimed at assessing our national educational effectiveness. No matter the size or cost, evaluations, if conducted efficiently and professionally, can yield results that will help improve the educational process.

EVALUATING INSTRUCTIONAL TEXT

Tuckman (1985) notes that educational (instructional) materials may be locally developed or purchased from a commercial vendor. Today's schools reflect, with increasing frequency, a greater variety of instructional materials in the classroom. The availability of an array of materials places great decision demands on teachers and administrators. In addition to the usual textbooks and other written materials we may now find personal computers, videodisc and VCR players, participating-games, manipulatives, and a variety of multimedia devices (e.g., slide-tape and videocassette programs). The most frequently used instructional mode, however, is still the written text.

Lipscombe (1992) has made a significant contribution to the literature with the development of a theory-based text evaluation form. Her intent was to develop a system for evaluating medical text materials. The particular theory used to serve as a foundation for instrument development was Gagné's events of instruction (Gagné, Wages, & Rojar, 1981). The six events used were: (1) Inform Learner of Objectives, (2) Stimulate Recall of Prerequisite Material, (3) Present Stimulus Material, (4) Provide Learning Guidance, (5)

Elicit Performance, and (6) Provide Feedback. Criterion questions were written for each event. These questions were critiqued by instructional and content experts. The content experts in this case were instructors in a variety of medical settings since the first application of the instrument was to evaluate instructional materials in medical science. A total of 32 questions were positively reviewed and grouped for convenience under four general headings: Purpose and Objectives, Content, Structure, and Helping Features. Readability and up-to-dateness (vintage) were also assessed. The "vintage" index (Pittinger, 1978, p. 279) is the difference between the latest copyright date and the cube root of the product of the mean, median, and mode reference dates. An agreement index of .92 was found among the judgments of instructional experts, and .95 among medical experts in evaluating the appropriateness of the criteria. The instrument, *Lipscombe Textbook Evaluation Form* (LTEF) was found to differentiate reliably among three microbiology and seven general medical textbooks (e.g., hematology, immunology). Readability ranged (in a grade level metric) from 19 to 24, and vintage from 4 to 6 years.

Following is a copy of the LTEF. As one can see, the questions are generic and could be used effectively with virtually any text material.

LIPSCOMBE TEXTBOOK EVALUATION FORM

Title_____

Author_____

Copyright_____

Select the response which best represents your evaluation of this specific text with regard to each of the criteria. Read each criterion carefully and place it in the context of an instructional setting. If you do not feel a particular criterion is relevant, simply indicate that by checking Not Relevant. Be sure to respond to each criterion.

	YES	SOMETIMES	NO	Not Relevant
PURPOSE AND OBJECTIVES				
1. Does the preface and/or introduction clearly state the purpose of the text?	___	___	___	___
2. Are objectives clearly stated or at least implied by the use of an advance organizer or an overview of each chapter?	___	___	___	___

LIPSCOMB TEXTBOOK EVALUATION FORM (Cont'd)

	YES	SOMETIMES	NO	Not Relevant
CONTENT				
3. Are the sources of current research findings and other information properly identified?	___	___	___	___
4. Does the author follow through in developing ideas specified in the introduction?	___	___	___	___
5. At the end of each chapter or section, does the author summarize the essential concepts covered?	___	___	___	___
6. Is the material presented to the reader logically through successive levels of difficulty?	___	___	___	___
7. Does the information reflect the current status of the field?	___	___	___	___
8. Is the difficulty level of the text appropriate for the intended audience or use?	___	___	___	___
9. Does the table of contents clearly reflect the organization of the text?	___	___	___	___
10. Is a bibliography included with each chapter or section?	___	___	___	___
11. Are a variety of sources listed in the bibliography?	___	___	___	___
12. Is there a glossary which defines terms new to the students using this level of text?	___	___	___	___
13. Does the index list both major and minor topics?	___	___	___	___

LIPSCOMB TEXTBOOK EVALUATION FORM (Cont'd)

	YES	SOMETIMES	NO	Not Relevant
14. Does the index provide multiple cross listings?	___	___	___	___
15. Is the context of each citation in the index listed?	___	___	___	___
16. Are there one or more appendices which contain useful data to facilitate independent work?	___	___	___	___
17. Is illustrative material such as graphs, tables, charts, and pictures placed appropriately in the text?	___	___	___	___
18. Are graphs and charts easily read and interpreted?	___	___	___	___
19. Are key concepts and terms emphasized by the use of italics or boldface type?	___	___	___	___
20. Are headings and subheadings used frequently and clearly?	___	___	___	___
21. Does the physical layout of the text clearly set off subsections?	___	___	___	___
22. Is color used effectively throughout the book?	___	___	___	___
HELPING FEATURES				
23. Is there a statement describing the knowledge and prerequisite skills needed before mastery of the text material?	___	___	___	___
24. Is there a preface at the beginning of each chapter or section which relates new material to previously learned material?	___	___	___	___
25. Does the text guide the learner to sources outside the text for further information if needed?	___	___	___	___

LIPSCOMB TEXTBOOK EVALUATION FORM (Cont'd)

	YES	SOMETIMES	NO	Not Relevant
26. Are all new terms and concepts defined when they are used or presented?	____	____	____	____
27. Are an adequate number of examples provided to illustrate the concepts presented?	____	____	____	____
28. Are intratextual clues used such as:				
a. first, second, etc., to indicate sequencing of ideas?	____	____	____	____
b. most of all, a key factor, etc., to emphasize important concepts?	____	____	____	____
c. however, on the other hand, etc., to emphasize comparisons?	____	____	____	____
d. for example, such as, etc., for illustration?	____	____	____	____
e. therefore, as a result, etc., to signal a conclusion?	____	____	____	____
29. Are there problems or questions representative of the principles and concepts covered at the end of each chapter or section?	____	____	____	____
30. Do the questions require the student to evaluate and analyze data?	____	____	____	____
31. Are answers provided for the previous questions and problems?	____	____	____	____
32. Are these answers explained so that the student can understand how they were derived?	____	____	____	____

In using the Lipscombe form one could weight the responses (Yes = 2, Sometimes = 1, No = 0) and attain a "score" which in turn could be subjected to some absolute standard setting procedure (see Chapter 3) or used to make comparative judgments among competing texts. The user may also want to employ a text-selection committee to help make the judgments. At either the local or state level, such textbook selection activities can prove to be very interesting events indeed. The interaction of politics, budget, community awareness, and instructional integrity can prove to be a volatile brew.

EVALUATING COMPUTER EDUCATIONAL SOFTWARE

Virtually every teacher and classroom in today's schools has access to a personal computer. As budgets are enhanced, more complex, comprehensive, and creative multimedia installations will be evident. The original applications of computers as *tool* and *tutor* has given way to the more meaningful use as *tutee*. It is in this later area of application where real "learning" can take place, by having the student create programs that teach. Well conceived computer software stimulates the use to interact truly with the system.

Traditional computer software evaluation approaches focus on the expected kinds of secondary characteristics such as content, user friendliness, and nature, extent, and appeal of the graphics. The terms *software*, *educational computer program*, and *courseware* will be used here interchangeably. The *appeal* dimension should not be treated lightly since the program should attract and maintain the user's attention if meaningful interaction is to take place. Different kinds of courseware will have different requirements. Ideally, each different kind of intended-use-courseware should have a custom-made evaluation form (Cohen, 1983). Criteria for evaluating computer software aimed at educational applications should (1) be responsive to what research suggests is sound educational practice, (2) focus on a specific knowledge or skill, and (3) exploit the full potentiality of the microcomputer environment (e.g., tracking progress).

The development of criteria for evaluating computer educational software should address at least two important dimensions of use, namely instructional and technical. The instructional or content criteria should focus on how well the objectives to be accomplished (both cognitive and affective) are achieved, the generalizability of accuracy and up-to-datedness skills developed, and appropriateness of approach relative to the objectives addressed. As with the

selection of any instructional device, the match of software objectives and curriculum objectives is critical. The use of enhancements, be they graphics, color, or sound, are also of concern. Are these enhancements simply window dressings or gimmicks that could prove distracting or do they utilize principles of good mirage design? Technical criteria would involve consideration of ease of interaction and activity level of uses, increments in complexity and skill development, nature of reinforcement and feedback (provision for review), and a whole host of "user friendly" characteristics. The screen format should be relevant, and the user should be able to "navigate" with ease through the program.

Personnel at the Northwest Regional Educational Laboratory in Portland, Oregon, have created a useful courseware evaluation form. The form is reproduced in Figure 10-1. This relatively brief but efficient form could be used by a single evaluator or a committee. Absolute standards could be established, again using some standard setting procedure. The best *evaluators* are, of course, the users. The dictum of sitting with users as they go through the program is very good advice. Nothing can substitute for hands-on experience. It would also be nice to be able to document whether learning took place as a result of using the program. In that regard, vendor support materials are important. Are criterion measures built into the programs or are they available as separates?

The evaluation forms presented in this chapter, although extremely useful at the outset of a systematic examination of effectiveness, have tended to focus narrowly on the instructional rationale of the programs and operational characteristics. Instructional effectiveness will, however, be the ultimate criterion. Assessing that will require an actual field-test.

COGITATIONS

1. What are some characteristics of instructional materials that are important to you as (a) an educator, and as (b) a consumer?
2. Should we evaluate text and computer software with different criteria? Why or why not?
3. How can we assure that there is a link between instructional materials and the curriculum?
4. How does the nature of the instructional material interact with how we assess student learning?

microSIFT COURSEWARE EVALUATION

NORTHWEST REGIONAL EDUCATIONAL LABORATORY

Package title _____ Producer _____
Evaluator name _____ Organization _____
Date _____ ☐ Check this box if this evaluation is based partly on your observation of student use of this package.

SA - Strongly Agree A - Agree D - Disagree SD - Strongly Disagree NA- Not applicable
Please include comments on individual items on the reverse page.

CONTENT CHARACTERISTICS

(1)	SA A D SD | NA	The content is accurate.
(2)	SA A D SD | NA	The content has educational value.
(3)	SA A D SD | NA	The content is free of race, ethnic, sex and other stereotypes.

INSTRUCTIONAL CHARACTERISTICS

(4)	SA A D SD | NA	The purpose of the package is well defined.
(5)	SA A D SD | NA	The package achieves its defined purpose.
(6)	SA A D SD | NA	Presentation of content is clear and logical.
(7)	SA A D SD | NA	The level of difficulty is appropriate for the target audience.
(8)	SA A D SD | NA	Graphics/color/sound are used for appropriate instructional reasons.
(9)	SA A D SD | NA	Use of the package is motivational.
(10)	SA A D SD | NA	The package effectively stimulates student creativity.
(11)	SA A D SD | NA	Feedback on student responses is effectively employed.
(12)	SA A D SD | NA	The learner controls the rate and sequence of presentation and review.
(13)	SA A D SD | NA	Instruction is integrated with previous student experience.
(14)	SA A D SD | NA	Learning can be generalized to an appropriate range of situations.

QUALITY

Write a number from 1 (low) to 5 (high) which represents your judgement of the quality of the package in each division:

— Content
— Instructional Characteristics
— Technical Characteristics

Figure 10-1. Northwest Regional Educational Laboratory Courseware Evaluation Form

TECHNICAL CHARACTERISTICS

(15)	SA	A	D	SD	NA	The user support materials are comprehensive.
(16)	SA	A	D	SD	NA	The user support materials are effective.
(17)	SA	A	D	SD	NA	Information displays are effective.
(18)	SA	A	D	SD	NA	Intended users can easily and independently operate the program.
(19)	SA	A	D	SD	NA	Teachers can easily employ the package.
(20)	SA	A	D	SD	NA	The program appropriately uses relevant computer capabilities.
(21)	SA	A	D	SD	NA	The program is reliable in normal use.

RECOMMENDATIONS

- ☐ I highly recommend this package.
- ☐ I would use or recommend use of this package with little or no change. (Note suggestions for effective use below.)
- ☐ I would use or recommend use of this package only if certain changes were made. (Note changes under weaknesses or other comments.)
- ☐ I would not use or recommend this package. (Note reasons under weaknesses.

Describe the potential use of the package in classroom settings.

Estimate the amount of time a student would need to work with the package in order to achieve the objectives: (Can be total time, time per day, time range or other indicator.)

Strengths/Weaknesses/Other Comments

Source: Reprinted by permission. Developed by the Northwest Regional Laboratory under contract No. 400-83-0005 with the National Institute of Education.

SUGGESTED READINGS

Eash, M.J. (1972). Developing an instrument for assessing instructional materials. In J. Weiss (Ed.), *Curriculum evaluation: Potentiality and reality* (pp. 193-205). Ontario, Canada: The Ontario Institute for Studies in Education.

Flagg, B.N. (1990). *Formative evaluations for educational technologies*. Hillsdale, NJ: Erlbaum.

Merrill, P.F., Hammons, K., Tolman, M.N., Christensen, L., Vincent, B.R., & Reynolds, P.L. (1992). *Computers in education*. Boston: Allyn and Bacon. Chapter 7 addresses the major instructional and presentation criteria, helpful in evaluating educational software.

Siegel, M.A., & Davis, D.M. (1986). *Understanding computer-based education*. New York: Random House. See particularly Chapter 9, "The Computer in the Classroom," for a discussion of criteria for evaluating educational software.

Tuckman, B.W. (1985). *Evaluating instructional programs*. (2nd ed.) Boston: Allyn and Bacon. See in particular Chapter 6, "Surveying the Inputs and Processes from the Classroom."

REFERENCES

Alkin, M. C. (1969). Evaluation theory development. *Evaluation Comment*, 2, 2-7.

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations (Does evaluation make a difference?)*. Beverly Hills, CA: Sage.

Allen, G. R. (1979). *The agile administrator*. Tempe, AZ: Tempe Publishers.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, D.C.: American Council on Education.

Berk, R. A. (1986a). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56(1), 137-172.

Berk, R. A. (Ed.) (1986b). *Performance assessment: Methods & application*. Baltimore: Johns Hopkins University Press.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain*. New York: David McKay.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 100-120). Baltimore: Johns Hopkins University Press.

Braskamp, L. A., & Brown, R. D. (Eds.) (1980). *Utilization of evaluative information* (New Directions for Program Evaluation, No. 5). San Francisco: Jossey-Bass.

Brinkerhoff, R. O., Brethower, D. M., Hluchyj, T., & Nowakowski, J. R. (1983). *Program evaluation (A practitioners guide for trainers and educators)*. Boston: Kluwer-Nijhoff.

Brookover, W. B., LePere, J. M., Hamachek, D. E., Thomas, S., & Erickson, E. L. (1965). *Self-concept of ability and school achievement: II* (USDE Cooperative Research Report, Project No. 1636). East Lansing: Michigan State University.

Brown, C. L. (1980). Issues and concerns in the selection of a control group for project/program evaluation. Paper presented at the Georgia Educational Research Association, Georgia Southern College, Statesboro, GA.

Brown, M. J. M. (1991). Validity and the problem of reality: An issue of trust. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Brown, M. J. M. (1992). Issues in educational reform: How triangulating qualitative and quantitative evaluation methods can enhance understanding. Paper presented at the American Evaluation Association. Seattle, WA.

Brown, R. D., Braskamp, L. A., & Newman, D. L. (1978). Evaluator credibility as a function of report style: Do jargon and data make a difference? *Evaluation Quarterly*, 2, 331-341.

Brown, R. D., & Newman, D. L. (1992). Ethical principals and evaluation standards. Do they match? *Evaluation Review,* 16(6), 650-663.

Brownell, W. A. (1965). The evaluation of learning under different systems of instruction. *Educational Psychologist*, 3, 5-7.

Bruner, E. M. (1984). Introduction: The opening up of anthropology. In S. Plattner & E. M. Bruner (Eds.), *Text, play, and story: The construction and reconstruction of self and society* (pp. 1-16). Prospect Heights: Waveland Press.

Campbell, D. T. (1957). A typology of tests, projective and otherwise. *Journal of Consulting Psychology*, 21(3), 207-210.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (New Directions for Program Evaluation) (pp. 67-78). San Francisco: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 15(2), 195-207.

Cattell, R. B., Heist, A. B., & Stewart, R. G. (1950). The objective measurement of dynamic traits. *Educational and Psychological Measurement*, 10, 224-248.

Cohen, V. B. (1983). Criteria for the evaluation of microcomputer courseware. *Educational Technology*, 23(1), 9-14.

Conner, R. F. (1980). Ethical issues in the use of control groups. In R. Perloff & E. Perloff (Eds.), *Values, ethics, and standards in evaluation*. San Francisco: Jossey-Bass.

Conoley, J. C., & Kramer, J. J. (Eds.) (1989). *The tenth mental measurement yearbook*. Lincoln: University of Nebraska Press.

Cook, D. L. (1966). *Program evaluation and review techniques: Applications in education* (Monograph No. 17). Washington, D. C.: U.S. Office of Education, Office of Education Cooperative Research.

Corey, S. M. (1943). Measuring attitudes in the classroom. *Elementary School Journal*, 43, 437-461.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research*, 56(3), 331-364.

Cox, G. B. (1977). Managerial style: Implications for the utilization of program evaluation information. *Evaluation Quarterly*, 1(3), 499-508.

Crandall, V. C., Katkovsky, W., & Crandall, V. J. (1965). Children's beliefs in their own control of reinforcements in intellectual-academic achievement situations. *Child Development*, 36, 91-109.

Cronbach, L. J. (1963a). *Educational psychology* (2nd ed.). New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1963b). Course improvement through evaluation. *Teachers College Record*, 64, 672-683.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. J., Ambron, S. R., Dronbush, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cronbach, L. J. & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. New York: Macmillan.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examination. *Journal of Educational Measurement*, 21, 113-130.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research,* 53(3), 285-328.

Davis, H. R., & Salasin, S. (1975). The utilization of evaluation. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research*, Vol. 1. Beverly Hills, CA: Sage.

Denzin, N. K. (1970). *The research act: A theoretical introduction to sociological methods*. Chicago: Aldine.

Diener, E., & Crandall, R. (1978). *Ethics in social and behavioral research*. Chicago: University of Chicago Press.

Directory of selected national testing programs. (1987). Phoenix, AZ: Oryx Press.

Eash, M. J. (1972). Developing an instrument for assessing instructional materials. In J. Weiss (Ed.), *Curriculum evaluation: Potentiality and reality* (pp. 193-205). Ontario, Canada: The Ontario Institute for Studies in Education.

Educational Testing Service test collection catalogs. (1989). Phoenix, AZ: Oryx Press.

Edwards, A. L. (1957a). *The social desirability variable in personality assessment and research*. New York: Dryden.

Edwards, A. L. (1957b). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Eisner, E. W. (1976). Educational connoisseurship and criticism: Their form and function in educational evaluation. *Journal of Aesthetic Education*, 10, 135-150.

Eisner, E. W. (1982). An artistic approach to supervision. In T. Sergiovanni (Ed.), *Supervision of teaching: 1982 ASCD Yearbook*. Alexandria, VA: Association for Supervision and Curriculum Development.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Eisner, E. W. (1992). *The educational imagination: On the design and evaluation of school programs* (2nd ed.). New York: Macmillan.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 119-161). New York: Macmillan.

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 162-213). New York: Macmillan.

Fabiano, E. (1989). *Index to tests used in educational dissertations*. Phoenix, AZ: Oryx Press.

Fetterman, D. M. (1984). *Ethnography in educational evaluation*. Beverly Hills, CA: Sage.

Firestone, W. A. (1993). Alternative arguments to generalizing from data as applied to qualitative research. *Educational Researcher*, 22(4), 16-23.

Franck, F. (1973). *The Zen of seeing: Seeing/drawing as meditation*. New York: Vintage Books.

Freed, M. N. (Ed.) (1991). *Handbook of statistical procedures and their computer application to education and the behavioral sciences*. New York: American Council on Education/Macmillan.

Gagné, R., Wages, W., & Rojar, A. (1981). Planning and authoring computer-assisted instruction lessons. *Educational Technology*, 21(9) 17-21.

Galton, M. (1987). Structured observation. In M. J. Dunkin (Ed.), *International encyclopedia of teaching and teacher education* (pp. 142-146). Elmsford, NY: Pergamon.

Geertz, C. (1973). Thick description: Toward an interpretation of culture. In C. Geertz (Ed.), *The interpretation of cultures*. New York: Basic Books.

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Gold, R. (1958). Roles in sociological field observation. *Social Forces*, 36, 217-223.

Goldman, B. A., Saunders, J. L., & Busch, J. C. (Eds.) (1974-1982). *Directory of unpublished experimental measures*, Vols. 1-3. New York: Human Sciences Press.

Green, J. A. (1970). *Introduction to measurement and evaluation*. New York: Dodd Mead.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255-274.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco: Jossey-Bass.

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

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hammond, R. L. (1972). Evaluation at the local level. In P. A. Taylor & D. M. Cowley (Eds.), *Readings in curriculum evaluation* (pp. 231-239). Dubuque, IA: W. C. Brown.

Heath, R. W. (1969). Curriculum evaluation. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.) (pp. 280-283). New York: Macmillan.

Herriott, R. E., & Firestone, W. A. (1983). Multisite qualitative policy research: Optimizing description and generalizability. *Educational Researcher*, 12(2), 14-19.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Horan, J. J. (1980). Experimentation in counseling and psychotherapy. Part I: New myths about old realities. *Educational Researcher*, 9, 5-10.

House, E. R. (Ed.) (1983a). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation*. Boston: Kluwer-Nijhoff.

House, E. R. (Ed.) (1983b). *Philosophy of evaluation*. (New Directions for Program Evaluation, No. 19). San Francisco: Jossey-Bass.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, 3, 1-23.

Huberty, C. J (1988). Another perspective on the role of an internal evaluator. *Evaluation Practice*, 9(4), 25-32.

Hycner, R. H. (1985). Some guidelines for the phenomenological analysis of interview data. *Human Studies*, 8, 279-303.

Instructional Objectives Exchange (1972). *Attitude toward school (K-12)*. Los Angeles: Instructional Objectives Exchange.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 485-514). New York: Macmillan.

Jenkins, T. M., & Dankert, E. J. (1981). Results of a three-month PLATO trial in terms of utilization and student attitudes. *Educational Technology*, 21, 44-47.

Johnson, R. B. (1993). An exploratory conjoint measurement study of selected variables related to innovative educational evaluation participation and instrumental utilization. Unpublished doctoral dissertation, University of Georgia, Athens.

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

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*. Newbury Park, CA: Sage.

Kaluba, J., & Scott, J. (1993). A design for evaluating a computer assisted tutorial statistics course. Unpublished paper, Department of Educational Psychology, University of Georgia, Athens.

Katz, J. (1972). *Experimentation with human beings*. New York: Russell Sage Foundation.

Kaufman, R. A., & English, R. W. (1979). *Needs assessment: Concepts and applications*. Englewood Cliffs, NJ: Educational Technology Publications.

Kennedy, M. M. (1979). Generalizing from single case studies. *Evaluation Quarterly*, 3, 661-678.

Keppel, G. (1991). *Design and analysis (A researcher's handbook)* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keyser, D. J., & Sweetland, R. C. (Eds.) (1988). *Test critiques* (2nd ed.). Kansas City, MO: Test Corporation of America.

Kourilsky, M. (1973). An adversary model for educational evaluation. *Evaluation Comment* (Published by the Center for the Study of Evaluation). University of California at Los Angeles, No. 2, 3-6.

Krueger, R. A. (1988). *Focus groups: A practical guide for applied research*. Newbury Park, CA: Sage.

Kulik, J. A., Bangert, R. L., & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. *Journal of Educational Psychology*, 75(1), 19-26.

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

Lipscombe, M. J. (1992). The use of instructional analysis in the development and application of an instrument to assist in the selection of textbooks used in the training of medical professionals. Unpublished paper, University of Georgia, Athens.

Locatis, C. N., Smith, J. K., & Blake, V. L. (1980). Effects of evaluation information on decisions. *Evaluation Review*, 4(6), 809-823.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin,* 68, 304-305.

Mathison, S. (1988). Why triangulate? *Educational Researcher*, 17(2), 13-17.

McCarty, D. J., Kaufman, J. W., & Stafford, J. C. (1986). Supervision and evaluation: Two irreconcilable processes? *The Clearing House,* 59 (April), 351-353.

McGreal, T. L. (1983). *Successful teacher evaluation.* Alexandria, VA: Association for Supervision and Curriculum Development.

Medley, D. M. (1982). Systematic observation. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (5th ed.) (pp. 841-851). New York: The Free Press.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance.* New York: Longman.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high stakes tests. *Applied Measurement in Education*, 5(3), 265-283.

Merriam, S. B. (1988). *Case study research in education.* San Francisco: Jossey-Bass.

Merriman, H. O. (1972). Evaluation of planned educational change at the local education agency level. In P. A. Taylor & D. M. Cowley (Eds.), *Readings in curriculum evaluation* (pp. 225-230). Dubuque, IA: William C. Brown.

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

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis.* Beverly Hills, CA: Sage.

Mills, C. M. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement,* 20(3), 283-292.

Mitchell, J. W., Jr. (Ed.) (1990). *Tests in print.* Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Morgan, D. L. (1988). *Focus groups as qualitative research.* Newbury Park, CA: Sage.

Morris, L. L., Taylor Fitz-Gibbon, C. J., & Freeman, M. E. (1987). *How to communicate evaluation findings*. Newbury Park, CA: Sage.

National Diffusion Network. (1993). *Educational programs that work* (19th ed.). Longmont, CO: Sopris West.

Niehaus, S. W. (1968). The anatomy of evaluation. *The Clearinghouse*, 42, 332-336.

Owens, T. R. (1973). Educational evaluation by adversary proceeding. In E. R. House (Ed.), *School evaluation: The politics and process*. Berkeley, CA: McCutchan.

Pancer, S. M., & Westhuer, A. (1989). A developmental stage approach to program planning and evaluation. *Evaluation Review*, 13(1), 56-77.

Patton, M. Q. (1986). *Utilization focused evaluation*. Newbury Park, CA: Sage.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Newbury Park, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Payne, B. D. (1976). The influence of passion on strength of matrimonial bonding. *Journal of Interpersonal Relationships*, 17, 44-58.

Payne, D. A. (1982a). Portrait of the school psychologist as program evaluator. In C. R. Reynolds & J. B. Gutkin (Eds.), *Handbook of school psychology* (pp. 891-915). New York: John Wiley & Sons.

Payne, D. A. (1982b). Diary of a mad evaluator. *Educational Evaluation and Policy Analysis,* 4(4), 543-545.

Payne, D. A., & Brown, C. (1982). The use and abuse of control groups in program evaluation. *Roeper Review* (A Journal on Gifted Education), 5(2), 11-14.

Payne, D. A., & Hulme, G. (1988). The development, pilot implementation, and formative evaluation of a grass-roots teacher evaluation system-or the search for a better lawnmower. *Journal of Personnel Evaluation in Education*, 2(1), 259-267.

Payne, S. L. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Perloff, R., Perloff, E., & Sussna, E. (1976). Program evaluation. In M. Rossenzeig & L. W. Porter (Eds.), *Annual review of psychology*. Palo Alto, CA: Annual Reviews, Inc.

Pittinger, C. B. (1978). How up-to-date are current anesthesia and related textbooks? *Anesthesiology*, 49, 278-281.

Popham, W. J. (1987a). Preparing policymakers for standard setting on high-stakes tests. *Educational Evaluation and Policy Analysis,* 9(1), 77-82.

Popham, W. J. (1987b). The shortcomings of champagne teacher evaluations. *Journal of Personnel Evaluation in Education,* 1(1), 25-28.

Popham, W. J. (1990). *Modern educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Popham, W. J. (1993). *Educational evaluation* (3rd ed.). Boston: Allyn and Bacon.

Popham, W. J., & Carlson, D. (1977). Deep dark deficits of the adversary evaluation model. *Educational Researcher,* 6, 3-6.

Prescott, D. A. (1957). *The child in the educative process*. New York: McGraw-Hill.

Preskill, H. (1991). Metaphors of educational reform implementation: A case study of the Saturn School of Tomorrow. Paper presented at the Annual Meeting of the American Evaluation Association, Chicago.

Provus, M. (1971). *Discrepancy evaluation: For educational program improvement and assessment*. Berkeley, CA: McCutchan.

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, Program Effectiveness Panel Recognition Division.

Reichardt, C. S. (1979). The statistical analysis of data from non-equivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Remmers, H. H., & Silance, E. B. (1934). Generalized attitude scales. *Journal of Social Psychology*, 5, 298-312.

Renzulli, J. S. (1972). Confessions of a frustrated evaluator. *Measurement and Evaluation in Guidance*, 5, 298-305.

Renzulli, J. S. (1975). *A guidebook for evaluating programs for the gifted and talented*. Ventura, CA: Office of the Ventura County superintendent of schools.

Ripley, W. K. (1985). Medium of presentation: Does it make a difference in the reception of evaluation information? *Educational Evaluation and Policy Analysis*, 7(4), 417-425.

Rippey, R., Geller, L. M., & King, D. W. (1978). Retrospective pretesting in the cognitive domain. *Evaluation Quarterly,* 2(3), 481-491.

Rist, R. (1980). Blitzkrieg ethnography: On the transformation of a method into a movement. *Educational Researcher*, 9(2), 8-10.

Ross, J. A. (1980). Decision-rules in program evaluation. *Evaluation Review*, 4(1), 59-74.

Rossi, P. H., & Freeman, H. E. (1989). *Evaluation--A systematic approach* (4th ed.). Newbury Park, CA: Sage.

Sanders, J. R., & Nafziger, D. H. (1976). *A basis for determining the adequacy of evaluation designs* (Occasional Paper No. 20). Bloomington, IN: Center on Evaluation Development and Research, Phi Delta Kappa.

Schermerhorn, G. R. & Williams, R. G. (1979). An empirical comparison of responsive and preordinate approaches to program evaluation. *Educational Evaluation and Policy Analysis*, 1(3), 55-60.

Scriven, M. S. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation*. American Educational Research Association Monograph Series on Evaluation, No. 1. Chicago: Rand McNally.

Scriven, M. S. (1972). Pros and cons about goal-free evaluation. *Evaluation Comment* (Published by the Center for the Study of Evaluation), University of California at Los Angeles, No. 4, 1-4.

Scriven, M. S. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process*. Berkeley, CA: McCutchan.

Scriven, M. S. (1974a). Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in education*. Berkeley, CA: McCutchan.

Scriven, M. S. (1974b). Standards for evaluation of educational programs and products. In G. D. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, NJ: Educational Technology Publications.

Scriven, M. S. (1978). Goal-free evaluation in practice. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Scriven, M. S. (1991). *Evaluation thesaurus*. Newbury Park, CA: Sage.

Scriven, M. S. (1993). *Hard-won lessons in program evaluation* (New Directions in Program Evaluation, No. 58). San Francisco: Jossey-Bass.

Sherwood, C. C., Morris, J. N., & Sherwood, S. (1975). A multivariate, nonrandomized matching technique for studying the impact of social interventions. In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research*, Vol. 1 (pp. 183-224). Beverly Hills, CA: Sage Publications.

Simpson, R. H. (1944). The specific meanings of certain terms indicating different degrees of frequency. *Quarterly Journal of Speech,* 30, 328-330.

Sinacore, J. M., & Turpin, R. S. (1991). Multiple sites in evaluation research: A survey of organizational and methodological issues. In R. S. Turpin & J. M. Sinacore (Eds.), *Multisite evaluations* (New Directions for Program Evaluation Issue, No. 50). San Francisco: Jossey-Bass.

Sjoberg, G. (1975). Politics, ethics, and evaluation research. In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research*, Vol. 2 (pp. 29-51). Beverly Hills, CA: Sage Publications.

Stake, R. E. (1967). The countenance of educational evaluation, *Teachers College Record*, 68, 523-540.

Stake, R. E. (1970). Objectives, priorities, and other judgment data. *Review of Educational Research*, 40, 181-212.

Stake, R. E. (Ed.) (1975). *Evaluating the arts in education: A responsive approach*. Columbus, OH: Merrill.

Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation*, 2, 19-22.

Stake, R. E. (1983). Program evaluation, particularly responsive evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models (Viewpoints on educational and human service evaluation)* (pp. 287-310). Boston: Kluwer-Nijhoff.

Steinmetz, A. (1976). The discrepancy evaluation model. *Measurement in Education*, 1(7), 1-7.

Steinmetz, A. (1977). The discrepancy evaluation model. *Measurement in Education*, 2(7), 1-6.

Straton, R. G. (1977). Ethical issues in evaluating educational programs. *Studies in Educational Evaluation*, 3, 57-66.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models (Viewpoints on educational and human service evaluation)* (pp. 287-310). Boston: Kluwer-Nijhoff.

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

Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage Foundation.

Thurston, P. W. (1978). Revitalizing adversary evaluation: Deep dark deficits or muddled mistaken musings. *Educational Researcher*, 7, 3-8.

Thurston, P. W., Ory, J. C., Mayberry, P. W., & Braskamp, L. A. (1984). Legal and professional standards in program evaluation. *Educational Evaluation and Policy Analysis*, 6(1), 15-26.

Tuckman, B. W. (1985). *Evaluating instructional programs* (2nd ed.). Boston: Allyn and Bacon.

Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research*, 35 (March), 492-501.

Tyler, R. W. (1973). Testing for accountability. In A. C. Ornstein (Ed.), *Accountability for teachers and school administrators*. Belmont, CA: Feardon Publishers.

Webb, E. J., Campbell, D. T., Schwartz, R. O., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.

Willis, B. (1978-1979). A design taxonomy utilizing ten major evaluation strategies. *International Journal of Instructional Media*, 6(4), 369-388.

Wolf, R. A. (1969). A model for curriculum evaluation. *Psychology in the Schools,* 6, 107-108.

Wolf, R. L. (1979). The use of judicial evaluation methods in the formulation of educational policy. *Educational Evaluation and Policy Analysis*, 1(3), 19-28.

Wolf, R. M. (1984). *Evaluation in education* (2nd ed.). New York: Praeger.

Worthen, B. R., & Owens, T. R. (1978). Adversary evaluation and the school psychologist. *Journal of School Psychology*, (4), 334-345.

Worthen, B. R., & Sanders, J. R. (1987). *Educational evaluation (Alternative approaches and practical guidelines)*. New York: Longman.

Yelon, S. L., & Scott, R. O. (1970). *A strategy for writing objectives*. Dubuque, IA: Kendall/Hunt.

Yin, R. K. (1984). *Case study research: Design and methods*. Newbury Park, CA: Sage.

APPENDIX A

Multiple Criterion Measures for Evaluation of School Programs

1. Indicators of Status or Change in Cognitive and Affective Behaviors of Students in Terms of Standardized Measures and Scales

   Standardized achievement and ability tests, the scores on which allow inferences to be made regarding the extent to which cognitive objectives concerned with knowledge, comprehension, understanding, skills, and applications have been attained.

   Standardized self-inventories designed to yield measures of adjustment, appreciations, attitudes, interests, and temperament from which inferences can be formulated concerning the possession of psychological traits (such as defensiveness, rigidity, aggressiveness, cooperativeness, hostility, and anxiety).

   Standardized rating scales and checklists for judging the quality of products in visual arts, crafts, shop activities, penmanship, creative writing, exhibits for competitive events, cooking, typing, letter writing, fashion design, and other activities.

   Standardized tests of psychomotor skills and physical fitness.

2. Indicators of Status or Change in Cognitive and Affective Behaviors of Students by Informal or Semiformal Teacher-made Instruments or Devices

   Incomplete sentence technique: categorization of types of responses, enumeration of their frequencies, or rating of their psychological appropriateness relative to specific criteria.

   Interviews: frequencies and measurable levels of responses to formal and informal questions raised in a face-to-face interrogation.

   Peer nominations: frequencies of selection or of assignment to leadership roles for which the sociogram technique may be particularly suitable.

   Questionnaires: frequencies of responses to items in an objective format and numbers of responses to categorized dimensions developed from the content analysis of responses to open-ended questions.

Reprinted with permission of publisher and Dr. Michael from Metfessel, N.S., & Michael, W.B. (1967). "A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs." *Educational and Psychological Measurement*, 27, 931-943.

Self-concept perceptions: measures of current status and indices of congruence between real self and ideal self--often determined from use of the semantic differential or Q-sort techniques.

Self-evaluation measures: student's own reports on his perceived or desired level of achievement, on his perceptions of his personal and social adjustment, and on his future academic and vocational plans.

Teacher-devised projective devices such as casting characters in the class play, role playing, and picture interpretation based on an informal scoring model that usually embodies the determination of frequencies of the occurrence of specific behaviors, or ratings of their intensity or quality.

Teacher-made achievement tests (objective and essay), the scores on which allow inferences regarding the extent to which specific instructional objectives have been attained. Teacher-made ratings scales and checklists for observation of classroom behaviors: performance levels of speech, music, and art; manifestation of creative endeavors, personal and social adjustment, physical well being.

Teacher-modified forms (preferably with consultant aid) of the semantic differential scale.

3. Indicators of Status or Change in Student Behavior Other than Those Measured by Tests, Inventories, and Observation Scales in Relation to the Task of Evaluating Objectives of School Programs

Absences: full-day, half-day, part-day, and other selective indices pertaining to frequency and duration of lack of attendance.

Anecdotal records: critical incidents noted including frequencies of behaviors judged to be highly undesirable or highly deserving of commendation.

Appointments: frequencies with which they are kept or broken.

Articles and stories: numbers and types published in school newspapers, magazines, journals, or proceedings of student organizations.

Assignments: numbers and types completed with some sort of quality rating or mark attached.

Attendance: frequency and duration when attendance is required or considered optional (as in club meetings, special events, or off-campus activities).

Autobiographical data: behaviors reported that could be classified and subsequently assigned judgmental values concerning their appropriateness relative to specific objectives concerned with human development.

Awards, citations, honors, and related indicators of distinctive or creative performance: frequency of occurrence or judgments of merit in terms of scaled values.

Books: numbers checked out of library, numbers renewed, numbers reported read when reading is required or when voluntary.

Case histories: critical incidents and other passages reflecting quantifiable categories of behavior.

Changes in program or in teacher as requested by student: frequency of occurrence.

Choices expressed or carried out: vocational, avocational, and educational (especially in relation to their judged appropriateness to known physical, intellectual, emotional, social, aesthetic, interest, and other factors).

Citations: commendatory in both formal and informal media of communication such as in the newspaper, television, school assembly, classroom, bulletin board, or elsewhere (see Awards).

"Contacts": frequency or duration of direct or indirect communications between persons observed and one or more significant others with specific reference to increase or decrease in frequency or to duration relative to selected time intervals.

Disciplinary actions taken: frequency and type.

Dropouts: numbers of students leaving school before completion of program of studies.

Elected positions: numbers and types held in class, student body, or out-of-school social groups.

Extracurricular activities: frequency or duration of participation in observable behaviors amendable to classification such as taking part in athletic events, charity drives, cultural activities, and numerous service-related avocational endeavors.

Grade placement: the success or lack of success in being promoted or retained; number of times accelerated or skipped.

Grade point average: including numbers of recommended units of course work in academic as well as in non-college preparatory programs.

Grouping: frequency and/or duration of moves from one instructional group to another within a given class grade.

Homework assignments: punctuality of completion, quantifiable judgments of quality such as class marks.

Leisure activities: numbers and types of; time spent in; awards and prizes received in participation.

Library card: possessed or not possessed; renewed or not renewed.

Load: numbers of units or courses carried by students.

Peer group participation: frequency and duration of activity in what are judged to be socially acceptable and socially undesirable behaviors.

Performance: awards, citations received; extra credit assignments and associated points earned; numbers of books or other learning materials taken out of the library; products exhibited at competitive events.

Recommendations: numbers of and judged levels of favorableness.

Recidivism by students: incidents (presence or absence or frequency of occurrence) of a given student's returning to a probationary status, to a detention facility, or to observable behavior patterns judged to be socially undesirable (intoxicated state, drug addiction, hostile acts including arrests, sexual deviation).

Referrals: by teacher to counselor, psychologist, or administrator for disciplinary action, for special aid in overcoming learning difficulties, for behavior disorders, for health defects, or for part-time employment activities.

Referrals: by student himself (presence, absence, or frequency).

Service points: numbers earned.

Skills: demonstration of new or increased competencies such as those found in physical education, crafts, homemaking, and the arts that are not measured in a highly valid fashion by available tests and scales.

Social mobility: numbers of times student has moved from one neighborhood to another and/or frequency with which parents have changed jobs.

Tape recordings: critical incidents contained and other analyzable events amenable to classification and enumeration.

Tardiness: frequency of.

Transiency: incidents of.

Transfers: numbers of students entering school from another school (horizontal move).

Withdrawal: numbers of students withdrawing from school or from a special program (see Dropouts).

4. Indicators of Status or Change in Cognitive and Affective Behaviors of Teachers and Other School Personnel in Relation to the Evaluation of School Programs

Articles: frequency and types of articles and written documents prepared by teacher for publication or distribution.

Attendance: frequency of, at professional meetings or at in-service training programs, institutes, summer schools, colleges and universities (for advanced training) from which inferences can be drawn regarding the professional person's desire to improve his competence.

Elective offices: numbers and types of appointments held in professional and social organizations.

Grade point average: earned in post-graduate courses.

Load carried by teacher: teacher-pupil or counselor-pupil ratio.

Mail: frequency of positive and negative statements in written correspondence about teachers, counselors, administrators, and other personnel.

Memberships including elective positions held in professional and community organizations: frequency and duration of association.

Model congruence index: determination of how well the actions of professional personnel in a program approximate certain operationally stated judgmental criteria concerning the qualities of a meritorious program.

Moonlighting: frequency of outside jobs and time spent in these activities by teachers or other school personnel.

Nominations by peers, students, administrators, or parents for outstanding service and/or professional competencies: frequency of.

Rating scales and check lists (e.g., graphic rating scales or the semantic differential) of operationally stated dimensions of teachers' behaviors in the classroom or of administrators' behaviors in the school setting from which observers may formulate inferences regarding changes of behavior that reflect what are judged to be desirable gains in professional competence, skills, attitudes, adjustment, interests, and work

efficiency; the perceptions of various members of the total school community (parents, teachers, administrators, counselors, students, and classified employees) of the behaviors of other members may also be obtained and compared.

Records and reporting procedures practiced by administrators, counselors, and teachers: judgments of adequacy by outside consultants.

Termination: frequency of voluntary or involuntary resignation or dismissals of school personnel.

Transfers: frequency of requests of teaches to move from one school to another.

5. Indicators of Community Behaviors in Relation to the Evaluation of School Programs

Alumni participation: numbers of visitations, extent of involvement in PTA activities, amount of support of a tangible (financial) or a service nature to a continuing school program or activity.

Attendance at special school events, at meetings of the board of education, or at other group activities by parents: frequency of.

Conferences of parent-teacher, parent-counselor, parent-administrator sought by parents: frequency of request.

Conferences of the same type sought and initiated by school personnel: frequency of requests and record of appointments kept by parents.

Interview responses amenable to classification and quantification.

Letters (mail): frequency of requests for information, materials, and servicing.

Letters: frequency of praiseworthy or critical comments about school programs and services and about personnel participating in them.

Participant analysis of alumni: determination of locale of graduates, occupation, affiliation with particular institutions, or outside agencies.

Parental response to letters and report cards upon written or oral request by school personnel: frequency of compliance by parents.

Telephone calls from parents, alumni, and from personnel in communications media (e.g., newspaper reporters): frequency, duration, and quantifiable judgments about statements monitored from telephone conversations.

Transportation requests: frequency of.

APPENDIX B

CONTRACT FOR PROFESSIONAL SERVICES

BETWEEN

_____

AND

_____

THIS AGREEMENT made by and between the City of _____ _____, hereinafter designated as the City and _____ hereinafter designated as the Evaluator.

ARTICLE 1

Description of Services to Be Performed. The Evaluator shall upon the basis of his training and experience enter upon the elementary and/or secondary level of the _____ Public School System and more particularly at the following schools and/or Programs:

and conduct a survey, analysis, and study of the same in action in order to evaluate and assess the efficiency and productivity of same in action, and upon the completion of the said survey, study, and analysis, to file with _____ _____ a comprehensive and detailed written report thereon in _____ ( ) copies, setting forth the results, conclusions, and findings thereon together with comprehensive recommendations in respect thereto.

ARTICLE 2

Information to be Furnished by City. The City shall furnish to the Evaluator whatever information, data, and statistics are in its possession which may be useful to the Evaluator in carrying out his survey under this contract.

Source: Renzulli, J.S. (1975). *A guidebook for evaluating programs for the gifted and talented*. Office of the Ventura County superintendent of schools, Ventura, CA. Reprinted by permission of author.

ARTICLE 3

Time for Commencement and Completion of Services. The Evaluator shall commence his services under this agreement forwith upon the acceptance of this contract by the City and shall complete his services and file his final report not later than thirty (30) calendar days after the termination date of the contract. Progress reports shall be submitted every _____ days and cover completely all the services performed since the date of the last progress report. Upon receipt of the final report, the Evaluator shall be required to amplify in writing any phase of the report which may not, in the opinion of _____ be sufficiently comprehensive. The Evaluator and the City shall perform their respective activities and responsibilities in accordance with the dates outlined in the attached Statement of Responsibilities.

ARTICLE 4

Compensation to Be Paid by the City. It is expressly agreed and understood by the parties hereto that in no event shall the City pay the Evaluator as full compensation for everything furnished or done by or resulting to the Evaluator in carrying out this agreement a total sum in excess of $_____.

ARTICLE 5

Time for Payment to Evaluator. Application for payment shall be made in accordance with the attached work phase and Payment Schedule. Each application shall show the name of the Evaluator and/or his employees engaged in the study, the number of hours (or days) worked, and his (or their) total wages at the contract rate(s), such abstract to be sworn to by the Evaluator.

ARTICLE 6

Discontinuance of Service. The contract may be terminated by either party at any time by a notice in writing duly mailed or delivered by one party to the other. In the event of contract termination, Evaluator will paid in accordance with Article 4 for all of his services duly performed up to the date of discontinuance.

ARTICLE 7

Laws and Regulations. This contract is subject to all laws of the State of _____, applicable to the administration of public schools and likewise to all Rules and Regulations of the School Committee of the City of _____, as amended to date. A copy of the aforesaid Rules and Regulations is on file in the office of the Secretary of the School Committee.

ARTICLE 8

Final Release. In consideration of the execution of this contract by the City, the Evaluator agrees that, simultaneously with the acceptance of what the City tenders as the final payment by it under this contract, Evaluator will execute and deliver to the City an instrument under seal releasing and forever discharging the City of and from any and all claims, demands, and liabilities whatsoever of every name and nature, both at law and in equity, arising from, growing out of, or in any way connected with this contract, save only such claims, demands, and liabilities as are expressly excepted in said instrument.

ARTICLE 9

Assignment. Neither the School Committee nor the Evaluator shall assign or transfer the respective interests in this contract without the prior written consent of the other.

_____
Evaluator

_____
Title: (If corporation affix seal)

Approved:

_____
Business Manager

Accepted:_____ 19 ____

THE SCHOOL COMMITTEE OF THE CITY OF

_____

_____
Chairman of the School Committee

CONTRACT FOR EDUCATIONAL PROGRAM AUDIT

1. Introduction and Purpose

This contract sets forth the terms and conditions for the provision of and payment for services involved in the conduct of an independent educational program audit (Summative External Evaluation) of the Dropout Consortium Project.

This audit is intended to verify the results of the project's evaluation process and to assess the appropriateness of the project's procedures for evaluating its product, operational process, and management process. In addition, an effort will be made to help insure that the relevant evaluation data are used to influence the modification and improvement of the program.

The types of auditing services to be supplied are specified below and will include the review of pertinent project documents, records, evaluation reports, evaluation instruments, and evaluation data to determine their appropriateness for project evaluation on-site visits to assess the appropriateness of evaluation procedures, and the preparation of written reports containing commendations and recommendations concerning the project's evaluation of the product, process, and management of each program component as well as the overall project.

The Grant Terms and Conditions of Grant Number OEG-0-9-270001-3417(281) between the Contractor and the U.S. Office of Education and the Manual for Project Applicants and Grantees, Dropout Prevention Program, are incorporated herein by reference and made a part of this contract.

2. Audit Personnel

The professional auditing services specified in this contract will be provided by the persons whose resumes of experience and qualifications are provided in Attachment B which is hereby made a part of this contract.

The approximate percentage of time over the period of this contract which each listed person will devote to providing these services is as follows:

Dr. Sig Nificant (Team Leader)	85%
Dr. Seymour Clearly	15%

Should it become necessary to have substitute or additional persons render auditing services during the period for which this contract is in force, the selection of such persons will be agreed to beforehand by all parties to this contract in accordance with the Grant Terms and Conditions referenced above.

All written reports will be the responsibility of the team leader. His substantive contributions will be in the areas of evaluation design and analysis. Dr. Clearly will focus on instructional and staff development operations.

3. Obligations of the Contractor for Document and Persons Availability and for Services

The Contractor agrees:

a. to provide the Audit Team with the following documents prior to the beginning of the audit:

1. Federal regulations
2. U.S. Office of Education guidelines and policy statements pertaining to educational program auditing.
3. Complete program proposal as approved including preliminary proposal, formal proposal, and revisions.
4. Pertinent correspondence between the Contractor and the U.S. Office of Education concerning the proposal, such as evaluative comments and recommended revisions.
5. Copies of the program evaluation design including copies of any available evaluation instruments that are to be used.

b. to make the following available to the Audit Team at the site at any time during the contract period:

1. All data collected through program evaluation procedures including data obtained from the administration of tests, questionnaires, interview schedules, rating scales, and observation schedules.
2. All materials products which are developed or procured for purposes of evaluation or instruction such as curriculum material, videotapes, films, etc.
3. All records concerning students, parents, program staff, and program operation.
4. Students, teachers, program staff, parents, community leaders, and school administrators for interview or observational purposes.

c. to make the following available to the Audit Team at specified times prior to the Audit Team's on-site visits:

1. Copies of any locally constructed or otherwise procured evaluation instruments that are to be used and that were not made available prior to the beginning of the audit.
2. All tabulations, data analyses, and written summaries and interpretations of the results of program evaluation including internal progress reports and quarterly reports submitted to the U.S. Office of Education.
3. Descriptions of the data analysis techniques and procedures used for program evaluation.
4. Evaluator's recommendations for revisions of the evaluation design and all recommendations for program modifications based on evaluation results.

d. to make secretarial service, office space, and local transportation available to the Audit Team during on-site visits and to facilitate or arrange the Audit Team's access to persons for interview or observational purposes as requested within the terms of this contract.

4. Obligations of the Audit Team for Providing Auditing Services

The auditing activities to be performed are based upon certain identified evaluation activities as specified in the Evaluation Design covering the period September 1 - June 30, which is hereby made a part of this contract.

The Audit Team agrees to perform the auditing activities listed below with respect to each of the categories of evaluation activities for each component and for the total program. It is understood that the auditing activities are dependent upon the accomplishment of specific evaluation activities. Should any evaluation activity on which an auditing activity is based not be accomplished or implemented during the period of this contract, it will be noted in the Audit Report, and the Audit Team will not be held responsible for accomplishing the corresponding audit activity. In this event, the amount due to members of the Audit Team will be reduced by a mutually agreed upon amount commensurate with the cost of performing the audit activity.

The auditing activities specified below will provide for assessment of evaluation activities related to the product and process of the project. No attempt will be made within the terms of this contract to perform auditing activities related to program management. The contract may be amended to include these auditing activities; however, should it be deemed advisable by all parties to the contract.

Listed below are the auditing activities to be accomplished with respect to each of the components.

a. Collection of Baseline Data

The Team will assess the procedure for collecting baseline data as specified for each objective by means of at least one of the following:

1. examining a written description of the procedure provided by the program evaluator
2. interviewing program staff members who are responsible for collecting the baseline data
3. examination of the data resulting from the collection procedure
4. direct observation of the data-collecting procedure

b. Collection of Assessment Data

The team will assess the procedure for collecting assessment data as specified for each objective by means of at least one of the following:

1. examining a written description of the procedure provided by the project evaluator
2. interviewing project staff members who are responsible for collecting the assessment data
3. examining the data resulting from the collection procedure
4. direct observing of the data collecting procedure

c. Data Processing and Analysis

The team will assess the procedures for processing and analyzing the evaluation data for each component and for the total program as specified by:

1. examining a written description of the procedures provided by the project evaluator and/or
2. examining the imput and output data for a sample of five cases in order to compare the results of an analysis of the five cases with the results for the total group.

d. Implementation Audit On-Site Visit

The team leader will visit project location in two sites for a minimum of three days (total) to accomplish activities (a-c above). The visit will fall between October 1-13. The team will observe, interview project staff, teachers and students, sample, review data collection devices and procedures, confirm data collection dates, and in general determine degree to which program is operational relative to the proposed plan.

e. Implementation Audit Report

On the basis of the on-site visit a report by the team leader summarizing findings, general reactions, and recommendations will be made to the project director. This report will be made on or before October 18.

f. Interim Audit On-Site Visit

During the period February 13 - February 29, another total three day on-site visit will be made. Again, confirmation of data collection, processing, and analysis activities will be sought. Areas in need of clarification as suggested by the project Interim Evaluation Report will be investigated.

g. Interim Audit Report

A report by the team leader will be submitted on or before March 14, which summarizes the results of the Audit Team's interim on-site visit. In addition, the report will include specific consideration of the Interim Evaluation Report to determine:

1. agreement of reported progress with audit findings
2. accuracy of reported data
3. validity of preliminary findings and recommendations

h. Final Audit Report

The team leader will examine the Final Evaluation Report to determine:

1. the accuracy of presented data with respect to the audit findings
2. the reasonableness of interpretations
3. the validity of conclusions drawn from the data
4. the validity of recommendations based on the conclusions

Focus of this Report will be on the confirmation or questioning of needed program modification that has been prepared as a result of project evaluation. This report will be made on or before August 30 or 30 days after receipt of Final Evaluation Report.

5. Obligation of the Team Leader for Reports

The audit team leader will prepare and submit to the contractor two copies of the following report:

a. Implementation Audit Report (on or before October 18)
b. Interim Audit Report (on or before March 14)
c. Final Audit Report (on or before August 30)

These reports will summarize reactions to evaluation reports where appropriate and on-site evaluation of evaluation activities related to:

a. specification of objectives
b. identification of instruments or data sources
c. collection of baseline data

In particular, the reports will include:

a. Introductory and general comments concerning the quality of the program evaluation and the comparative findings of the program evaluation and the audit.
b. Detailed critique of the process, product, and management evaluation conducted for each component based on an assessment of the instruments used, data collection procedures, data analysis techniques, and data analysis presentation.
c. Description of the audit team's on-site visit findings and their relationship to the evaluation data and reports, on a component-by-component basis; together with a summary of the consistencies and discrepancies and an interpretation of the discrepancies.
d. Recommendations for revisions in the Evaluation Design, including a rationale for each recommendation. (Since auditing objectivity can be maintained only if the selection of a specific corrective action is a project decision, the recommendations will be general rather than specific, suggesting several alternative actions or possible sources of assistance to correct the deficiency.)
e. Confirmation or questioning of the need for project modifications which have been opposed as a result of program evaluation.

Prior to the submission of the final audit report, the Audit Team will be available to discuss the report with the project director. The report will be submitted to the superintendent. The team leader will be responsible for the preparation and certification of both the preliminary and final audit reports.

6. Time and Travel Required to Perform Auditing Services

The number of man/days of professional time required to perform the auditing services specified in this contract are listed in Attachment A.

7. Schedule of Auditing Activities

The approximate schedule for accomplishing the auditing activities specified in this contract have been previously specified and are hereby made a part of this contract. It is understood that this schedule is approximate and flexible. Since the auditing activities are dependent on the evaluation activities, this schedule can be adjusted to meet the specific auditing needs of the program. Arrangements for the two on-site visits will be made at least one week in advance of the visit and at the mutual convenience of the project and the auditor.

8. Confidentiality and Dissemination

a. The Audit Team agrees to maintain the confidentiality of sensitive information acquired as part of the audit such as that contained in student records, teacher ratings, and similar documents.

b. The contractor agrees to assume responsibility for dissemination of the audit findings and reports.

9. Period of Performance, Cost, and Payments

a. The work specified in this contract will begin on September 1, and be completed by June 30, with the exception of the Final Audit Report which will be due August 30, or 30 days after receipt of Final Evaluation Report.

b. In consideration of the services provided as specified herein, the contractor agrees to pay the Audit Team for services rendered. The first payment will be due and payable upon submission of the Implementation Audit Report; the second payment, upon submission of the Interim Audit Report; and the third, following submission of the Final Audit Report.

c. The total amount of payment is based upon the estimates contained in Attachment A which is hereby made a part of this contract.

d. The responsible agents for this contract are those persons whose names appear below as authorized representatives of the Contractor and the Auditor.

e. A 10% penalty will be assessed per each late Audit Report.

10. Auditor and Independent Contractor

In performing services under this contract, Drs. Sig Nificant and Seymour Clearly are independent contractors and nothing herein is to be construed as establishing an employer-employee relationship.

11. Governing Law

The validity, construction, and effect of this contract shall be governed by the laws of the State of Veracity.

12. Entire Agreement

This contract constitutes the only and entire agreement between the parties hereto.

IN WITNESS WHEREOF, the parties hereto have set their hands by proper persons duly authorized on this date, September 1,

for the Board of Education	for Audit Team
BY_____	BY _____
_____	_____
Superintendent	Team Leader

SUBJECT INDEX

NAME INDEX