DOCUMENT RESUME

ED 036 913 EC 004 386

AUTHOR	Swisher, Ginny, Ed.; And Others
TITLE	Evaluation: Processes and Practices. Selected Papers from the Conference for the Evaluation of Instructional Materials (Washington, D.C., April 5-6, 1968).
INSTITUTION	Mid-Atlantic Region Special Education Instructional Materials Center, Washington, D.C.
PUB DATE	Apr 68
NOTE	55p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$2.85
DESCRIPTORS	Decision Making, *Evaluation, Evaluation Criteria, Evaluation Methods, Evaluation Needs, *Evaluation Techniques, *Exceptional Child Research, *Instructional Materials, Models, Objectives, *Program Evaluation, Research Needs, Research Problems

ABSTRACT

Selected papers from the Conference for the Evaluation of Instructional Materials treat the area of evaluation by describing Richard Dershimer's three-part evaluative schema, the Educational Products Information Exchange approach to evaluating instructional materials, the evaluation procedures in Montgomery county (Maryland), the Consumers Union Model, and a proposed management model for evaluation data. References accompany the first two papers, and commentary remarks regarding the middle three papers are provided. (LE)

Evaluation: Processes and Practices

Selected Papers from
The Conference for
the Evaluation of
Instructional Materials,
April 5–6, 1968

MID-ATLANTIC REGION SPECIAL EDUCATIC [illegible] STRUCTIONAL MATERIALS CENTER

Evaluation: Processes and Practices

selected papers from
The Conference for the
Evaluation of Instructional
Materials, Washington, D.C.

April 5-6, 1968

edited by

Ginny Swisher
Carol Gross
Ellen Cramer

Table of Contents

Preface

The present monograph consists of selected speeches given at the Conference for the Evaluation of Instructional Materials held in Washington, D. C., on April 5 and 6, 1968. This conference was sponsored by the Special Education Instructional Materials Center,* Department of Special Education, George Washington University. Although the primary aim was to focus specifically on the evaluation of materials for handicapped children, it was recognized that evaluation procedures involve techniques of a general nature: consequently the conference concerned itself with evaluation procedures *per se*.

Those invited to the conference included staff members of the fourteen federally funded special education instructional materials centers and CEC-ERIC, which are members of the national network of instructional materials centers for special education. The centers are funded by the Bureau of Education for the Handicapped, U. S. Office of Education. The general purpose of the conference was to provide this network with information and guidelines which would be pertinent to the evaluation of instructional materials.

*As of September, 1968, the name of the George Washington Special Education Instructional Materials Center was changed to the Mid-Atlantic Region Special Education Instructional Materials Center.

Foreword

REVISIONISM

In retrospect, I now wish that my presentation, contained elsewhere in this monograph, had started with the origin of evaluation needs. The ultimate goal of evaluation activities is to acquire sound, necessary, and sufficient information ("data") for the decision maker. We need to ascertain whether those people seeking information are the ones who will make the decisions; we need to know what decisions are to be made, whether additional information is necessary or relevant, and so forth. We need to look at evaluation from the point of view of the decision-making processes constituting the real system which includes who and what are affected by the decision.

Whether or not we go back and start at the beginning, we must focus more of our attention on the consumer of evaluative information. Recognition of the importance of functional dissemination of decision processes and skills *and* information for achieving effective change has yet to progress beyond the level of lip service.

Our attention is trained upon models and data. This is our communal heritage from the field of research. Research has yet to achieve an effective functioning dialogue with the educational scene. Research has reached that level of systematization where flexibility for investigating unknown interactions becomes greatly restricted. A "system" has internal governors and the "navigator" steers through chartered seas.

I fear evaluation may achieve system status at a level of incompleteness. Evaluation is pragmatic. It is "for" not "is". Can it "be"?

M. H. Moss
Conference Chairman

*Evaluation and Decision Making**

RICHARD A. DERSHIMER, PH.D.
Executive Officer, American Educational Research Association (AERA)
Washington, D.C.

This paper is to encourage and improve the sophistication of evaluations of all kinds. I am focusing on *educational programs* because of my background. Far too many supervisors and administrators are frustrated because they feel increasing pressure to assess their programs–or segments of them–only to be reminded by the specialists that the methodologies of evaluation are growing increasingly complex. Federal directives specify that programs must assess what benefits federal funds have produced, but research consultants explain that test scores and opinion questionnaires will not provide the answers.

The dilemma is heightened by statements like the one made by Louise Tyler (1968) that evaluation is now "imperative because curriculum and instructional materials development has become centralized." Yet, we are told that even the national curricular projects, with great educational leaders and hundreds of thousands of dollars from foundation and federal funds, have not been adequately assessed. Today's school administrators hear school boards use terms like *cost effectiveness, objectives*, and *assessment* with increasing frequency. They hear men like Scriven (1967) state, "Business firms can't keep executives or factories when they know they are not doing good work, and a society should not have to retain textbooks, courses, teachers, and superintendents that do a poor job when a good performance is possible."

The solution far too often is to "go through the motions," that is, gather some test score data, tabulate questionnaire results, obtain some testimonies, and write a flashy report; the data are not respected and only slightly used–if at all. A frequently used alternative is to postpone any serious evaluation until a fully qualified man can be found. The problem with both of these alternatives is that the potential for evaluation that exists in so many districts is overlooked, and valuable data are lost or ignored.

This paper will provide a rationale for the "let's get started" approach. It is meant to be a supporting document for the supervisor who is arguing for even more evaluation with his immediate supervisor, whether he is an assistant superintendent, superintendent, or chairman of the board. It will attempt to summarize and interpret the major disputes about evaluation that are most relevant to the local scene and to provide a plan that should help districts improve the way they now evaluate.

*I acknowledge the generous contributions of Robert Stake, Robert Panos, Malcolm Provus, and Glenn Boerrigter, all of whom made valuable suggestions, most of which have been incorporated in the final draft.

DEFINITIONS

Probably the most frequently used definition of evaluation, and the one which we shall accept for the purposes of this paper, is provided by Cronbach (1963). He sees evaluation, broadly conceived, as "the collection and use of information to make decisions about an educational program." It is interesting that he uses the term *information,* and not the more precise term, *data.*

As valuable as this definition is, it begs a very critical issue. Does evaluation merely attempt to describe the situation under study, or does it attempt to judge how adequate, effective, or valuable something is? Most evaluators in the past have only described matters like how well or how poorly children achieved compared to certain norms, or how adequate the environment of the learner or the preparation of the professional was. Stake (1967) takes a firm stand on this issue and claims that *both* [italics his] description and judgment are essential—in fact, they are the two basic acts of evaluation. Any individual evaluator may attempt to refrain from judging or from collecting the judgments of others. Any individual evaluator may seek only to bring to light the worth of the program. But their evaluations are incomplete. To be fully understood, the educational program must be fully described and fully judged."

Again, for the purpose of this paper, I concur with this extension of the definition by Cronbach, with one modification: the judgments of the evaluator must be limited to the question under study and kept within the boundaries for which there are supporting data. The evaluator is a technician and should have special insights and perceptions. But in nontechnical matters his judgment should be only one of many sought by the ultimate decision maker—supervisor, superintendent, or school board. He can be asked how effective a new elementary school science program is, for example. His reactions should be weighed against those of the business manager who may have been analyzing the costs of the new program, the assistant superintendent who knows the feelings of the parents about the existing elementary science program, and the principal who knows the qualifications and reaction of the teachers.

The term *educational program* used in this sense refers to those aspects of any school situation involving the direct interaction between a group of professionals and a group of students in a school. Educational materials can be included in this definition; budgets cannot. Organization of teaching teams can be included; administrative reorganization of the central office cannot. Evaluation of the progress of a single student or even several students by a single teacher is not included in our definition.

Evaluation is an activity or set of activities initiated or utilized to provide data for major operational decisions in the schools. It is only one of several sources of data for those major decisions even though, on matters relating to the educational program, it may be most relevant and most significant. Evaluation serves three other purposes, however: (1) providing more systematic ways of gathering data for many other purposes; (2) sharpening the objectives of any organization; and (3) identifying and clarifying aspects of the situation where revision or improvement is most needed and/or desired.

A SCHEMA FOR EVALUATION

All too frequently, in the mind of the administrator, evaluation is equated with checklists, tests, questionnaires—in other words, with techniques rather than an overall plan. I am proposing a three-part plan, or schema, the three major components of which are (1) antecedent conditions or input variables, (2) intervening events or treatments, and (3) outcomes or objectives.

Ideally, no evaluation should be considered complete, and no decision made, until meaningful data from all three sources are available. But, typically, data are uneven or nonexistent. The socio-economic position of parents or their child-rearing attitudes, critical personality characteristics of the children or teachers—these are only some of the many input variables that are difficult at best to obtain. What does the pretty, young teacher down the hall do that causes her pupils to perform better than children from other classes in the same building year after year? And how can we know whether we really changed the behavior of children without following them for the next 25 years? These questions are merely illustrative of the limitation in designating intervening variables and outcome data.

Even though there may be large gaps in the data, the schema is still useful, in that it will help the evaluator and administrator understand the limitations of the data that are on hand. For example, the fact that there are no records for one-third of the pupils in a given class because their parents are migrant workers (a highly relevant piece of datum) should prevent educators in that school from generalizing whatever success they have had with a particular project.

The schema has two other advantages: (1) It should allow evaluators and educators to break out from a myriad of activities smaller bits of data on behavior (performance) or relationships between people and materials, etc., so that they will be better able to focus on those variables which are most likely to be relevant to the program under study, and (2) It should lead to the collection of more and more varied data.

The simplicity of the three-part schema is deceptive. Any one part really must be thought of as a link in an interlocking chain. Since the schema must be kept dynamic, that is, kept as part of a constantly shifting scene, the final outcomes from one schema may serve as antecedent conditions for another—or intermediate objectives for yet a third. How many links in the chain are used in a given situation obviously will depend on the degree of competence of the evaluators and the complexity of activities being studied.

Before leaving the schema, I must point out that it is a fairly common approach and first appeared in print with the New York State Quality Education Project in the middle 1950's. Many writers, of whom Stufflebeam* and McQuire (1967) are excellent representatives, advocate a fourth component—the setting or context. These teams include the identification of the problem and the assumptives that undergird the problem. While I agree with the importance of these factors, I still see them as input variables, as I point out later on. But let's examine each component of the schema in greater detail.

INPUT VARIABLES OR ANTECEDENT CONDITIONS

The importance of input variables has been excellently summarized by Stake (1967):

> What one finds when he examines formal evaluation activities in education today is too little effort to spell out antecedent conditions and classroom transactions (a few of which visitation teams do record) and too little effort to couple them with the various outcomes (a few of which are portrayed by conventional test scores).

Inputs may be defined as all the relevant characteristics that the principals involved in

*See the CIPP Evaluation Model by Daniel I Stufflebeam presented on page 35, "Columbus Report," BIG CITY TITLE I EVALUATION CONFERENCE, 1967 REPORT, Pittsburgh Public Schools.

any educational enterprise under study bring with them at the beginning of the enteiprise. Training, age, experience, level of competence are typical of the antecedent data usually gathered about teachers. But a wise evaluator will attempt to gather data about the interrelationships of teachers–the pecking order for example–and the attitudes of teachers toward "outsiders", even if they come from the downtown, central office.

A frequently overlooked input datum is the information that led to the initial identification of the problem; evaluations have a way of becoming stereotyped. A school board member's child is having difficulty. The result may be the assessment of the performance of a single teacher rather than examination of the performance of all the children with characteristics similar to the single example. Conversely, an irate PTA meeting might lead to evaluation of the public relations program in the district rather than analysis of the educational problems highlighted by the parents. The assumptions that lead to quick interpretations must be questioned early and often.

TREATMENT OR INTERVENING VARIABLES

The data that are to be gathered in this category are from the relevant transactions that have taken place among the students, professionals, media, and materials in the situation under study. The term *relevant* in this case should be considered quite loosely and should depend in large measure on the type of data an evaluator wants to collect.

The role the evaluator plays in specifying the treatments frequently causes problems. Brickell (1961) points out that evaluators require two restrictions: "(1) procedures must not be changed in midstream, otherwise it will not be clear what is being evaluated; and (2) the circumstances in which the procedures are used must be kept compaiable, otherwise it will not be clear what is determining the outcome." In order to move every student toward the desired outcomes, the teacher will often use any promising means, regardless of treatments specified. If a student is not able to comprehend the concept of integers, for example, the teacher will want to find a new text, or a programmed text, or extra help from another student for him.

This conflict has been and remains one of the major conflicts that prevent many schools from conducting more evaluations. I place most of the blame on the evaluators themselves, for reasons I shall discuss in more detail later on. Evaluators should become more ethnographic. That is, they should be more willing to back off and observe what treatment teachers actually employ under given circumstances with given children and attempt to formulate hypotheses or gather other data based on these observations. This somewhat heretical recommendation is made because of the continued inadequate state of knowledge about teaching and learning in formal school settings, and with the realization that most evaluations in local schools are not aimed at–at least should not be aimed at–increasing the woild's knowledge of major educational issues. The concern with control has not provided the resulting trade-off in knowledge. As Gage (1963) states the case, "research in teaching, employing the most commonly accepted paradigm, has been abundant–hundreds of studies yielding thousands of correlation coefficients have been made. In the main, these studies have yielded disappointing results: correlations that are nonsignificant, inconsistent from one study to the next, and usually lacking in psychological and educational meaning."

We must be aware of the difficult position in which we place the local director of research if we do not permit him to tightly structure his designs and controls. Usually he has been prepared for his work by professors of research who continue to serve as gateposts for career opportunities. The reference groups and the professional associations to which he belongs expect him to contribute scientific papers, and they evaluate his work by the standards of research. For these reasons, it is obvious to me that persons who use research skills primarily to improve what is done within a given school system or institution badly need a separate, more clearly identifiable professional group. It should be a group whose members will be just as concerned as their scholarly counterparts with quality and quantification, but who will recognize the difference between gathering data for decisions, which is basically an engineering function, and attempting to add to knowledge, a research function.

There is yet another reason, however, why the teacher's concern for the individual frequently conflicts with the evaluator's concern for structure: individualization of instruction has not been adequately conceptualized in a way that would permit any kind of a systematic evaluation. A new approach to individualizatio.., "mastery learning," may supply this conceptual structure. Bloom (1968) provides a clear description of the term, "teaching fc .nastery." He cites evidence from pilot studies that by individualizing instruction and varying the time allotted for learning tasks, and to individuals, up to 95 percent mastery becomes a goal, it follows logically that evaluators must be expected to evaluate different kinds of treatments than they have in the past. The evaluation problems of mastery learning have not yet been explored in any depth, but the approach, I posit, still holds great promise. The approach should have value both for home-grown courses and curricular innovations and for courses and innovations packaged by outside sources.

OBJECTIVES

The importance of objectives to evaluation has not been challenged since the eight-year study (see Smith and Tyler, 1942). In recent years, objectives have assumed ever greater importance as can be seen in this quote by Gagné (1965):

> For the person who wishes to study the process of education, to analyze it, to perform research upon it for the purpose of understanding and improving it, statements of educational objectives as human performance are an *absolutely essential starting point*. [italics mine].

There is an emerging dissent to the emphasis placed on objectives—particularly the need for behavioral terms. Robert Travers, one of the most persistent critics, bases his objections on two points: (1) There are too many "behaviors" like creativity and abstract reasoning that cannot be specified and measured, and these then are usually underemphasized in educational programs. (2) The behavioral objectives approach rests on the false assumption that children are like "plastic masses" (1968) of raw materials that are inexorably molded into the shape foreseen by the planner. The only conclusion (which will be of little help to the practitioner, I fear) that seems compatible with the points of view represented by Gagné on one hand and Travers on the other is that behavioral objectives should be used for those teaching situations where detailed objectives are possible, remembering that there are "higher" objectives that cannot be treated in the same way.

Another issue that must be highlighted is how to balance locally derived objectives with those established for larger populations. In a private school that sends large numbers of graduates to leading universities, the two may be synonomous. In a slum school it may be unreasonable to expect children to be judged by national norms.

This leads to the separate issue of how to evaluate the objectives that teachers adopt for themselves, or for a school. If a junior high school staff wants all students to develop a written proficiency in Swahili, proficiency defined at a fairly advanced level, the central staff should be very interested to know why. Given this kind of priority, the previous objectives should be reexamined thoroughly to see if the rationale for all other courses can still be justified.

Although I realize I have treated the concept of objectives very superficially, the literature is now so extensive that I prefer to concentrate on aspects of the schema.

THE PROBLEM OF SOPHISTICATION OF DATA

The degree to which any evaluation will help improve decision making will depend to a large extent on how valuable the data are. Essentially there are five sources of data which may vary in the degree of sophistication and the degree to which they are empirically derived. They are: folklore, anecdotes, expert opinion, descriptive data, and research data. Each of these can help an evaluator make certain judgments and administrators make certain decisions, but each has its limitations. For this reason it is important that we examine each source of data.

FOLKLORE

Astin and Panos (1968) call folklore "any widely accepted but empirically untested assumption concerning a causal relationship between an educational program or operation and an educational outcome." We can name many bits of folklore, but a few will suffice. The educational justification for the junior high school is largely composed of folklore about the unique requirements of early adolescents. The belief that children should not be introduced to reading until they are six is folklore (that some teachers surprisingly still believe). Some coaches also swear that intercollegiate athletics produce better citizens.

But folklores are helpful to the decision maker because they are legitimizing beliefs. Folklores enjoy wide consensus. Therefore, the professionals are able to get on the job because they have standards for action.

The single most bothersome problem about folklore in education is that it is so often substituted for empirical evidence. It is easy to see our mistakes in the past, like the reading readiness concept. Bussing students from the slums to more favored schools is a plan born from folklore and rests on untested assumptions. So long as educators and evaluators remember this fact and are willing to question these assumptions when they can, then folklore is a valuable source of information.

ANECDOTES

Anecdotes, especially dramatic ones, can have profound influence on decisions. The student who stabs a teacher and the demonstration for a project that is about to be discontinued for lack of funds, are both examples of events that focus on issues and suggest courses of action. They help call to the attention of extremely busy administrators and school boards events that may not otherwise rise to command their attention.

Since the limitations of anecdotal information should be well known, the only other comment here is to enter a plea for more effective record keeping. Just as the recording of anecdotes for individual pupils can help sketch a profile of his behavior, similar records for an entire school can reveal problems where none are believed to exist. Daily or weekly logs or diaries can serve this purpose.

EXPERT OPINION

The problems of "outside" consultants are well documented and need not be reiterated here. Similar problems are arising with the increased prestige accorded to research specialists on the staffs of all large school systems. Knowledge about research design and methods and statistical treatment and analysis is extremely valuable but is only a means to a larger end. The opinions of these individuals must be weighed against the opinions of many others on the staffs.

DESCRIPTIVE DATA

Descriptive data document what is happening in a school system and provide a systematic inventory of conditions, the incidence, distribution, and, to a certain extent, the relationships of phenomena. (For a more detailed treatment see Van Dalen, 1963, Chapter 10). I am referring to regular collection of data such as attendance records, achievement test scores, broken windows, police arrests, and to *ad hoc* studies such as parental opinions about pending issues, adequacy of school buildings in light of possible population shifts, and percentage of students who have smoked marijuana!

One of the problems associated with descriptive data is deciding where to focus. Schools are admonished to undertake "continuous gathering of data" (AASA, 1959). State departments of education and federal agencies seemingly have insatiable appetites for information. It will do no good to rail against the trend; it is far better to automate so as to stay on top of critical demands. Schools must adopt procedures followed today by many large industries that have their record keeping so automated that only exceptions or deviations from regular reporting are fed to the decision makers. The computers handle everything else.

As valuable as descriptive data are for evaluators and administrators, they frequently are misused. Elaborate statistical treatments, multiple regression analyses, for example, seem to imply causality. (See, for example, R. C. Nichols' review of the Coleman study, *Science*, 154, 1966.) If the records of 10,000 students who have been given reading instruction with female teachers show that five times as many boys have reading problems as girls, while another 10,000 students who had male teachers show just the reverse, the implications for action are clearly suggested. But the cause of the reading problems for either girls or boys still has not been established.

Only one other comment is needed about descriptive data. The results of standardized testing frequently are equated with descriptive data. Much more imagination is needed to provide insights into the inner workings of schools–at the staff and student levels. Much information is available from secondary sources and unobtrusive data sources, two terms that will be discussed later. The number and kinds of books placed on reserve shelves in the library, kinds of equipment and supplies ordered by teachers, the frequency and variety of audio-visual equipment used, are records available to any principal and should help to supplement other pieces of information about the characteristics of teaching taking place in his school.

RESEARCH DATA

The term *research* used in the context of this paper refers to a body of procedures and methodologies borrowed from the social and behavioral sciences that allow the evaluator to gather certain kinds of empirical data that cannot be gathered in any other way. However, the focus is still on the kinds of data that will aid in decision making, and not data that will add to new knowledge. As such it cc..centrates on what Kerlinger (1965) calls the "shorter range goal of finding specific relations." Consequently, the researcher is restricted in the way he can generalize the results.

Astin and Panos (1968) use the criterion of replication to distinguish research procedures from descriptive procedures even though they acknowledge that it is "rarely feasible to replicate the (classical) experimental conditions on any substantial scale." To replication I prefer to add the concept of intervention, that is designing a set of experiences and testing and observing the reactions of the individuals to those experiences.

There is little agreement on the most appropriate procedures to obtain the desired data. In classroom settings it should help administrators to use comparative data whenever possible. That is, the data gathered about the behavior of students or staff should be compared both to a set of absolute standards as well as to data gathered from comparable groups. Research also is the best way to obtain clues about differentiated effects of treatments. Administrators must be concerned with discovering the various reactions of different kinds of students under differing conditions, not just with gross data like group means and correlation between means.

The problems related to the use of experimental and quasi-experimental methods are too complex to take up in this paper. (For one of the most comprehensive overviews see Campbell and Stanley, 1966.) To obtain the most sophisticated data possible, obviously a specialist in research design and methodology must be employed. School systems which do not have these specialists should either hire them or take the responsibility for retraining some of their competent young men and women. Hiring part-time consultants is a very poor substitute.

USING MORE IMAGINATIVE PROCEDURES

School systems should insist that their evaluators develop and use more imaginative procedures for gathering data for decisions and not try to meet criteria from the research community necessary to add to new knowledge. There are five "rules" that should lead to richer data:

1. *Use more than a single measure.* One quote from Webb and others (1966) will support this point.

> So long as we maintain, as social scientists, an approach to comparisons that considers compensating error and converging corroboration from individuality contaminated outcropping, there is no cause for concern. It is only when we naively place faith in a single measure that the massive problems of social research vitiate the validity of our comparisons.

2. *Use more than a single mode of measurement.* The beginning researcher almost inevitably thinks in terms of pencil-and-paper tests and questionnaires. He forgets that it is possible to observe behavior, to use logs and diaries, interviews, projective techniques--to mention only a few possibilities.

3. *Differ treatments among treatment groups.* Again, if the emphasis is on generating meaningful data for decisions, more than one group should be used, and the treatments should be differentiated. Cronbach's appeal (1963) for the use of double blind experiments in educational research has gone largely unheeded. In other words, there should be at least

one experimental group, one control group, and a third group to whom is given some kind of placebo, that is, an alternative treatment that has different ingredients but affords the third group as much attention as the other groups.

4. *Make your instruments and methods sensitive to associated data and possible serendipitous interpretations.* If you want to evaluate changes in behavior among teachers who are sent off to NSF summer institutes, it is important to find out not just if they teach the new materials they were taught, but what materials they no longer use. It is as important to determine the new sources of information they use and, possibly, the individuals with whom they now consult.

5. *Use more secondary data, or unobtrusive measures.* I have borrowed the term *unobtrusive measures* from a book of the same name by Webb, Campbell, Schwartz, and Sechrest (1966). They argue that if you are to use multiple measures, you just find ways of discerning behavior without artificially altering that behavior. For recording behavior they advocate using physical evidence (erosion of tile around certain exhibits in museums and archives), observation (conversation sampling and type of dress), and hidden hardware.

Using these measures opens up many exciting possibilities for evaluators in schools. Consider just one example: A principal has urged his teachers to undertake more joint planning as a way of moving his school toward team teaching. He makes it possible for them to meet during school time, but he would like to know who is really undertaking some joint planning. He can ask them, or give them a questionnaire, but he does not want to interfere, just obtain a clue as to progress.

He assumes that if teachers are going to plan future activities, they will need certain books in advance from the library (cards can be checked). They should request certain pupil records that are kept in the office in order to compare notes on individual students. As the plans unfold, they should order different materials than they have in the past and request permission to use the small auditorium for large-group meetings.

The quantification of many of these variables will be very difficult, if not impossible. However, this should not excuse any evaluator from measuring, as precisely as possible, as much data as possible. He must determine the reliability of his observations and the validity of the attempted improvements. So regardless of the level of statistical sophistication used in the beginning, the tendency should be over any period of time to use ever more sophisticated and complex measurements.

CONCLUDING STATEMENTS

The "let's get at it approach" can be criticized in several ways. It is quite likely to generate and perhaps condone poorly conceived and sloppily executed evaluations as well as provide only evidence of the kind we have used in the past. It is an oversimplified approach to highly complex problems and may tend to focus the educators on superficial aspects of those problems.

There are two ways to compensate for these weaknesses. First develop a broad data base, focus it toward major program decisions and keep it current. Secondly, theories should be used as frameworks and as guides for the collection and interpretation of data whenever possible. Let me elaborate somewhat on the second point.

In spite of the unquestioned value of theory to the sciences, it remains alien to practitioners in education. And with some justification, since theories in our field remain weak (Bloom, 1966). Nevertheless, without theories, investigations will be random and diffused. As Kerlinger (1965) defines theory it is "a set of interrelated constructs (concepts), definitions, and propositions that presents a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting the phenomena."

The importance of theory was reaffirmed recently in the study of the utility of research completed by the Defense Department. In this highly mission-oriented agency it is only natural that they favor research that will lead as directly as possible to technological developments. Project Hindsight, the title of the Department of Defense study, confirmed that "focused research" does have great payoff. But it also pointed out the debt owed to theory. Sherwin and Isenson point up this fact in their summary of the the study in *Science* (1967):

> None of our science Events [their term for discoveries] could have occurred without the use of one or more of the great systematic theories—classical mechanics, thermodynamics, electricity and magnetism, relativity and quantum mechanics. These theories also played an important role in many of the technology Events. If, for example, we were to count the number of times that Newton's laws, Maxwell's equations, or Ohm's law were used in the systems we studied, the frequencies of occurrence would be so high that they would completely overshadow any of the recent Events we identified.

Educators who doubt the value of theories should reread Robert Merton's *Social Theory and Social Structure* (1957). Even theories at the lowest level of generalizability give greater meaning to data and provide insights for improving the studies that follow the present investigation.

But the final argument in support of the "let's get started" approach has nothing at all to do with the technical nature of evaluation or research. It is rooted in the reality that decisions must be made in schools about program matters. They can be based on very limited data and large amounts of experience-based opinions, or just the reverse. The relationship of the evaluator to the decision maker in schools is precisely that described by Etzioni (1968):

> In short, the relationship between the social sciences and a societal decision-maker is not very different from that between the natural sciences and a medical practitioner: Even if either practitioner had mastered all knowledge which the scientific discipline contains, he still would have to interpret, project, and make connections—on the basis of fragmented information and in accordance with the canons of the applied world.

In short the evaluator should concentrate on providing the most valid, the most reliable and the most relevant information *and* the best judgments concerning the information he gathers. After that it becomes the administrator's responsibility to act or not to act as he sees fit. The recognition of this differentiated responsibility should be the final, clinching argument for getting started with more formal evaluations of educational programs.

American Association of School Administrators. *Judging schools with wisdom.* 1959.

Astin, A.W. & Panos, R. J. The evaluation of educational programs. Manuscript in process for the revision to *Educational Measurements*, American Council on Education, Washington, D.C., 1968.

Bloom, B. S. Mastery learning for all. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1968.

Bloom, B.S. Twenty-five years of educational research. *American Educational Research Journal*, 1966, 3, 211-221.

Brickell, H.M. Organizing New York State for educational change. State Education Department, Albany, New York, 1961.

Campbell, D.T. & Stanley, J.C. *Experimental and quasi-experimental designs for research on teaching,* New York: Rand McNally. 1966.

Cronbach, L.J. Course Improvement through evaluation. *Teachers College Record,* 1963, 64, 672-683.

Etzioni, A. How may congress learn? *Science,* 1968, 159, 170-172.

Gage, N.L. Paradigms for research on teaching. 1968, In N.L. Gage (Ed.) *Handbook of Research on Teaching.* New York: Rand McNally, 1963.

Gagné, R.M. Educational objectives and human performance. In *Learning and the Educational Process.* J.D. Krumboltz, (Ed.), New York: Rand McNally, 1965.

Kerlinger, F.N. *Foundations of behavioral research.* New York: Holt, Rinehart, & Winston, 1965.

McGuire, C.A. A proposed model for the evaluation of teaching. *The Evaluation of Teaching,* a report of the Second Pi Lambda Theta Catena, Pi Lambda Theta, Washington, D.C. 1967.

Merton, R.L. *Social theory and social structure.* New York: The Free Press, 1967.

Scriven, M. The methodology of Evaluation. *Perspectives of Curriculum Evaluation,* American Educational Research Association, 1967.

Smith, E.R., & Tyler, R.W. *Appraising and recording student progress.* New York: Harper & Row, 1942.

Stake, R.E. The countenance of educational evaluation. *Teachers College Record,* 1967. 68, 523-540.

Travers, R.W. Evaluator of new instructional procedures. Paper presented at the meeting of the American Educational Research Association, Chicago, 1968.

Tyler, L. & Klein, M.F. Recommendations for curriculum and instructional materials. Report, University of California, Los Angeles, 1967.

Van Dalen, D.B. *Understanding educational research.* New York: McGraw-Hill, 1962.

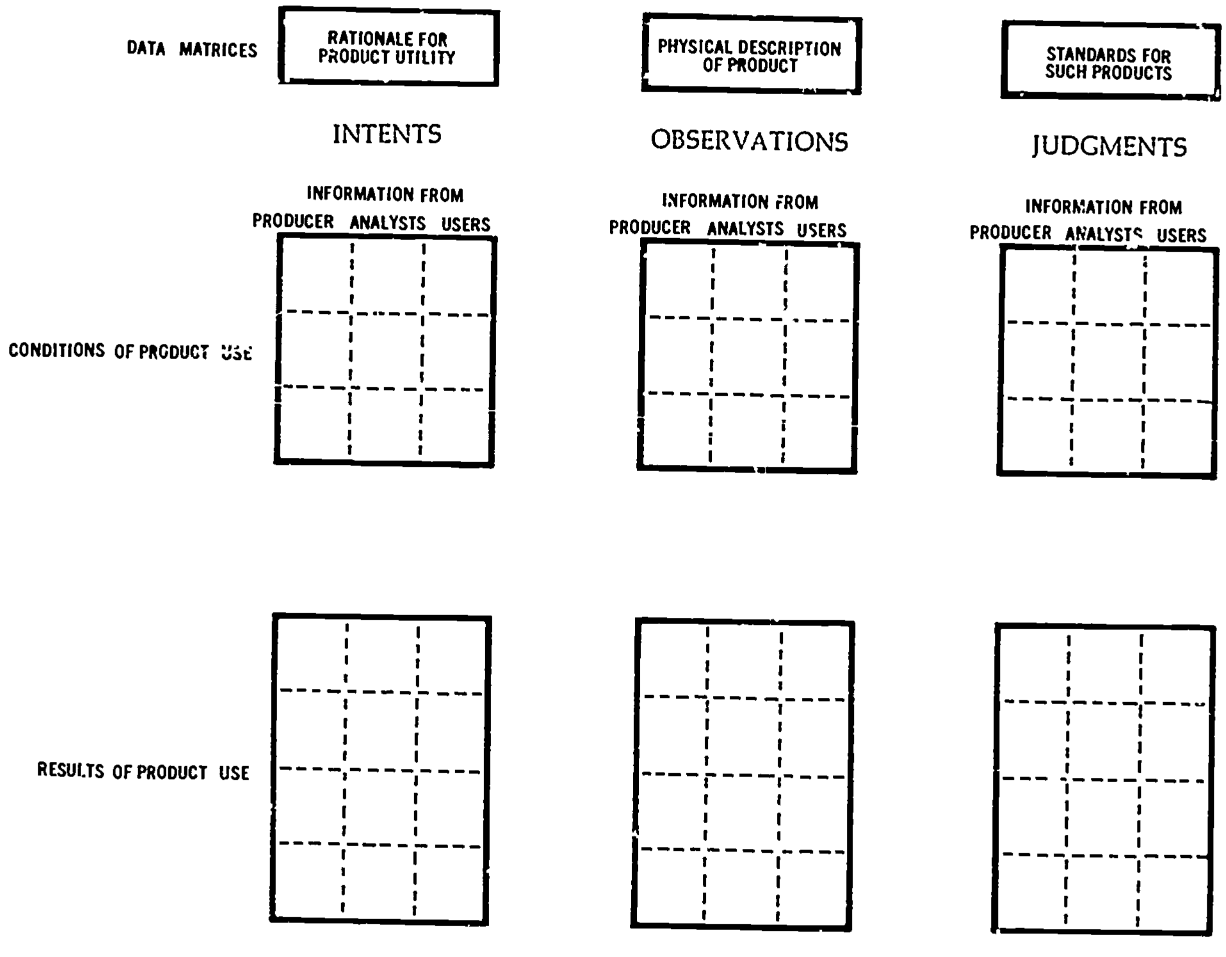
Webb, E.T., Campbell, D.T., Schwartz, R.D. & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences.* New York: Rand McNally, 1966.

The EPIE Approach to Evaluating Instructional Materials

TERRY DENNY, ED.D.
Head, Educational Products Information Exchange (EPIE)
Research Office, Urbana, Illinois

In my dark ages Robert Stake (1967a) began talking about an approach to evaluation in education, an approach designed to reveal the countenance of education. I heard about it and was interested, read about it and was intrigued, thought about it and was possessed. During my pleasant years on Purdue University's educational psychology staff I had been involved in a variety of short- and long-term experiments in the name of evaluation. When

Figure 1



Matrices to guide the collection of evaluation information about an educational product.

the going was rough, I hugged my Linus blanket, the venerable Tyler rationale. Then, too, I had a few AERA-APA model-building papers in a special file marked "do something with these, someday" and had been dimly aware of a few elements of a way of thinking about educational evaluation that made sense to me. It was pretty chaotic. I used to answer "research" instead of "evaluation" whenever a colleague asked me what I was doing . . . no matter what I was doing. Stake's countenance paper clouted me.

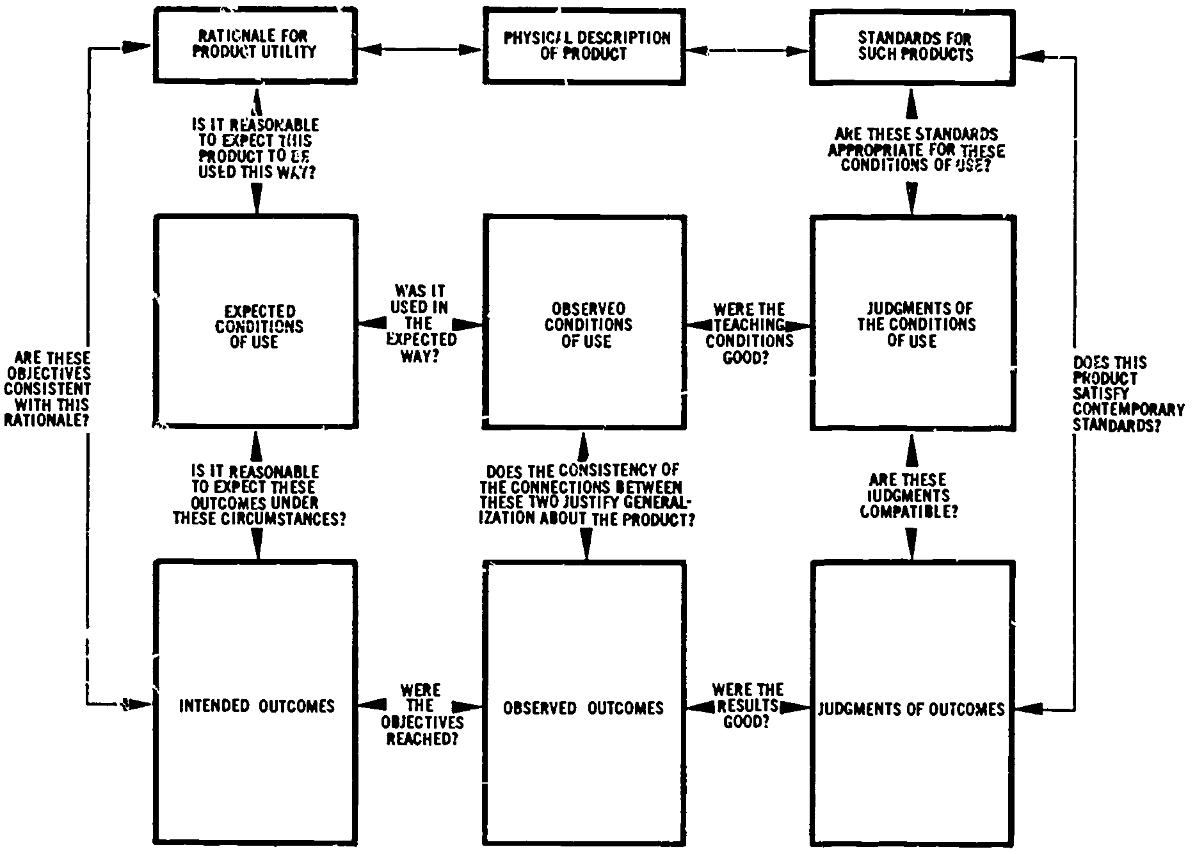
About a year later I began hearing about a systematic attempt to institutionalize the evaluation of the thing-side of education: educational products. P. Kenneth Komoski (1967) was starting the EPIE Institute and had articulated a captivating set of technological propositions about what needed to be done and hov· it might be done. The Institute has evolved an evaluation model principally through the thoughts of Stake and Komoski, and it serves us well. Indeed, I have talked and written so extensively about it I have come to believe I have proprietary rights to discussing it. This conference signals an end to that notion.

EPIE RESEARCH RATIONALE

Another conference paper, the one by Dr. R. Dershimer, shows his understanding and appreciation of the work of Stake and Komoski and, as a consequence, of the EPIE Institute's approach to evaluating educational products. This approach we visualize in the manner found in Figure 1. The complete explication of the relationships within and between data-collection matrices found in Figure 1 has been detailed elsewhere (Stake, 1967b).

To our way of thinking, the elements of any evaluation are bits of information. Each bit is identified according to dimensions or characteristics that help to describe the product. In the EPIE matrix designed to help the local decision-maker evaluate products, each dimension or characteristic is assigned a row. Each source of information is assigned a column. A

Figure 2


RATIONALE FOR PRODUCT UTILITY
PHYSICAL DESCRIPTION OF PRODUCT
STANDARDS FOR SUCH PRODUCTS
IS IT REASONABLE TO EXPECT THIS PRODUCT TO BE USED THIS WAY?
ARE THESE STANDARDS APPROPRIATE FOR THESE CONDITIONS OF USE?
EXPECTED CONDITIONS OF USE
WAS IT USED IN THE EXPECTED WAY?
OBSERVED CONDITIONS OF USE
WERE THE TEACHING CONDITIONS GOOD?
JUDGMENTS OF THE CONDITIONS OF USE
ARE THESE OBJECTIVES CONSISTENT WITH THIS RATIONALE?
DOES THIS PRODUCT SATISFY CONTEMPORARY STANDARDS?
IS IT REASONABLE TO EXPECT THESE OUTCOMES UNDER THESE CIRCUMSTANCES?
DOES THE CONSISTENCY OF THE CONNECTIONS BETWEEN THESE TWO JUSTIFY GENERALIZATION ABOUT THE PRODUCT?
ARE THESE IUDGMENTS COMPATIBLE?
INTENDED OUTCOMES
WERE THE OBJECTIVES REACHED?
OBSERVED OUTCOMES
WERE THE RESULTS GOOD?
JUDGMENTS OF OUTCOMES

bit of information, then, has its own sub-subcell, squared off by row and column, identified by type and source of information. We collect information from the producer, from independent analysts and from the users of the product. Figure 2 is a representation of the processing of product information collected by EPIE.

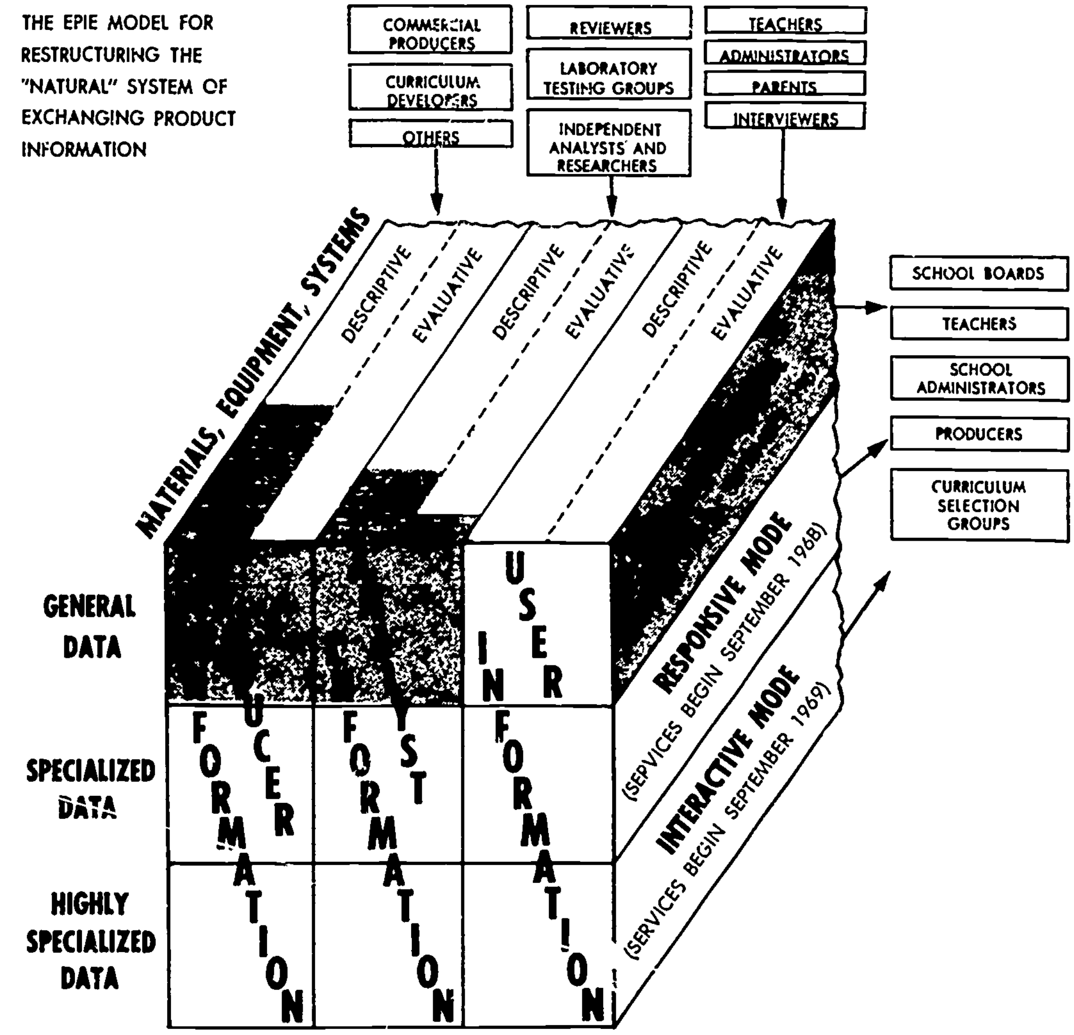
So, how will EPIE make its attack on evaluating instructional materials? Our most extensive search among the data will be for (1) congruence between what was expected of the product and what actually occurred and (2) contingency relationships between outcomes and conditions-of-use which reveal the limits of a product's effectiveness. EPIE relies on researchers and analysts with a broad range of talents and diverse methods of inference to bring about some orderly confluence of data.

TALK ABOUT EVALUATION

Let us assume that the case has been made for the need to evaluate instructional materials. Let us assume also that such evaluation is wanted by several interested publics. The EPIE evaluation model is one such way of doing a job. *It can be done.* We hear, understandably enough, voices of vested interests who say that the educational research and evaluation community has insufficient tools at hand, few skills perfected, no strategies appropriate for beginning the job.

The catch phrases are so appealing: "Researchers can't even agree among themselves; study after study shows no significant differences; the problems under attack are trivial; the methodologies needed to do the job right aren't available." These assertions are appealing

Figure 3



THE EPIE MODEL FOR RESTRUCTURING THE "NATURAL" SYSTEM OF EXCHANGING PRODUCT INFORMATION because if mistakenly elevated to the status of warranted assertion they relieve us of our obligation to evaluate: "It can't be done anyway..." "... and keeps the market the way it 'ought' to be." "The decisions should be made by the producer." That sort of thinking may have served someone well in the past. I don't know very much about how educational materials came to be marketed in the past. The little that I do know suggests that there has been considerable nonsense going on in the name of producers meeting the instructional needs of teachers. We can change the producer-controlled market of doing what comes naturally, to a user-centered market of doing what comes necessarily (Denny, 1967).

The task of evaluating instructional materials on a systematic, comprehensive scale requires a large disinterested group of competent professionals cooperating to gather, analyze, scrutinize and report on their use of educational products. It requires an enormous passel of information, reliable information, synthesized in a defensible public fashion, and a receptivity to its importance and utility by potential decision makers. We conceptualize such a strategic undertaking as shown in Figure 3. We are very excited about its possibilities, impatient with our slow progress, prayerful about necessary funding, and delighted by our reception in a few school systems with which we soon will be working closely to develop our services.

Figure 3 represents the EPIE Institute's program for improving the natural system of information exchange about educational materials. The shaded areas represent information now in the system and being disseminated through our "broadcast" mode: *The EPIE Forum.*

The schoolman knows that out there somewhere is descriptive information about all elementary science series or all secondary school social studies materials, and about all closed-circuit TV equipment or all the overhead projectors now on the market. He also knows that in some cases analytical reviews of materials have been made by competent professionals and that laboratory tests have been conducted on some types of equipment. Further, he knows that other schoolmen have had experience with many of the products among which he must choose. Being aware that this information does exist, he does his best systematically to bring what he can together in order to make a decision. But given the number of products, their increasing complexity, and the time and staff limitations within which he works, his best efforts too frequently end in frustration and failure.

In order to carry out its objectives, the Institute must collect in a systematic fashion the information which it will process and disseminate among educational decision-makers in schools and in industry. The three columns on the face of the model refer to the three types of information (identified by the three sources, Producer, Analyst, User) that are being gathered and will eventually be disseminated by EPIE. On the left side of the model presented in Figure 3 are three levels of information sophistication: General Data, Specialized Data, and Highly Specialized Data. Down the right side of the model are represented the three basic "modes" of information services to be provided by the Institute.*

SOME PROBLEMS

EPIE has been working on some problems in which I think you are interested –problems which confront you in your evaluation projects with the materials of special education. Our primary methodological problem is one of relating goals and conditions-of-use to outcomes. Other problems vie for attention: What are our target populations? To whom shall we ask our questions? What products do they want to know about? What questions

*For a complete discussion of this model see *EPIE Forum,* Vol. 1, No. 5, January, 1968, p. 28-33.

aren't they asking that could be relevant to the solution of their problems? What steps must be taken to assure that classroom conditions are representative?

We have the problem of comparing products with different purposes. No two instructional aids have identical objectives. We ask some questions which show each product in its best light, as well as in light which may be best for its competitors. One product appears better under certain limited conditions, poorer elsewhere. We must aid the potential user to see what objectives, which conditions of use complement which products for his instructional needs.

Still another problem has to do with standards. Most operating standards are idiosyncratic and unconscious, serving to shape personal preferences, perhaps very consistently, but avoiding public exposure. The advocate of this or that standard may adhere to still others in his own practice. EPIE's purpose is not to show what is popular, but to reveal–to expose–the various expectations that exist. Surveying every expectation is all but impossible; utterances both pertinent and suitably documented are hard to find. A thorough presentation of existing standards is a formidable obligation, but a necessary one for a nationwide evaluation project.

Identifying school goals is important, too. Are products differentially useful depending on school goals? Of course they are. Along with other statements of opinion and judgment, school goals have a translation problem. Each goal has implications for practice, but spelling out what practices are consistent and what are inconsistent is not an easy task. How can goals be quantified unambiguously? EPIE is working on the development of more definitive scales for goal priorities.

It is necessary to take the individual teacher's and the individual school's goals into account. The accounting procedure we employ compares the congruence of hopes with realizations. When they are matched, or nearly so, our report to our client might be, "At this point in our deliberations this is the match or mismatch." This is not necessarily good or bad, by the way. People can intend foolish things which fortunately go awry. One also might have quite narrow intentions for a given program which has been observed to have rather far-reaching benefits–another fortuitous mismatch.

An observed lack of congruence between *intended* antecedents, transactions or outcomes and *observed* antecedents, transactions or outcomes may be undesirable, may be merely tolerable, may be quite nice indeed, depending on the context in which the incongruity occurs. When the highly prized aims are not realized in action, it is always an unhappy experience. When unanticipated events transpire which are of some seeming potential but which could be of possible disservice to the program, the findings are indeterminate; and when outcomes exceed our more ambitious goals and are accompanied by grand benefits, the best of all possible mismatches occurs. The judgment depends on the contingencies and congruence throughout the system and not upon the presence or absence of congruence. Stake's model reflects considerable concern for the wishes of the individual instructor. But he says more.

> Consciously and unconsciously, people have different expectations of the products they use.
>
> The EPIE research rationale has no stronger commitment than the commitment to record and to honor this diversity of values. No product evaluation can be complete without a survey of the preferences and priorities of

> the many groups of people who use the product, or who may benefit or be injured by it.
>
> But this commitment does not preclude forthright statements of relative values (Stake, 1967b).

The congruence of local intents and observed events must be judged against external norms. general standards. Evaluation is required. These steps are required in an evaluation study but may be quite entirely absent from experimental research and assessment or status studies.

If such judgmental procedures reveal little commonality for the local school's purposes and the external criteria, difficulties are imminent. Difficulties of the sort experienced by any evaluator who tells a client that which the client would rather not hear. Difficulties of the sort which cause strained evaluator-client relationships. Hopefully evaluators can gain sufficient respect from their clientele to avoid the stigma of *persona non grata.* I have now come to view as vital a relatively unexplored dimension of educational evaluation, namely that of the psychodynamics of advice and information giving. We need some generous help from brethren in the business of clinical counseling.

RESEARCH OR EVALUATION

I would like to turn now to a few problems I have been worrying about of lat
regarding distinctions between evaluation and basic research, between generalizability goals and evaluation goals.

Basic research is much in the news of late. An articulate basic research proponent, Kerlinger, has written, "It just seems to be too hard to understand the nature and purpose of basic research and too easy to talk easily about applied research, research and development, product-oriented research and similar kinds of jazzy things. Besides, one gets money for the latter and not for the former" (1967, p. 5). In his acerbic commentary he charges USOE with hindering the development of basic research in education and commends Wittrock (1967) who wearies of product research "that gets us nowhere." Wittrock calls for support of basic research of educational phenomena, research for conceptual space points, for theoretical experiments in naturalistic settings. But Kerlinger regards the likelihood of getting anyone but a handful of kindred souls to listen to their pleas as unlikely. Bright's (1968) pro-basic research comments are widely printed and discussed; Cronbach's *Kappan* (1966) statement is quoted by everyone and now I have joined them; Gagné (1966) gags on loose molar discussion by educators about curriculum reform, evaluation, revolution; and Bruner's bombshells periodically explode about us in search of a theory of instruction. All of this could be seen as a "movement" of sorts. It is the sort of movement I am not very excited about. Basic research as I understand its potential for evaluation problems or educational decision making is not what EPIE and SEIMC really need. Basic research will of necessity take a long time to come to fruition. The vineyards need to be worked in the meantime. If not basic research, then what? More theory? More information? Yes and no. I am much taken up with the notion that we are about to be glutted with information. We are information rich. We may be headed for a period of superabundance. People are being told the information will soon be at our fingertips.

Unfortunately EPIE may be seen by some to be another member of a growing list of do-gooder agencies committed to the proposition that "what the educational world needs now is data, more data." We dislike this image very much. The educational world *can* do without EPIE, IBM, ERIC, SEIMC tooling up to collect, store, and retrieve still more information. I think we presume too much when we presume the usefulness of data *per se.* I think we presume too much when we assume the rationality of educational decision making. Dershimer has expressed a concern for the amount of folklore in our profession. Others wish to quickly supplant the folklore with empirical evidence. This could be an incorrect ploy for us to take. Kenneth Boulding (1966) suggests that empirical evidence may not always be the better evidence for a variety of human endeavors. The pluralism of the educational enterprise may need folklore, vagueness, alogical custom to provide a sort of elastic bond to keep the melange intact.

It may be that some teachers' behavior deteriorates, when their intuition, folklore and unthinking habits are supplanted or seriously challenged by scientific information. The ingestion of information may not effect or may even produce undesirable results on educational decision making. What a frightening thing if true! I think what all this means in part is we need more data, if, and only if, we develop more and different procedures for information assimilation by our users.

I have no doubts that educators need to exchange information among themselves about instructional tools and techniques. Yet there is apprehension among educators (and certainly among producers) about organized efforts to obtain that information–and with good reason. Stake has put it, "The hazards of prejudice are not less than the hazards of ignorance."

It is obvious to the supporters of EPIE that the need for information justifies the risk of prejudice, i.e., the possibility of encouraging an occasional unwarranted innovation or maintaining some out-dated standards. The risk can be kept small; but the need for evaluative information cannot be made small, for it grows out of the imperative need for rational decision making.

What constitutes a useful distinction between evaluation and research for those concerned with evaluating instructional materials? At the core of the educational researcher's purposes is his aim to generalize his findings beyond the people and setting utilized in his research. He works hard at maximizing the likelihood of similar findings being replicated by other educational researchers doing studies of the same type with other people, other instruments, in still other ecological settings. It's a fierce task. He is deeply concerned with the questions of external validity as well as internal validity sources (Bracht and Glass, 1967). External validity sources may be grouped into two classes: first, those which deal with questions of generalizing findings to other people, or population validity; and second, those which deal with questions of generalizing findings to other settings, ecological validity. The researcher who wants to build instructional theory, to work at curriculum evolution and revolution, to standardize tests and methodologies, is necessarily primarily concerned with population-ecology validity problems.

While the educational evaluator will be aware of these matters and will do what he can to safeguard against invalidity contaminants, the crux of his concern is not that of discovery and building principles or relationships with high generalizability to other people in other settings than those under his investigation. Which is not to say that he cavalierly disregards such concerns–rather, it is a case of priorities, of alternative emphasis. His concern will be with the applicability of extant measurement techniques and regimens to the particular population and setting being evaluated. (Large-scale evaluations must of course deal with large-scale sampling procedures, data handling and statistical inference.) But the evaluator's principal charge remains one of fully describing and fully explaining the program being evaluated so as to facilitate his evaluation of matters related to *that population.*

The evaluator of the local school social studies curriculum is not particularly interested :n whether his findings are generalizable to still other social studies instructional programs across the state or nation. Certainly the local evaluator should not be examining the behavior of a small number of people principally to acquire predictive power to make statements about a large number of other people outside his research sample. When he achieves internal validity through random sampling procedures within his available or accessible research population his validity problem is licked.

The educational researcher on the other hand must do more. He has another inferential leap to make, that of discerning the relationship of his available sample to his target population. What practical difference grows out of this distinction?

It is better for the evaluator to invest his resources in obtaining full, reliable descriptions about a restricted set of circumstances to be judged and to let uncertainty prevail about the generalizability of his findings. This latter posture is achievable only by agencies with huge capital resources, personnel and expertise necessary to assess the large target group.

EPIE will have to conduct some basic investigations and some evaluative studies. Any research study seeks generalization, but studies differ as to the level of generalization they seek. The "basic research" study in education is usually indifferent to personnel, subject matter, locality, and time. The practitioner's inquiry usually calls for minimum generalization, because a purchase to meet some given need is in the offing. But EPIE has many clients. EPIE's studies will specify the product, and search for generalization or limits related to types of pupils, teachers, schools, and so forth.

Why have I labored the point? First, because I wanted to try it out today as a possibly useful distinction for examining the tasks of evaluating educational materials and for training educational researchers and evaluators. Second, it may help those who may be using a consulting evaluator to understand better and to participate more meaningfully in evaluation projects. Third, I hope it encourages discussion and thought about the nature of evaluation, the nature of basic research and the resultant tradeoffs inherent in the conduct of an evaluation study versus a research study.

I am not certain that my notions about the differential function of generalization in experimental and comparative educational research versus situational or local evaluation studies can be or should be defended. There are some forces that could be marshalled. I have argued for theoretical research in naturalistic settings (Denny, 1967). Recall Brunswick's (1956) strong arguments for naturalistic research and consider Cattell's (1966) assertion that the progress of psychology as a science depends increasingly on nonmanipulative designs. Bracht and Glass (1967) encourage me. They claim that while basic research serves a vital function in contributing to our knowledge of the human organism such studies are not the basis for generalization to a variety of situations in which humans normally interact with their environment. Such generalizations are fraught with indeterminable risks. If carefully controlled experimental designs really do not yield generality of findings—or more so than do more local, naturalistic approaches—I feel less constrained to strive hard for more and more basic research at all costs.

QUESTIONS ARE INDICATORS

I would like to have you consider another distinguishing characteristic of the evaluation situation as the attention we must pay to questions such as:

a. How good is this elementary science kit?

b. How can I justify keeping this literature program next year?

c. How helpful was the parental involvement in getting that program in modern math on the road last year, and should I do something similar for my modern biology curriculum revision?

d. The superintendent who asks "which program would be best for my four low-achievement schools insofar as beginning reading materials are concerned: that linguistically based series, the one that is nearly all programmed instruction, or the experimental model that's coming out of the individually prescribed instruction research shop at the university?"

Now, consider some questions which are better answered by education researchers; questions of the type listed by Bracht and Glass (1967) in their treatment of external validity of the ecological variety:

a. Is the treatment equally effective with all teachers?

b. Is the treatment effect independent of the size of group?

c. Is the treatment dependent to some extent on the use of certain audio-visual aids?

d. Is the treatment effect independent of the time of day?

These are important questions. They require careful attention to safeguarding against contaminants of internal and external validity to permit the questioner to generalize to the larger target population he has in mind. They are, clearly, quite different questions from those cited as being better answered by educational evaluation.

There are many common concerns of the evaluator and researcher. Often the questions confronting educational decision makers demand approaches that are one part research to one part evaluation. Consider the social studies research project director who wants to assess the effects of value preferences held by secondary school teachers of the social studies on student concepts of the role of the social studies in upgrading American society when he asks me where to begin his evaluation, ahead are several hours of conversational give and take, coaching, structuring for both of us. But one thing we are likely to settle upon is the need for looking at the interaction of teaching style with personological variables of the learner. We will be concerned with designing that particular facet of the study to reveal the presence of ordinal or disordinal interactions of the sort that Cronbach (1957) called for over ten years ago. Hence, this facet of the evaluation project is likely to be traditional educational research methodology pure and simple, and there will be still other research elements to be included. But when the final selection of what we shall or shall not do comes about, I will be willing to trade off maximal generalizability (or concern for sampling procedures to maximize external validity) in order to complete my picture of the local social studies instructional milieu under investigation. Of course I shall exercise extreme caution to control internal validity, as must the program researcher or the experimental researcher. But there are distinctions to be made.

I have a feeling that there is too much high-powered research motion going on in the name of evaluation. Conversely, in the worst sense of the term, that of defining evaluation as the act of imputing personal preference to matters of choice, there may be too much evaluation going on in the name of research in our schools. But let me end this tangential probe by saying there's not enough research *or* evaluation going on in the name of anything. The documentation for this assertion has been made by others, elsewhere, to my satisfaction, at least, (Scriven, 1967; Kerlinger, 1967; Stufflebeam, 1966; Cronbach, 1966; Stake, 1968).

OBSTACLES

There are school people deeply concerned by educational researchers' unwillingness or inability to listen to the real questions coming from the practitioners—the kinds of questions which should interest educational evaluators, which very likely turn off educational researchers, and . . . alas . . . are being turned down for supportive consideration by a fair number of us as impossible for study. These messy matters are the stuff on which decisions must be made. Some of us are taking all this quite seriously. Stufflebeam (1966) and Hammond (1967) have been listening to such questions, to such cries of the anguished educational decision maker, and have been developing new paradigms for coping with the content and contextual realities of evaluation problems. The research and theory of Guba (1965) encourages me to believe that the tried and untrue old tribal research dance, the pre- and post-test shuffle, can be challenged, improved upon, and where need be, discarded. The critical work of Hastings (1966), Glass (1967, 1968), Stake (1967a) and Webb (1966) and his associates support EPIE's contention that help is coming. Not tomorrow. And if gigantic financial support agencies don't help, not for many days after tomorrow either.

Evaluation is a bad word in several kingdoms. If the word "money" can be translated into "interest", "affection" or "confidence", please check "none of the above" as your answer to the question, "What fascination, love or faith do the money givers have for educational evaluation and evaluators?" Clearly they misunderstand our message; or understand and disvalue it; or understand, value and are about ready to give us a chance to develop and practice our instruments and techniques to increase our usefulness to those who want to evaluate instructional materials.

Could it be that this evaluation conference is the first trickle of a new wave of financial support about to engulf the educational community for the exploration and conduct of evaluation activities? I'm about ready for a good swim.

Boulding, K. *The impact of the social sciences.* New Brunswick: Rutgers University Press, 1966.

Bracht, G.H., and Glass, G.V. *The external validity of experiments:* Research paper No. 3, Laboratory of Educational Research, University of Colorado, 1967.

Bright, R.L. Bright seeks 'basic' educational research. *Educational Researcher,* 1968, 8(2), 1-2.

Brunswik, E. *Perception and the representative design of psychological experiments.* Berkeley: University of California Press, 1956.

Cattell, R.B. Guest editorial: Multivariate behavioral research and the integrative challenge. *Multivariate Behavioral Research,* 1966, 1, 4-23.

Cronbach, L.J. The role of the university in improving education. *Phi Delta Kappan,* 1966, 539-545.

Cronbach, L.J. The two disciplines of scientific psychology. *The American Psychologist,* 1957, 12, 1957, 671-684.

Denny, T. Educational product evaluation. *First Annual Project ARISTOTLE Proceedings,* 1967, in press.

Gagné, R.M. Elementary science: A new scheme of instruction. *Science,* 1966. 151(3706), 49-53.

Glass, G.V., Editor. Research Papers, Laboratory of Educational Research. Boulder: University of Colorado, 1967-date.

Guba, E.G. Methodological strategies for educational change. Paper presented to the Conference on Strategies for Educational Change, Washington, D.C., November, 1965.

Hammond, R.L. A design for local evaluation. *EPIC Forum,* 1967, 1, 3-6.

Hastings, J.T. Curriculum evaluation: The why of the outcomes. *Journal of Educational Measurement,* 1966, 3, 27-32.

Kerlinger, F.N. Letter to the editor, *Educational Psychologist,* 1967, 5(1) (7).

Komoski, P.K. An invitation. *EPIE Forum,* 1967, 1 (1), 2-3.

Scriven, M. The methodology of evaluation. *AERA Monograph Series on Curriculum Evaluation,* No. 1. Chicago: Rand McNally, 1967, 39-89.

Stake, R.E. The countenance of educational evaluation. *Teachers College Record,* Vol. 68, No. 7, April 1967a, 523-40.

Stake, R.E. A research rationale for EPIE. *EPIE Forum,* 1967b, 1, (1), 7-15.

Stake, R.E. Testing in the evaluation of curriculum development. *Review of Educational Research,* 1968, 38 (1).

Stufflebeam, D.L. A depth study of the evaluation requirement. *Theory into Practice,* 1966, 5, 121-133.

Webb, E.J. et. al. *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally, 1966.

Wittrock, M.C. Paradigms in research on instruction. *Educational Psychologist,* 1967, (1) 1-3.

Comments

Dr. James Moss: Would you be prepared to do evaluation if the money were available?

Terry Denny: I don't think there is too much danger of our not being able to tool up in due time. If it were dumped on us today would we be embarrassed by these riches, yes; would we have the cadre to pull it off? I think I'd have to say "no." I think it could be embarrassing. But I participated in creating an undergraduate research training program at Purdue University for a couple of years. We took juniors and seniors, some in education, some not. They were bright – make no mistake about that. They were bright, they had a good track record – I assured myself that I'd get some workers – bright people who were casually interested in matters educational. And we said to them, "What is this world of educational research? What is education doing in American society anyway? How can we answer these questions? How can it be changed and ought it to be?" Very loose, molar questions; and we also bootlegged in about a third of our graduate curriculum in those junior and senior years, without much trouble at all. We gave them descriptive statistics and found ourselves right in the content of a second course in measurement before four months were through, with these bright, aggressive kids. I think that there is one thing we could be doing more of, identifying kids with research interests much earlier, rather than hoping merely to revitalize gas station education for the inservice cadre.

I said "rather than." If one has to "either/or it" – I don't usually like either/or propositions – I think that if we had the kind of money you're talking about we could do that *as well* – the kind of money I *hope* you're talking about. I hope you're talking like Scriven talks, which is a minimum of 10% of allocations for evaluation and research. And so let's talk about a $5 million project, with $500,000 to a million, please, to find out what in the world we are doing. Not $55,000 so I can buy a research director and a couple of graduate assistants and a secretary and send out pretty memos on letterhead stationery.

Yes, I think I know how to train educational product evaluators. (I think there are a lot of people who know better than I how to train them.) It is very difficult to work within institutional constraints at universities who can't and don't want to commit themselves to this important role in teacher education. The problems are in the schools: they just don't have money to assign to this. That's why I'm hoping that the "great-money-givers-out-there" will be able to engage in some kind of a productive dialogue, so we start talking with one another – so they start trying to understand what we are saying. I know some very bright people who have tilled long and hard in the field of educational evaluation who are about ready to give up. And it would be a sad thing indeed if these people were to go back doing their neatly counterbalanced designs, doing their experimental classroom research – because that's where the money is – and desert educational evaluation work.

Mrs. Moss: I have a question both for Dr. Dershimer and Dr. Denny. The basis for this question comes from an experience of going out into the field and talking to people and finding that many teachers feel they don't need materials.

Now both of you developed models today about disseminating evaluation. I would like some suggestions about how you think you might go about establishing some rapport with your target audience. The specific question is: How am I to disseminate evaluation if teachers are not interested in the information?

Dershimer: I don't know. I can think of two ways very quickly. The first is to begin packaging evaluation procedures to give the people with a minimum of knowledge something to work with so that they can have more data than they have now. And secondly I want to come back to the point I touched on earlier, which may not seem to be very appropriate for this group or at this time, but I think it is terribly important: we do not yet have a network of professionals who can support the kind of approaches that Mr. Denny and I are espousing here today, and the evaluator feels very lonesome I'm sure at times, particularly when he is confronted in the university campus with his colleagues. And therefore I am advocating that if it's not a formal, professional association, someone should start working very quickly at pulling these men into a kind of invisible college, or network or whatever term you wish to use. I frankly think that if a given association like AERA does not become more responsive to these demands that these men should very quickly form another group of their own so that they get that kind of support and the opportunity to learn from and support each other. These are the only responses I can think of.

Mrs. Moss: I would like to comment on your first point. I see the problem not in terms of physical dissemination but in terms of rapport which has to do with the possible ultimate effect of the information. I feel that many teachers (and I have no evidence for this) probably have no value about receiving evaluative information, so that we have a real communication problem and we are going to want to do evaluation and we don't have anybody to use the information.

Denny: I would encourage you to write Allan Hartman at the ERIE lab, who did a curriculum materials products survey of users, asking them, "What do you know? What don't you know? What do you want to know more about?" He's thought a lot about this and he may be working on field testing to change users' attitudes, because what he found is just what Mrs. Moss is alluding to: he found, "Things are fine! I don't want any more information." To assume that people are dying to have you dump a ton of data on 'em is a dangerous assumption. And to assume that you can coax them into a readiness position by telling them that they *ought* to be more responsible and more professional and *ought* to be more systematic in their problem-solving, and *ought* to conceive of education more as a rational process, is foolhardy.

Dershimer: This relates to something I wanted to question you about, and comment on further. You are perfectly correct; however, I refer you to some of John Goodlad's experiences at the UCLA lab school, where he launched into a non-graded school that, as you can well imagine, re-shuffled the relationships among the staff. The only system they had was that the children would be moved up the ladder based upon bits of data not previously considered. The existing school enables a teacher to judge children with a minimum of data and even a minimum of judgment, as far as that goes; you know, the only judgments you make are when you have to deal with exceptions, and so if the child is not exceptional in any way you don't have to judge him at all, just move him along. And what he found and is now concentrating on – one of his three major thrusts – is what happens to teachers when you suddenly move them from what I call an experience base for judgments to a data base for judgments. Change the school setting and you will have to change the demand for data and the place of evaluation in their work.

Denny: Good prototype. Highly improbable that that degree of freedom, that kind of power, is in the hands of most school administrators; the people with whom you'll be working. Goodlad, good for him! He has many dollars to construct a totally new milieu, and I believe that he can knock out a lot of walls, do a lot of house cleaning, and even shoot a few, if that were necessary, but I don't see how that is implementable now. I don't see how that provides us with a viable model for changing the ways things are.

Much depends on the extent to which people in an embryonic network can have that kind of freedom, can destroy old role relationships that are getting in the way of functioning in the way they want to -- I don't know what they might be, titles and pecking order – I really hate to open this up because it gets to be awfully messy, you know, sex roles determining who advances further fastest; how they are perceived in their community; their reference groups; what the people with whom you are supposed to be working think you're supposed to do; what your image is. Maybe in a new organization like this, the IMC network, you'll *have* a chance to experiment. I heard you say that you gave considerable freedom to individual centers within the net to try different evaluation models, different dissemination models, and that's good. If you all settle on *one* prototype you'll rise to the stars collectively or all run down in the mire.

Dershimer: I'm sorry. I cannot let you get away with this. For years we have admonished teachers to gather and use anecdotal data, keep them in the files and build them up – to very little avail. I know of very few teachers who will do it. I did not do it when I was a teacher; and the reason is that I was not called upon to use the data. Very little data were required in the decisions I was called upon to make for the children or for what I was teaching. All I'm saying is, the present system we have, that is the self-contained classroom with a teacher who seldom herself is ever questioned about what she does, is not going to change. You must do something in order to alter the system and inevitably you will have to move, if you change the system, to a data-based model. The individually prescribed instruction system is one way of doing it. John Goodlad's non-graded primary is another way of doing it. Any variation in the present system inevitably makes data more critical. I think this is what we've got to aim toward. But we will not have a glut of data for a long, long time; in fact, there will be too little of it, Terry, so this is why I really take exception to the way I think you are cavalierly treating the whole question of data now.

Comment from the audience about teachers' needs for help in decision-making. The person remarked upon a communications gap between teachers and researchers-evaluators and said this was aggravated by the latters' use of jargon, which was ascribed partially to the need for ego-gratification.

Denny: You said that teachers with whom you work want some help in making those decisions – they don't want a lot of esoteric argot. I play two roles here: I'm making general evaluation comments, and I'm also talking about the EPIE research rationale. EPIE will never *make* those decisions for the teachers or even get close to making them for them. What we're going to do is alert them as to what their choices are: give them the evidence that exists in relationship to how materials have worked with other teachers, with other kids under various conditions of use, report the producer's field testing analysis, make-ready activities, and we'll encourage and employ independent analysts to do analyses of these materials. But the decision ultimately lies with the user. I hope he'll never abdicate that. I hope nobody is going to make decisions for him.

We're finding more and more there isn't any one way to analyze instructional materials. We're looking at elementary science materials extensively and we have six analysts' groups identified with quite different points of view. Let us pretend like you want to make a decision about purchase. When we talk you through these six analytical approaches and the findings of analysis, you may say to me, "But which one is appropriate for me?", and I'll say, "That's for you to decide." "But won't you order them according to priorities? Will you give me a best buy? Checkrate it for me?" "No. Here are your alternatives. Now what is it *you* want again?" Now that might be the opening dialogue. We can help people come to see what they are about, what they value. And then, if they choose one analytical group's approach over another, that's their choice. But then they will do so with some foreknowledge as to how these materials have been analyzed, how they have been developed, where they have worked and haven't worked, and under what conditions of use: we regard that as a service.

Now let me say a word about jargon: I think if the teachers (about whom we are talking) had known and understood that you were in fact going to talk about techniques "one through five" you could have told them in two minutes with "esoteric argot" like: "I now refer to the VanRostov technique and the Dulson technique and you all know the Husenberg technique, and we have found that there is an interaction..." and sat down. It's a tremendous savings, you know, when you and I know what interaction effects are, and I don't have to talk three or four pages about it, do I? But if I do, then I'd darn well better learn how to do it, to coach you. Some of you know Robert Stake's evaluation model. Now I'm not going to have to talk you through antecedents-transactions-outcomes by congruence-contingency relationships. I'll say, "Do you remember Stake's model?" That's jargon: model; Stake is a reference somewhere. But sometimes I think that which you would ascribe to a need for ego-enhancement may be a shortcut to aid communication. (I know it's also the former as well as the latter.)

Finally, there are very few of us, by the way, who have the kind of talent that Mrs. Moss is calling for. Tom Hastings and Robert Stake at CIRCE, University of Illinois have it. Terry Denny doesn't. I think Mr. Dershimer may. I'm not talking about talking beneath teachers, administrators, or down to them; he can reconceptualize these highfaluting concepts into practical language. I think that is a very useful talent.

Dershimer: We're known as troubadours.

Evaluation Procedures in Montgomery County Schools

by RICHARD L. DARLING, PH.D.
Director, Department of Instructional Materials
Montgomery County Schools, Rockville, Maryland

A scant six years ago in the Montgomery County Public Schools there were no procedures for evaluation of instructional materials. Perhaps this is an extreme statement, but it is true in the sense that there was no organized and systematic program to identify the best materials to support the county curriculum. No doubt individual schools, within the limitations of their staffs, made a valiant effort. But even then, the number of items of instructional materials appearing on the market made it impossible for a single teacher or even a single school to examine and evaluate ail of the potentially useful materials available.

In 1 '62 a committee called the Superintendent's Advisory Committee on Instructional Materials recommended that a program be initiated. Here are portions of their report.

> Instructional materials are those items which are designed to impart information to the learner in the teaching-learning process. A wide variety of instructional materials is essential for the best instructioı.. Instructional materials may be consumable or expendable but are generally non-consumable and fairly durable, such as books of various sorts, charts, films, filmstrips, globes, maps, models, magazines, newspapers, pictures, recordings, program materials, slides, specimens, transparencies, workbooks, etc.
>
> Second, the most important objective in all evaluation procedures is to locate and make available for teachers and pupils the most suitable materials that can be found in the various subject areas.
>
> Third, materials should be evaluated by those who are to use them. Group evaluations are best when they are based upon the actual experience of using the materials in a teaching-learning situation. Instructional materials should be considered in terms of the total curriculum and should be closely coordinated with curriculum revision to assure current and suitable materials.
>
> Fourth, general criteria to be applied when evaluating all types of instructional materials are: appropriateness to a particular curriculum, appropriateness to a particular grade level, authenticity, contribution to learning, quality, good value in terms of purchase price. The above criteria are also to be

applied when evaluating free or inexpensive materials. In addition, the following items must also be considered when evaluating instructional materials: characteristics of students relative to their interests, attitudes, experiences, knowledge and skills, the learning pattern which will be most beneficial for the students, the total curriculum of the Montgomery County Public Schools and the specific educational goals of the Montgomery County Public Schools.

As a result of this report, the review and evaluation program was set up within the department of instructional materials, which included other instructional materials operations as well: supervision of the instructional materials programs within the individual schools of the district; a county central instructional materials center, largely a center for inventory of 16 millimeter films and other materials appropriate for central office holdings; a processing center to process library books and other instructional materials; and production facilities for producing new materials.

The first step after the decision had been made to create this service was to establish firm criteria for the evaluation of materials. As a result of a cooperative effort among the departments of supervision and curriculum development, the department of instructional materials, teachers from the classrooms, principals and others, the criteria and procedures were established and published in a document called *"Review and Evaluation Procedures for Textbooks and Instructional Materials."* This document has gone through several revisions and is distributed in each edition to every teacher, librarian and other educator within the school system.The basis of our program, taking the principles set down by the original advisory committee, is that group evaluation is better than individual evaluation, and as a result, committees were set up to do the bulk of the evaluating of the various instructional materials. There are approximately sixty committees. They are organized by subject and by level.

For example, there is an elementary mathematics evaluation committee, a secondary mathematics evaluation committee; there is a kindergarten through third grade social studies committee, a grade four through six social studies committee. On the secondary level there are committees which devote their efforts to evaluation of materials in specific subjects: a geography committee, an American history committee, and so on.

These committees are made up largely of teachers who volunteer for this service. The Division of Review and Evaluation secures volunteers by public announcement and by asking supervisors, principals, and area directors within the system to recommend teachers to serve on these committees. The majority of the members of every committee must be classroom based or school based teachers and the chairman, who is elected by the members, must be a school based teacher. Others who serve on the committees are supervisors in the appropriate subject, librarians, counselors, and some administrators.

The function of the committees is to review and evaluate all of the materials of instruction in the subject at the level at which they work. From their evaluation activities decisions are made as to what materials shall be purchased for the central inventory and/or what materials will be recommended for school purchase. They meet monthly or more frequently, and a schedule is posted on a regular basis in the Division of Review and Evaluation. When the committees come in to evaluate materials they find the materials waiting for them and the necessary equipment there so that they may devote their total effort to the professional task of evaluating materials. Some materials are evaluated on the spot, others, particularly printed materials, are taken away to be evaluated individually, but the committees come back together to discuss the materials and to arrive at final conclusions concerning them.

Textbooks and programmed materials require six signatures to be approved; al: others require only three and do not require the signature of the subject supervisor. In addition to the volunteer program which I have just described, (and these committees meet ordinarily after the end of classes on school days) we have a summer program. In Montgomery County Schools 25% of all the professional staff work on a twelve month assignment. During the summer months we have full time committees working on the bulk of the subject areas, devoting either full or part time to the evaluation of materials through the summer. This program in a sense is not voluntary, because the teachers are being paid to devote their full efforts to the evaluation of materials.

There is one exception which I must mention to this committee procedure: this relates to trade books. The volume of publication of trade books is so great in any one year that the committees in most subject areas, though not all, are unable to handle the full volume of publications. Therefore, with trade books, we provide for individual evaluation by individuals within the total professional staff of the system, and even provide a combination where one or two professional evaluations from within the system may be coupled with one or two professional evaluations from journals to provide us with the three approving signatures that we require.

This brings me to the Division of Review and Evaluation itself. Having said that the evaluation is done by committees and other individuals in the professional staff of the system, I have indicated that the staff of the Division of Review and Evaluation does not, in itself, do the reviewing. On the contrary, this staff is much too small to hope to encompass the reviewing job to be done. Instead, it is a coordinating staff responsible for the logistics operations in connection with the entire review and evaluation procedure.

A major job of the Division staff is securing materials. They must bring in the materials which committees and individuals will evaluate. They have been averaging, for the last four years, thirty thousand items of instructional materials per year. They come in through various ways: automatic samplings from textbook and trade publishers and from some producers of audio-visual materials; the staff may request materials which have been requested in turn by members of the staff of the school system; any member of the professional staff of Montgomery County Public Schools may ask that we bring materials in for evaluation. These requests are given top priority in the entire process. The other way in which the staff gets materials is by meeting with vendors, discussing with them new products, and requesting directly that their new materials be sent in to be evaluated by our staff.

Another function of the Division of Review and Evaluation is coordinating the committee work. I guess I left out the first one: it is organizing the committees. They bring materials before the committees, they schedule the meetings of the committees, and for some types of materials, they make an attempt to bring all the materials of that type together at one time. A good example of that is the map and globe evaluation, which is a one-shot activity in review done just once a year. They bring seventy to one hundred maps and globes together, and that, I believe, is something that no teacher and no school could possibly do for itself.

Still another function of the Division of Review and Evaluation is the dissemination of its information. This is done, of course, on an informal basis as teachers and librarians come to the Review and Evaluation Division or phone in or write memorandums requesting information; but it is done in a systematic way through the issuing of lists. The division issues four lists annually: the map and globe list (which reflects the activity performed in that once a year evaluation of maps and globes); the library book lists (which are issued once a year, and with supplements throughout the school year); elementary and secondary textbook lists (which are issued once a year and include thirteen or more titles for every subject at every grade level, so that teachers may select from a rich variety of approved texts); and the fourth list is the list of other instructional materials.

Everything goes onto these four lists from the evaluation process except those items which are for central inventory only. Sixteen millimeter films, for example, are purchased only by the Instructional Materials Center (IMC), another division of the Department of Instructional Materials. The Division of Review and Evaluation simply forwards those evaluations to the Instructional Materials Center where the staff in turn orders for central inventory.

Still another dissemination activity of the Division of Review and Evaluation is the creation and maintenance of an examination collection, so that teachers, librarians, and supervisors who have received the lists of approved materials may actually examine the materials before they select for their own collections. This examination collection includes a collection of textbooks and library books and a collection of other instructional materials of the non-print type. Every librarian in the school system is provided a minimum of one half day per month to come to the examination center to select materials to take back, to coordinate the evaluation of trade books, or to examine approved materials for selection for their own IMC.

The lists are used in the schools as ordering tools, but they are used in connection with the examination center. I am sure that there are things purchased blindly, but more and more, teachers and librarians, with their lists in hand, come to the Division of Review and Evaluation so that they can examine materials.

Now this must sound like a fairly rigid program. Let me remind you that we are talking about the 22nd largest school district in the United States and therefore, one where when we approve an item and every school decides to buy it, we are already talking about a lot of money. It sounds rigid, but there are things that keep it from being totally rigid. One of them is a tryout procedure so that new and different instructional materials may be used with children and evaluated with children before they are placed on approved lists. With this procedure, if the school requests permission to try out an item not on the lists, it goes through certain channels; to the Division of Review and Evaluation to make certain that it is not already on an approved list or has not already been rejected for use, and to the appropriate subject supervisor. Then the school is permitted to try it out. They make a written evaluation, which is in turn given to the appropriate committee, which evaluates the materials after the try-out has been completed.

The other exception to our regular procedure is nothing more than an exception. When teachers or principals identify needs with particular children or particular groups of children that cannot be met by materials on approved lists, and they justify the use of some other material, my office has the authority to grant them permission to buy materials not on the approved list.

This is our program, basically, of review and evaluation. I would like to say a little about what its effects have been. One of the things I can say for certain is that by having teachers and supervisors evaluate those things which we purchase for our central inventory, the films and other materials which schools cannot afford to own themselves, we have made sure that we are ordering the materials which teachers want and which teachers will use.

We have had an enormous annual increase in the use of centrally inventoried materials. For the last four years circulation of our 16 millimeter film collections has gone up 25% annually, while the size of our collection has been increasing by 6% to. 10%. Our research department has revealed some interesting statistics in connection with this too; I may not have the exact percentage right, but there is someone here who can correct me if I am wrong. In 1961 we knew that with 16 millimeter motion picture films, approximately 9% of classroom time was going to the use of these materials. This fall in one limited group, at least, of the teachers, this had increased to a little over 20%.

Another way in which we have been able to find a difference as a result of our review and evaluation program is that the curriculum and instructional materials are more closely related to one another. This is true not only in the classroom and within the school, but in terms of the central office departments of Curriculum and Instructional Materials as well.

The program has developed closer staff relationships. Teachers and librarians work more closely together. Teachers and supervisors, librarians and supervisors, work more closely together. The result is a far better staff relationship.

Still another effect is that far more of our staff have a direct say in the selection of materials than would be possible if central office supervisors and other so-called experts were to do all the selecting. With more than 10% of the total professional staff of the school district involved in the evaluation of materials, teachers have a major role in the decision as to what will be used. I believe I can truly say that this program has enhanced the prestige of the teacher in our system. The teachers are the people who decide what materials will be used. They are the ones, basically, who decide which ones ought to be tried out with children before a final decision is made.

With the issuance of lists of materials that have been pre-sifted, with those things which are poor or which are unrelated to our program of studies sifted out, our teachers, librarians and administrators can devote their time and efforts to more important aspects of selection. They can examine materials that have been pre-sorted and can pick those more appropriate to support the program of the school. We can be sure, also, that better materials are selected for every school. No longer is it a hit and miss effort but a professional selection from the best materials available for the purpose.

Finally, because these committees examine a great variety of instructional materials, ranging from the text book to the trade book through the whole variety and range of non-print materials in auditory and visual forms, our teachers are better able to find those materials which best suit the needs of the children. In other words, materials are brought to bear on the individualization of instruction.

These are a few of the advantages we feel we have gained. We still must improve the program in a number of ways. We must speed it up; I have avoided telling you how long it takes, but we must speed it up because it now takes too long. We must cover a far larger number of materials than we now do, although it is only fair to the staff of the Division of Review and Evaluation to say that the total of the four lists comes to well over 75,000 individual studies of instructional materials.

We must develop methods which will permit us to involve the children even more than we do in our present method. Now, we use the children when we have materials that are new and different; we think also it would pay if we could try out materials that are more traditional with children, before we make final decisions as to whether or not we should use them. We must devise procedures and secure staff for a more systematic coverage of all materials. Now, we depend upon the samplings provided by producers. We depend upon the requests of teachers and others, and we depend on what the vendors who come in tell us.

What we need to have is a staff adequate to make a systematic search for materials in every subject of the curriculum and at every level for which we need materials. We must learn how to disseminate more information to our schools. We issue lists now, but the lists are bare-bones lists. They give information on title, author where appropriate, publisher or vendor, price, whether it is black and white, in color, and a few additional things. But we need not only to provide lists, and lists with annotations, but actually to present materials to teachers. We are looking forward in the future, with the development of a television distribution system, to actually showing materials to our teachers, giving them opportunities to hear and see, so that teachers and librarians know every item through this kind of dissemination. There are copyright problems involved, of course, but I think they are insoluble since the intention here is not to substitute for purchase but help schools to learn how to purchase more wisely.

Another dissemination activity to which we must give more attention in the future, though it has certainly not been totally neglected in the past, is in-service education about instructional materials. We must help teachers and librarians learn how to select more effectively, to evaluate the materials they have, to select from those available, those they need to have, and to identify those tasks for which no materials exist, so they will know better which materials must be created in the school situation locally.

There are many defects in our program now, but we are identifying them and we are developing plans for their improvement. The millenium, I suspect, will never come, but the program will continue to get better day by day.

Comments

Question: Do you have special education committees?

Darling: Yes. We have special education committees and every list includes special education. This is supplement one – it looks like the whole thing is special education. It consists of 13 pages which are computer print-outs, and beginning on page 7 and running through to the end there are materials which our special education committees approved. They include materials in speech and hearing. There are such things as film strips, kits, workbooks, transparencies, tape recordings, flat pictures and so on – all identified by our evaluation committees for special education.

Comment: Montgomery County seems to be very well endowed. There are less fortunate counties. Perhaps taking advantage of the strength of this group and realizing of course that there are some cultural differences that might have to be translated in terms of other counties: could any of this be disseminated to us?

Darling: Let me say one thing about cultural differences. Our deaf kids are just as deaf as yours and our blind kids are just as blind as yours and our retarded kids are just as retarded as yours, and therefore when we come to our special ed. list I don't think that you can say there are cultural differences because you know they aren't relatively deaf in relation to our community. Where it might be true in our regular curriculum areas that there would be differences, I think when it comes to special ed. this wouldn't apply. Our evaluations don't go as deep as they ought to. These materials haven't all been used with children, and you have to recognize this defect. But we do print deliberately several hundred extra copies of our basic annual list every year, and they can be purchased at $5 a volume from the Division of Supply Management of Montgomery County Public Schools. They require that you send a check in advance because they want to avoid bookkeeping problems. This represents a little more than the editing of and preparation of copy, printing and binding costs. We're in a sense giving away the bulk of the professional activity that goes in. It sounds a little steep at $5 a volume but there are 250 to 300 pages in every list.

Comment by Montgomery County staff member: I would like to point out that the subject areas are in alphabetical order so that special ed. would always follow social studies alphabetically. Then the next breakdown after subject area – this is already done by level. We have an elementary one. The first breakdown is level, the second within the book is subject area alphabetically and the next breakdown is kind of material. At the back there is a vendor index alphabetically by vendor.

Question: (inaudible)

Darling: Well, our relationship with most of the commercial vendors is very, very good indeed. The staff of the Review and Evaluation Division must talk to 300 or 400 representatives of the companies each year and at least two thirds of them don't leave the building without stopping in my office, so we have lots of direct face-to-face discussion in which we talk about things we see as our needs and they talk about the things that they have, and I know they take back from us some ideas to their companies as suggestions for new materials that need to be produced. Occasionally, though not too frequently, they ask us to cooperate with them in trying out some materials which they have not yet put on the market. I think I mentioned to someone at lunch today that the 3M-*Newsweek* "News Focus" was tried out in one of our schools before it was marketed. Another thing that was tried out in our schools before it was put on the market was a new McGraw-Hill globe with transparencies for teaching geography which was tried out several months before it went on the market. Now, to come to the last part of your question – what about the people whose materials are turned down? Well, most of them are people who represent companies producing a lot of materials and they know that first of all we look at our curriculum and are looking for materials that suit it, and that we may turn down some things not because their quality is poor, but because they are remote from the level at which we try to teach those concepts or provide that information. We aren't always saying that they have a bad product and they understand that some things may be turned down for these reasons. Occasionally somebody is exceedingly unhappy. I have a letter on my desk right now from the representative of one company who says he is going to sue me, which I am waiting to see.

Comment: (inaudible)

Darling: ... You mean the one, two, and three ratings? Is that what you are referring to? Let me put it on here – [transparency]. This is the instrument, but this volume, which runs to 37 pages, lists in detail what these criteria stand for and describes the gradation. One is top-notch materials, two is average, and three is unsatisfactory; and they look at the eight categories, and on the basis of what they said in all eight, they decide on "approved" and "disapproved." Although, obviously, if something is number three in authenticity, a one in interest would hardly make up for that – this is a contradiction... I was going to say false facts.

Comment: (inaudible)

Darling: We do need to have this kind of evaluation; the one thing we do is to ask teachers to provide information and feedback so that we can re-evaluate items which they think are no longer appropriate, items which they may judge inappropriate even though they may have recently gotten on the list. And this is something we are exceedingly conscious of and should have always been more conscious of, and that is materials which are offensive to any one of several minority groups within our community.

Comment: (inaudible)

Darling: Right. There are separate criteria; that's correct. As you go through this booklet you will find out that there is a detailed listing for charts, globes, display specimens, library books and textbooks, films, filmstrips, slides and transparencies, tape and phonograph recordings, flat pictures, maps, workbooks, programmed materials; and then, and something I didn't dwell on at all, but which is equally important: criteria for the evaluation of audio equipment, visual equipment and audio-visual equipment, flannel boards, magnetic boards and so on; evaluation of screens, evaluation of teaching machines. Since we really look not just at the media program but at all instructional equipment, we have criteria for the evaluation of power tools, and so forth.

Mrs. Jett (Montgomery County): A comment about different groups. Montgomery County has a cross section, I think, which would match any area within the country, so that we are serving as many disadvantaged children in proportion as any other area. It is representative for all groups and all levels. My other comment is in answer to the questions dealing with vendors. On our form there is a space for comments and these are available to anyone who needs them, as well as to vendors. In working with the vendors I have found that they welcome comments of any kind that we have, as Dr. Darling mentioned, suggesting changes that might be made in materials we found were not good; and most of the vendors have asked when a material was rejected, why, and we have been able to say: "We have rejected it for this reason – or this" – and they have welcomed this and have gone back and actually made a change in the materials if there were wide enough use for it.

Comment: To this point there has been a lot of talk about a vendor taking the materials back for changes: as this would cost a lot of money to them is it ever possible for the vendor to come and present his ideas before the company actually produces the item and get your evaluation of the possibility of such an item's usefulness?

Darling: Well, you see this is what has happened with "News Focus." I think that most producers and publishers have had such a huge background of experience in publishing that they aren't inclined to do it. It's when they have a very new and different product that they want to try out. "News Focus" was a new kind of thing – where they were taking the current data from *Newsweek* and issuing it in – as you know – transparency master form, and they needed a different kind of reaction on this. We would encourage it. I mean any time someone says: "Will you help us to see if this material is appropriate?", we're happy to work out the arrangements for them.

Comment: Is there any other school district that utilizes your evaluation?

Darling: Well, not exactly a school district. But, under a contract with the Maryland State Department of Education we purchase and lend all of the materials that go to all the private schools of the State of Maryland under Title II, so they are using them. The state department buys from us our approved lists, issues them to the three hundred-odd private and parochial schools, and they use our lists as the basis for selecting those things which they're going to borrow under Title II.

Comment: There are something like 20,000 school districts in the country; how come there isn't more acceptance on the part of the county next door of [approved lists].

Darling: Well, I can only say this – that one well-known bibliographic publisher asked for copies which we gave to them. They were considering these lists – that is not the library or textbook lists but the other instructional materials list, which is the area in which there is the least information available, and they decided that the coding we used was too parochial – aimed at the course of study that we offer, although it's the state course of study in the secondary schools, and that our way of arranging was not appropriate. But we do – as I have said – print several hundred extra copies, and many places that know about them have purchased them, but we haven't made any attempt to advertise them. We do the job for ourselves and we don't want to pretend that we can speak for what everybody needs to have. If other schools examine our lists and find them useful we are delighted to share.

Comment: (inaudible)

Darling: ...I think that with 20,000 school districts in the country and all but fifteen or twenty of them smaller than we are, and many of them much, much smaller, vendors are not going to send materials for evaluation to smaller places. The reason we can get 30,000 items to look at is that the vendors know that those things which are approved they can count on to bring a sizable sale. I do think that if anyone wants our lists that we have a responsibility to share with them. We do sort out on a kind of absolute basis the no-good things that nobody ought to waste their time looking at. We don't say that what is on our list is all necessarily right for your needs, with a particular child in a particular place and time, but I think we do say that among those we have sifted out are some that you ought to take a look at if you can, because we have only discarded them because they don't fit our program. But there are some that nobody ought to waste their time looking at.

Comment: There was some discussion about whether teachers are or are not interested in evaluation; obviously only a small percentage of the teachers who volunteer to be your evaluators are necessarily interested. Do you have any feed-back from other teachers as to how they feel about evaluation in general; how many requests do you get from teachers?

Darling: We get more requests than we can feed back information properly on, but it is true that there are a large number who don't send in requests. I think more of them are really interested in having this evaluation done than the requests show. The gripes, I think, should be added onto the requests to get an understanding of how much teacher feeling there is about evaluation. There are those who feel that they know everything – 'I ought to be able to go out with a blank check in my hand and buy whatever I want, regardless' – and then there are many teachers who devote the utmost of professional skill and time to the process.

My own feeling is that every teacher ought to have free time paid by the school board in which they have time to go in and look at materials, and I think that the best way they can do it within a given school is for the faculty to come together, or, in the department in a secondary school where they are more likely to be organized into departments, for teachers in a subject area to come to an examination center together and actually loc.. at materials. There is nothing better than seeing them themselves, but they can't all see them. That's why we have used the device of having the librarians come to examine materials. They can take some materials back. With the library books they actually get teachers involved in reviewing them.

Comment (Montgomery County school librarian): *re* the first comment you made – the first reaction I had was dealing with teachers – I coordinate libraries with schools (library services for special ed. schools) – that one thing is that there is just not enough time. I try to bring things to teachers, but that just isn't possible; but we do have the committee for the trainables and the educables, and those with learning disabilities. We don't have children at home that we have to run to: we can take time to carry back what we have to the teachers – for which they are most grateful. Particularly in special education itself there isn't a teacher there who wouldn't love to have the opportunity to examine the materials.

The Consumers Union Model

by MR. MORRIS KAPLAN
Technical Director
Consumers Union, Mount Vernon, New York

Although it is not necessary to have a fine film in order to communicate one's ideas, given equal cost and equal energy one would prefer to find the best material on the market. If there were some way that the consumer could find out which item was the best, this would be an improvement over possibly having to choose an inferior item due to lack of information on the available items. It is to this problem that I would like to address myself.

There are characteristics of a good microscope, a good projector or a good tape recorder that are almost independent of the use to which it will be put. No matter how you use it, you want it to be safe. No matter how you use it you would like it to be durable, or convenient to use. And so there are always a host of questions, and they are at a most elementary level. They are not really at all in the area of which we have been talking. If you recognize that, I think we can proceed.

Let me tell about how Consumers Union attempted to answer the question that was raised here earlier – how to go about setting up an agency or process for evaluating products or processes. Who's to do this? Should it be the government in some form or other? Should it be the producers in some form or other? Or how? The solution which Consumers Union arrived at was to do it through a cooperative of the users. If there are 10,000 school districts and if they have any interest at all in products or process evaluation, then perhaps each of them would be prepared to contribute some small sum of money, assign the appropriate investigators to investigate these questions and distribute the information throughout the 10,000 school districts. This in general, is the approach Consumers Union used.

Some thirty-odd years ago a number of people became concerned about the poor quality of products which people were faced with and the choices that they had to make, and how difficult they were. They had the same set of questions I have just mentioned now and to answer them they chose this approach: to establish an independent, non-profit union of consumers. This has certain kinds of implications, and it suggests a single-mindedness of purpose: to provide consumers information on consumer goods and services, independent of a profit motive which could make profit-making more important than this purpose itself.

The fact that we chose a membership or union kind of organization made it a cooperative effort. Members elected a Board of Directors who had no financial interests in the products or any connection with producers or distributors of goods. They were independent. Generally, this Board of Directors consists of scientists and engineers and educators, social workers, and so on. The members of the organization also express their desires through an annual questionnaire on the kinds of things they are interested in. The members pay for it all thrc ıgh their subscription fees and through the purchase of special publications, and although non-members may subscribe, anyone :an become a member merely by electing to do so.

Most important was the notion of being independent. That meant that Consumers Union would accept no advertising; it would accept no gifts or donations, or subsidies or grants from any commercial source. It would accept no samples for tests; all of these are bought in the open market. It does not permit the commercial use of Consumers Union test results or Consumers Union's name. It does not permit more than ten copies of any single issue to be purchased by anyone except for educational or non-profit use.

If you are interested in these concepts, then, and find that they might be transferrable to your problem, you will recognize that we start out with subscribers, or members, or users of the service. They elect a Board of Directors who are non-paid and independent. They in turn appoint a director who is the Executive Officer. He hires staff and gets the facilities together. The staff produces a product which is distributed in the form of a published monthly magazine for use of the subscribers.

I will review briefly the general approach that we take to evaluating products. One of the questions we concern ourselves with is, "What products should we evaluate?" For this we set up the Operations Committee which consists of the heads of the departments involved in getting out the reports. There is a marketing group which purchases the samples; there is a library, or information group which feeds in that kind of input; there is the heart of the organization which is the technical department that does the technical product evaluations and then there is the editorial department which concerns itself with distributing this information in a usable form.

The input to the Operations Committee consists of information from an annual questionnaire which we send out to subscribers and from the voluminous correspondence that we get. The magnitude of this correspondence demonstrates the interest that people have in this kind of thing, and I expect that if there were a similar organization in the educational field the response would be similar. There would be lots of people who would be concerned with questions such as: What projector shall I buy? Which will be the easiest to maintain? Which will be the most effective?

We also subscribe to the trade press and read all the trade literature and have lots of similar inputs. The committee decides what projects to undertake and when to publish for timeliness and interest and proper balance of the issues, so that our magazine will be readable.

One of the next problems we concern ourselves with is deciding which brands and models to include. It is rare that we cover the whole field. The Marketing Department determines what products are the most widely available, which have the greatest interest for other reasons. For example, a product may be of interest because of its unusual claims or low cost, or other special features which might be important. This information is put together with a discussion of marketing practices, seasonality of sales, type of outlets through which the products are sold, delivery and installation problems, warranty practices, price information and so on. It assembles manufacturers' specification data. On the basis of its studies it recommends a scope for a project; the scope may be narrow and include a particular narrow category of products or it may be broad and include all kinds of products.

One of the most important questions involves deciding what to test and how to test. I was most impressed with the earlier presentation that discussed this problem and we worry about the same kinds of questions: Are we testing for the right things? Do we know the ambience in which this product is going to be used? And, in the context of that ambience are we evaluating the product properly? But, we have learned over a great many years that even if it is not possible to answer those questions well, there are so many other important questions that one can answer, it is clearly worth the candle to go through this operation and not be hung up by the things we cannot do. We concern ourselves with the things we can do. We can learn a lot by lining up twenty or thirty or forty products, one next to the other. And even if we know nothing more about these products than what we can see, there would be clearly demonstrable differences, sufficiently important to warrant the dissemination of this information. Now we, of course, do have a great deal of technological know-how that has been developed over the years to provide even better answers.

If someone were to press me in terms of the discussion earlier this morning – Do we know the process well enough? Do we understand all the uses to which these products are going to be put? Does our evaluation answer the question in terms of all of these processes, and wouldn't it be different if we were to answer it for other processes? I really would have great difficulty in answering. But I would argue that what we do is so valuable that one need not be disheartened by the difficulties of the problems that one can raise about such evaluations. If one merely wades in and starts he will find that there is a great deal of extremely useful information that can be obtained and that will help immensely with the decision-making.

What to test for then, and how to test? In terms of consumer products, our library information source supplies the standards and specifications that have been developed for such products all over the world; and if you look at them you realize how elementary, how inadequate, they are. It supplies publications of other studies on the subject, as well as other, more general, references. An engineer, chemist or technologist to whom the project is assigned writes to sources such as the manufacturers themselves, to other test laboratories which concern themselves with this problem, and to others, for information on criteria and methods. He may initiate a questionnaire to ascertain what users consider important. Asking the consumer is a useful kind of exercise, but one ought to know what its serious limitations are.

The sophisticated tester will devise test procedures of his own. There is, of course, the danger of the tyranny of the tester. One of the things he does is test for criteria he knows about. He tests by methods he knows how to use. These are not necessarily the ones that are responsive to the need, but he himself is a captive of the limitations of his craft, and unless he is fully aware of this (and it takes a sophisticated tester to be aware of this) he can mislead you and himself grossly.

Most of the methods in the field have been developed by industry, by the manufacturers, and are industry-oriented. I daresay this is more true in the area of educational materials than it is in the area of consumer goods. The characteristics we try to think about we classify under five headings, and they may be useful to you.

The main heading we refer to is *performance*, and this has to do with how well the product does the job for which it is designed. If it is a projector, we are interested in how sharp the picture is and characteristics of that sort.

There is a secondary category of characteristics that we refer to as *convenience* characteristics, and those are: How easy is it to get the product to perform what it is supposed to do? Are the knobs easily accessible? Are they hard to turn? Is it easy to load and unload?

Then there are a set of considerations under the heading of *safety*. These have to do with safety for the user, safety for the product that is put into the machine, like the films, safety of the machine itself, and safety of its environment; the table on which it sits, for example.

Another set of characteristics has to do with the *durability* of the product – that is, how long it is going to last, how trouble free the operation will be, etc.

And finally, we have to deal with *economic factors*. How costly is it to buy, to operate and to maintain?

These five categories do not include such things as style or appearance, which are also important considerations, but ones we think need to be left to the individual user of the product.

Having gone through this exercise of deciding what to test for and how to test, the project starts. Samples are ordered. We have shoppers, a large number of them around the country, who buy the samples for us at retail and send these samples to our laboratory. This concept is important, it seems to me. Accepting samples from manufacturers is not good practice, for many reasons. For example you may get a sample which is not representative of the product you are interested in evaluating. Also, there is a kind of obligation that you assume when you accept products from manufacturers that you do not want to be under.

Samples are bought and sent to the laboratories. In the laboratories testing equipment is procured, set up, and de-bugged, and then the testing process starts. The samples are subjected to many tests and the data is accumulated. We use a variety of types of testing that may also be interesting to you. We may use laboratory instruments of one sort or another. With that, as with all testing, there is the problem of validation – how well does a set of test conditions simulate what actually goes on in use? These are not easy questions, more difficult for some products than for others.

We have small panel "use" tests, which again have all kinds of theoretical problems, and are subject to all kinds of criticisms. Yet in practice they turn out to be extremely informative. You can learn a lot about a product by having even a small, unsophisticated group of people use it. The defects which show up even in this simple screening are amazing. Sometimes, one wonders whether the manufacturer even had his product tested by anyone before it went on the market.

Finally, there are elaborate field trials; another technique for evaluating products. These are the most useful, I think, if the field panel is adequately chosen and the test design is appropriate. They are also the most costly, the most complicated, and the most time-consuming, but the easiest to interpret because they involve real life situations.

Then there is the problem of having accumulated a vast amount of data on different characteristics, (sometimes as many as forty or fifty of them). There are twenty or thirty products which one has. And there is a mass of data which needs to be added up some way. It is not enough to merely present all of this information about all of these products, because most users would be confused by all of this. It is not possible to look at twenty times fifty bits of data and try to make some sense out of this. It is rare that one of these twenty products turns out to be good in all respects.

The relevance of data involves weighing of the factors, and it becomes most important to know how the product is going to be used. Some assumptions must be made. We have often been told, and if you are going to attempt to use this approach in solving your problems you will be told by all the producers and by the people who are opposed to this model of product evaluation, that each user is unique and therefore it is not possible to devise a scoring scheme, a way of integrating this data, except in terms of the uniqueness of the individual user. This is a lot of nonsense. Again it is the same kind of sophistication, over-sophistication that is way beyond the level at which we are dealing with these problems. *Nobody* wants a product that's unsafe. *Nobody* wants a product that will fall apart after a week of use. Everyone wants a product that performs well even at the most elementary level. The light must go on when you press the button. Now if you limit yourself to the most elementary kind of evaluation, you will still find clear and important distinctions. You will be able to reduce these twenty brands to three or four or five that have the virtues you are looking for; the others will all fall by the wayside for one or another of the kinds of deficiencies I have talked about that are so serious that no matter how the product is going to be used it will be deficient. So it is possible, I think, on a crude level, but still a very useful one, to devise rating schemes, scoring schemes of various types and a number of similar approaches.

Next we go through an elaborate process involving checking of the data. It is essential if you are to retain the credibility of the whole operation that you check the data. Credibility is a very important word in this union of consumers. If there is any question about your doing your work well, if there is any question about your lack of bias, your independence, then the whole operation falls by the wayside. What has made Consumers Union as unique and as successful as it is, is that it has been possible over the years to maintain this independence and reliability.

Finally, there is the writing of a technical report which is re-written by our editorial department; nothing gets published unless the engineers who have done the work are satisfied with the accuracy, tone, and emphasis of the published report. This is very important. This is not what happens in industry, even where industry does do its own product evaluations. The advertising or sales people are the ones usually who have the final word about what goes out to the user.

This, in very rough outline, reviews the kinds of things that we do. Our evaluations are comparative rather than absolute. This makes life a good deal easier, and for most purposes, that is the real life situation – the choice is among the existing products and you might end up saying that none of them is any good and that therefore one ought not to buy any of them. But this is beyond the kinds of questions I am raising.

Actually, it is rare that all products are unsafe, although we have had situations of that sort, and you might decide, if you know that, not to buy any of them and find some other way to achieve the objective. But the usual situation is that you are going to buy a projector anyway; you just have to have a projector, and the problem is which of the twenty or thirty or forty you are going to buy. Therefore, comparative evaluations are the most meaningful and useful ones.

Comments

Comment: It certainly would be interesting to me to know how an organization like this [Consumers Union] met the early problems they faced and overcame. I have a question on the ratings in *Consumer Reports:* acceptable–not acceptable. What has been the reaction from the producers of the materials?

Kaplan: Well, the reaction is either one of great joy or one of great pain, depending on whether your product was rated well or poorly. In terms of law suits, in the 22 years that I have been associated with Consumers Union there have been three or four law suits and we have done roughly 60 or 70 projects per year, each project averaging 25 products or so. None of them ever reached the courtroom. Most of the time after the evidence was presented the producer withdrew the suit.

Comment: Do producers ever come and ask for help before they produce something?

Kaplan: They always do and they are always turned down. It is a matter of philosophy with Consumers Union not to devote its energies to telling a producer how to make his product better. If he can deduce it from the information we have supplied, o.k., we'll make this information available. They do come to us; we talk to manufacturers all the time and they ask about methodology which, curiously enough, they often don't know.

Producers of vacuum cleaners don't have a method. Producers of washing machines don't have a method for determining how well the machine washes. There just is no accepted method; they've been working on it for 15 years. It is still not accepted in the industry – and the reason is very clear. The marginal producer doesn't want the method accepted because it will show his product to be poor. So, they often come to us at least for methodology which we usually make available. As for test data, we will make details of data available as it applies to their own product, but we will not tell them what we think is a good product or how to do it – except insofar as they can deduce it from the published report. One of the reasons is that we want the option of changing our minds, because we learn as we go along. And I would be very unhappy if I had to publish a report of a kind I did 20 years ago; I think we know a good deal more about it now. Also, I think producers can use other techniques for determining how to do this kind of thing; they don't need us to tell them how; they are far richer.

Comment: What did you run into in the first few years?

Kaplan: I wasn't there during the first ten years of C.U.'s existence, but I am told that people worked for love. We tested only products that were very inexpensive to buy, like milk or hosiery – we didn't buy washing machines or automobiles. We even found ways to talk about integration and fascism in the early days. But there were lots of problems; however, there was a felt need and it was not too difficult to get 100,000 subscribers after a relatively short time. To get up to 100,000,000 is another matter.

The original emphasis was supposed to be for people who were low income. It turned out that this was not the market for our kind of activity. It turned out that the techniques for evaluation, the cultural pattern of people – poor people – isn't to go to written literature as a source of information. They are not used to thinking in terms of research to solve a problem. Also, they don't read, and so we found that our market was the middle class – unfortunately the upper middle class.

Comment: Do you have a manual of evaluation for producers that might like to learn more about the evaluation process?

Kaplan: No, but they can always learn how we tested mattresses this time. It might be quite different the next time. And we would be glad to tell them.

Comment: With the variability that you find in so many of the commercial products, have you run into this – where you might buy an item and it is either very good or very bad, and then you get some feedback from your own people, and then...

Kaplan: Yes, it has happened. We try to avoid this wherever possible by proper sampling techniques and by proper interpretation of the data. This involves a kind of sophistication that again some of the fancy evaluating scientists would think is crude, and how could we possibly dare to do this kind of thing? But we dare, because it is a real life problem and we think that what we do on balance helps rather than hinders. But it doesn't guarantee. There are techniques – and they're kind of complicated – for trying to assure yourself that the product you're testing is not unique, that it is reasonably representative of what one ought to expect of the whole class of products. You sometimes make mistakes and when you do you have misled readers, but our record has been quite good on this.

The thing that makes this whole scheme possible is the uniformity resulting from mass production. If each producer made a single product you couldn't use a scheme like this because what you learn about that product wouldn't apply to anything else. But when the manufacturer has a run of 100,000 products then it is possible to find techniques for buying those, evaluating those and saying something about the other 99,995.

I think the thing one needs to learn about this is, that we do two things: we don't list the raw data alone and say: "You are on your own to interpret this." We try to list as much of the information as we can which might modify the conclusions in terms of your own needs. The report itself lists both kinds of things in the hope that you will read it carefully and make your own judgments, but you at least have some guidelines from us as to the things we consider important and how we decided to integrate this in terms of our typical user or users.

The most importan' thing I think is, one needs to be fearless about this kind of thing. It is an important need. The most significant defenses from C.U.'s point of view are the defenses of truth and fair comment. The laws of libel in this country are such that any statement about an individual or a product, however damaging, is immune from liability as long as the statement is true and as long as the comments made in connection with the statement are within the bounds of what the law would call "fair comment."

It is critically important, obviously, to maintain the most painstaking checking and rechecking of facts and findings so that the defense of truth can be securely relied on and sustained if necessary in court. Of equal importance is a certain measure of restraint and non-extravagance in the comments which are made based on these findings, so that a reasonable man under the circumstances, in a reasonable court, would agree that it was fair comment.

Comment: At this point it is clear that you want to make it clear to everyone that you are absolutely unconnected with producers. Now in the educational materials field I wonder how this applies; whether a group of evaluators should maintain an independence or...

Kaplan: As I said, you'll have to do the translation yourself, to your own needs. I can only tell you that the principle I consider important is the one that leaves you completely independent of the producer, in every possible way. You must establish your credibility to the users of your materials. If your material is going to go out and the people who get it are aware of the fact that you have worked with the producer – that he has supplied you with the product, and perhaps money for its evaluation – you will be suspect. At least, you would by me. That has nothing to do with your integrity or competence; you'd be suspect even if you didn't want to be – even if you tried as hard as you could to be objective. There is always the problem of – well, if I'm not careful – subconsciously – I won't get the second grant. And this is a problem with grants; it is a problem with any kind of funds you get from other sources. If you can be independent this whole question doesn't arise – in your mind or in the minds of the people you are serving. Now, how to achieve this in a particular area for you I don't know how to describe. I've thought for years that there was room for this sort of thing in the educational field. I'm pleased that EPIE started. I think they are moving very slowly, cautiously, as behooves people in the educational field and I think if I were doing it I would be more rash, and perhaps fall on my face sooner, but if you move too slowly there is always a risk it will take you too long to build up the momentum and capture the interest. On the other hand if you go too fast you might run way ahead of your users – so you have to know your own field. In the case of Consumers Union there was no problem – we just started, and the key thing was that we rated things by brand and model – by name. This was unique; nobody had ever done this before. And we did it on an independent basis – we were beholden to nobody.

Comment: I can certainly understand your insistence that an organization like yours or like EPIE's should purchase your products that you are going to evaluate, but I'm not sure that would apply to a large school system. We annually have a bid on educational equipment. One of the requirements of the bid specification is that the manufacturer must supply for testing purposes a copy of the exact model on which he is bidding, which we will, through a series of tests not as rigorous as yours, but rigorous for a school system, come up with a choice. I don't think we are suspect by the 160 some schools which we disseminate.

Kaplan: No, I should like to make a distinction, and again, your situation is a little different. One could make a case for a model kind of organization which accepts samples from producers if you establish a set of rules about how you accept them. If you get them from everybody, for example, and you *always* get them from everybody, and if you choose them in a way which assures their representativeness, one could justify this kind of scheme. In the case of bidding there is a problem: you could make this a condition of your bidding. But if you tried to evaluate all products on the market or just a large portion of them for which the producer is not interested in your evaluation – he wasn't interested in your buying it – he may have wanted the other school systems to buy it but he doesn't care about yours (maybe because you test too well) – you would be in a somewhat different situation. So there are circumstances under which this could work.

Comment: I would like to also comment on your remark that even a fairly unsophisticated committee could make a start, because they can discover very quickly when a switch is located in the least handy place which even an expert may have failed to see.

Kaplan: I may have over-emphasized all of that. It is possible to find the machine or gadget which has none of the deficiencies: it is safe, it seems to perform and so on, and which has no educational value – it's just wrong. That kind of evaluation is quite a different one – you may end up buying the wrongest one although it will last a long time and be cheap and effective, and so on. I'm not commenting on that kind of evaluation, but as to the other kinds, you're absolutely right and I can prove this out of twenty-some-odd years of experience. It would be very easy for most products for you to make this kind of elementary evaluation simply by getting all the products, lining them up and using them for some period of time under some sort of controlled, intelligent conditions. Now this is the beginning of an evaluation, and you can learn a great deal – far more than you would suspect – in terms of the effort involved.

*Proposed Management Model for Evaluation Data**

by M. H. MOSS, ED.D.
Associate Director, Mid-Atlantic Region SEIMC

The purpose of this presentation is to describe a management model for evaluation (see Fig. 1) that evolved from my fumbling with some of the problems of trying to see the issues and develop procedures for evaluating materials. I feel that it offers to the IMC's certain advantages for implementing some data collection procedures. The emphasis of my presentation is upon personnel and information sources regarding the collection and management of evaluative data.

Management Model for Evaluation

Type of Data	*Level*	*Data Source*	*Information obtained includes:*
Analytical	1	non-expert staff	physical characteristics including adequacy of directions or instructions
Analytical	2	expert staff, consultants	accuracy, scope, and sequencing; learning, motivational and other characteristics
Empirical	3	classroom use: teachers, pupils,	teacher-pupil and environmental characteristics and teaching and learning style interactions
Experimental	4	formal research: in-house, other on-going, research literature	1. Comparison of material A under various conditions and/or with a variety of applications 2. Material A compared to material B for process x. 3. Using material A with material B under a variety of conditions.
All of the Above	5	Industry	Rationale; population on which material was developed; conditions under which material was used; related research

(Figure 1)

*Transcription of a speech presented following the conference on Saturday, April 6, edited subsequently.

The evaluation components of the model I am going to present are not unique. Its value, if any, lies in the fact that it is management oriented. The basis for the model is the question "What are the sources of information we have for evaluating a material (or instructional system)?" I've looked at the data-collecting procedures we have developed at the IMC. I've looked at the questions we have asked. And it seems that we have five principal sources of information. These are as follows: a) non-expert staff; b) "expert" staff and consultants; c) teachers, pupils, "observers", consultants; d) research (in-house; other on-going; literature); and e) industry (see Figure 1). Each source of information provides a slightly different kind of information, with some overlap, most notably in the area of information obtained from industry. The first four sources vary in terms of the validity of the evaluative information obtained. This validity ranges from face, through content, to construct validity. The prime validity in terms of the pupil is that of "learning efficiency" and short term and long range "learning effectiveness."

The first source of information is analysis of the physical characteristics of the material, including adequacy of instructions for use of the material.[1] One can train almost any mentally competent person to rate physical characteristics for specific materials: for example, in terms of books, you want information about readability (i.e., print size, format), illustrations, or whatever. These are very physical, gross characteristics. In the case of instructions, the reviewer must see if he is able to easily and without confusion carry out the manufacturer's directions for using the material, playing the game, etc.

The next level of analytical information is the content. This analysis involves someone who is knowledgeable in the field which that material represents, and most materials represent several content fields. It may involve a language expert, an expert in human growth and development, and so forth. These experts have to look at the materials in a context, e.g., in terms of "the child" with respect to a subject matter or a task. They have to look at it analytically, relative to sequencing, appropriateness of level, motivation, and the like. They are able to suggest other uses for the materials. So that is the role of the consultant or the "expert" – some of whom are on the IMC staff, some of whom are brought in.

The third source of information is analysis of the material in use in the classroom. This involves the teacher variables, the child variables, the environmental variables – their interaction with each other and with the use of material. At the simplest information-of-use level it might only account for the teacher's reaction to or opinion of the material, hopefully in the context of classroom use with students. Many of the forms we have developed in the IMC network have to do with this: e.g., the teacher rating forms would fall into this category.

The fourth level is "research." All of these labels are difficult; I don't like any of them, I guess, but this is about the worst. I've tried calling it empirical, but this conflicts with category three – and so forth. This is experimental research I'm talking about. I feel that research has a certain image in our minds so that we can use that label, keeping in mind that I'm referring to formal, experimental research. Media research is a very broad field. A thorough design would include an assessment of the teacher's and pupil's reactions and attitudes.

The fifth source of information I see is industry. (If some of you are concerned that these are not exclusive categories of information, you're right.) Information from industry may be used at this level, by raters. They look at the manual, see if it has certain kinds of

[1] After the initial build-up of library materials this phase can be eliminated. The "expert" can do these tasks concurrently with content assessment.

data in it, things like this. Ideally a statement of rationale is included along with a description of the population with which the material was developed and any other relevant research data.

EVALUATION OF A SPECIFIC MATERIAL

Now, as an example, if you take a language kit, you can have someone go through it and see if it has what it says it has. This would be "internal evaluation" and categorically it would fall under information-source level I. You can see if it has adequate instructions. If it is something the teacher is going to have to do, does it have instructions, are they clear, are they pleasing, or whatever. This is a matter of physical content analysis and not a matter of the educational content. What are the physical characteristics? Are the materials durable – to the extent that we're concerned with durability? Are the materials that a child is expected to handle of an appropriate size? The rater can be trained to recognize physical appropriateness in the context of particular kinds and ages of children.

At level 2 (content, meaning, substance) you would need to have a language development person comment on whether or not the kit can be expected to achieve the specified goals. You would need a special ed. person to see if those goals are appropriate, and also perhaps to translate the information into the specific curriculum, and so forth.

At the third level you could do this: get your classroom data through observation checklists, teacher's rating, or observation of pupil's reactions. You can get information from an observer. It depends on the rating form; maybe it'll be the same as the teacher's.

Looking at it most superficially: Language Kit X provides an example of what information you can get from industry: you have the purpose and rationale stated explicitly. You have the population described on which it was developed, and quite a bit of related research mentioned in the manual, or the set of instructions, that goes with the material.

Beyond that, of course, you can research it by comparing it with another language development kit, by comparing it with another system which involves materials to achieve the same goal, or you can try it out in a variety of conditions, including use in conjunction with other media. You can keep the material constant and vary the conditions.

ISSUES

It seemed that within the first four levels of the model I'm presenting, the following issues must be dealt with: personnel, validity and reliability of data, data collection instruments, criterion measures, and inservice training. Regarding personnel, who's going to do it, what staff, which teachers? The typical teacher, the master teacher, the volunteer? Look at all the kinds of error these things involve. Who's going to do it?

In terms of over-all reliability of ratings – how many raters do you have using what forms? In terms of validity, the questions you are asking raise questions about the use and dissemination of the information collected. How are you going to use it? How are you going to disseminate it? How are you even going to store it? These questions are related to matters of personnel recruitment, selection, and training.

Rating forms, questionnaires and checklists need to be developed for all of these, including the research. There may already be appropriate "criterion measures," (tests or other data collection instruments) for a particular piece of planned research, but sometimes there are not. You need to have checklists, you need to have rating forms for the classroom observers. And then, ultimately, you may have to code the data for computer analyses.

And you also need to provide inservice training. I was interested to read in one of the newsletters that they had inservice training for their abstractors. Well, if you're going to get reliable and valid data: inservice. How much inservice is involved? These are important questions.

Priorities have to be evolved: priorities of materials to be evaluated, priorities of techniques. One reason why I'm sort of partial to this way of organizing (i.e., the management model) is that I think you can take a piece of material and go through the various levels of data sources. For example, on the forms you can have as one of the items, "recommendations for further evaluation." It may have to be filled in by someone other than the rater. "Should it be tested and evaluated further?" "Can it be evaluated?" From our practical experience with the library acquisition materials, I have found that there are some things that come back to my desk with the answer that someone doesn't know whether the physical properties are adequate. In other words, they're asking for something relative to the concept of what they see the material as. This particularly applies to things that we're using for reasons which come under the folklore approach (see Dershimer, p. 12) that was presented yesterday, things that we feel are traditional and are good. And so we aren't being as discriminating with them as we are with others. The 'expert' sees it: does it then need to go on for further information? Can you say: "Can it stop here?" Maybe it can stop more easily at one of the earlier levels than another material can. For a lot of materials you may have to say, "I don't know if it is any good until I 'research' it." In terms of what was said this morning, in many cases the final answer may or may not rest at an earlier level. I feel it does in many cases. I think that you're left with so many other questions unanswered, even with the research data, that it may not be worth the time and effort to get results because of some of the criterion measures you'll have to qualify and cannot really use, but in some cases it certainly is very desirable. What things should be brought to any particular level: the things that are the most widely used? The things that are the newest? Those which will cost the most? Those which are economical in terms of the teacher's time?

EVALUATION BY "BAD" CHARACTERISTICS

I mentioned computer data forms and at each data source level the question of recommendations for further evaluation. I also want to mention the fact that no one has talked about "bad" information. One could evaluate by setting up a "bad" rating scale and if something had so many characteristics that were "bad" you would throw it out. You may want to know "bad" characteristics. These may be very important for you to look for. I don't know. It's something that had occurred to me.

EVALUATION PROCESS DISSEMINATION

One question a participant mentioned to me was that after you go through most of the models and a teacher has the material in the field, does she need to have a built-in test so that *she* can be evaluating her results when she gets a different group of kids. Should we just disseminate evaluation results? Should we also disseminate process and tools? Do we need to take specific steps such as inservice training to create awareness in the consumer, or can this be achieved without such direct effort?

INDUSTRY AND THE NETWORK

In terms of industry, this isn't quite parallel, I warn you. I feel from some of the experiences I have had that what I would like to do is start developing some real dialogue with industry – to start out by saying, "This is the information we need": communicating with them and having them want to communicate with us at a much broader level than we have, I'm sure, or than we have had an opportunity to. I would also like to see us (perhaps) come out with a "catalogue of acceptable materials" for the network. Is this desirable? This is one of the steps in getting to the point – which is the ultimate hope for me – where we say to industry: "You either provide us with such-and-such or we don't buy your product. These are our standards." I see that there must be lots of intervening steps, partly to give industry a sort of sign-of-the-times as they move along. If we have an "accepted list," how are we going to handle it if we have different "levels" of evaluative information? Are :ve going to have different kinds of criteria for these materials or are we only going to include those we do at a particular level, or what? Those are all questions.

OVERVIEW

I propose that the components of this model can be dealt with initially independently. When I thought of the conference, I thought that we might have a workshop on evaluation, or something like that. So I was trying to think of how one could set up some structure to get started – at least to get off the ground and be productive the first day rather than the second – and the second day as well. Most of the ways for running a workshop that came to me at the time depended upon hierarchy. Somebody would have had to be working on objectives while someone else was working on priorities, and this sort of thing – each step depending on a prior one. I rather like the framework of the management model, and I have gone into it partly because I feel this can provide an organizational scheme for independent IMC collaboration. A couple of centers have been saying, "Well, we've been doing this, and we've got some stuff." Great. Well, work out some criteria. Work out the problem of personnel as well and the other problems I haven't even envisaged.

A number of centers are working on teacher ratings, and Bob McIntyre tells me that he is field-testing some now. What if all of these were worked out and then offered to the network? I don't know whether it is desirable or undesirable to have a common evaluation procedure for the network. I haven't thought of all the problems involved. Some of the advantages are that we could exchange information more reliably if we were talking about the same kind of inputs. This is very desirable, otherwise we're in it alone. We cannot share unless we are willing to work around the risk of differences of validity and reliability and, therefore, maybe, even lose some of the potential effectiveness for the evaluation. Another thing we have to face in terms of ı.etwork effort is, if one of the centers evaluates something, if we have an approved list and another center disagrees, what are we going to do about that? These are some of the problems, in terms of support, in terms of cooperation. I personally favor some kind of general similarity in a relationship either at the criteria level or at the software level. And I feel that if people did work intensively on these areas, and come up with some criteria and things like this then a center would probably use them because of the time and effort that they wouldn't have to put into it, and they would have something available that has merit, that has evolved within a setting analogous to their own, and so forth. It would seem that with the exception of the last two data sources, it would be desirable that two to four IMC's, for each of the first three levels, undertake to develop the criteria, point out the issues involved, develop policy statements and determine priorities. This is one possible way to divide up the work load. Criteria and priorities also need to be established for research. The big problem in research is not how to do research, but questions of priorities and cost, center resources, personnel, availability of populations and the like. With research one has the problem, unsolved to date, of communicating with the practitioner. So I propose that we can establish criteria, develop standards for software, approach the relevant population – field-test the software and disseminate these and the results within the network. I think for instance, in looking at the teacher rating forms that are being used in the network that we could have a long form and a short form as a minimum, depending on the purpose, and maybe the priorities are such that you are willing to forego certain information. You want a short form, so to speak, and it may even be used to collect different kinds of data. There could be 20 or 30 (at least) basic necessary bits of information. This information may be collected with different rating forms and the population sampled accordingly. I think though that initially it would be nice to have several forms and evolve from there. Or perhaps I'm even being wishful when I imply (or pretend) that there might be an abundance of forms – because of our commitments. I also suggest that these efforts be given sufficient priority that such developments and products be available a year from now.