

1: OBJECTIVES, PRIORITIES, AND OTHER JUDGMENT DATA

ROBERT E. STAKE

University of Illinois

This review is embedded in a plea. It is a plea to treat educational objectives as data. Fallible data. The review itself surveys methods for those particular data that reflect judgment of what education *should* accomplish.

Evaluation requires judgment. Decision-making requires judgment. Both are judgmental in themselves but also depend on judgments previously made. A school and a curriculum are where they are because of judgments from within and from without. Judgments are made early, and late, and in between times. To understand what a school is doing requires an understanding of what a school is expected to do.

In education, as elsewhere, judgments will continue to rest on incomplete knowledge, imprecise measurements, and inadequate experience. No error-free system is possible, but improvements are within easy reach. The evaluator may lessen the arbitrariness of judging and decision-making by introducing data-gathering methods already developed by other social scientists. Social psychologists, behavioral scientists, economists, political scientists, and historians routinely study opinions, preferences, and values. Many of their methods can be used to measure the judgments that shape an educational program.

Judgment Data

In this section several different kinds of judgment data are identified. Emphasized here are judgments of what educators should do rather than judgments of what educators have done. Some contemporary designers of evaluation studies call for the collection of judgment data, others do not. Their works are cited.

What Educators Should Do

Personal value-commitments, educational aims, goals, objectives,

*Leonard Cahen, Educational Testing Service, and Dennis Gooler, University of Illinois, served as consultants to Dr. Stake on the preparation of this chapter.

priorities, perceived norms, and standards—in one form of expression or another—are judgment data. An analysis of these data should reveal what someone wants school people to do. Many data-messages are only indirectly informative but still valuable. When a teacher says, "I believe a child should learn to be responsible to himself for what he does during study period"; when a parent says, "I want my son to go to college"; when a letter-to-the-editor writer says, "Kids can't be trusted"; or when a high official says, "Reading deficiencies are a national disgrace," judgment has been passed. Sometimes the subjective character of the utterance is obvious; sometimes the speaker is unaware that he has expressed a value. But in either case he is making a contribution to a value background that is, in fact, a definition of success and failure.

Success does not mean hitting a bulls-eye; success means coming *acceptably* close to a valued target. The responsibility of the evaluator is not only to find a good target-test and to tag the discrepant shots; he should also learn what accuracy is appropriate. He should learn which people hold the goal in high regard and which do not. But often an evaluator reports gain-score data with decimal precision and no data at all on the suitability of the instructional goals.

Compared to many educational-research data, judgment data are messy. They seem particularly susceptible to the obtrusiveness of formal evaluation. But even in natural conversation they are half shrouded, ambiguous, and imbued with emotion. Preferences are often inconsistent and arguments are circular. Whatever the clarity or confusion, that clarity or confusion should be acknowledged. Whatever the conflict and diversity, that too should be acknowledged. Whether people should think more rationally is not the issue since it is not the evaluator's task to reform human-judgment processes. The issue here is whether evaluators should treat these judgments as relevant data.

Many of those who write about educational processes take objectives as a starting point. *Objectives* are high-value targets. Objectives presumably identify the outcomes that someone thinks are most worthy, and all the unmentioned outcomes presumably are less worthy. Listing objectives can be thought of as selecting a few more-valued goals from a vast multitude of possible goals. The list is always an oversimplification. Goal-stating succeeds to whatever extent it succeeds because people are tolerant of omissions: particularly omitted superordinate goals and omitted statements of conditions. With any statement of objectives there are assumptions about more basic needs being met. In the classroom it is assumed that certain essential educational skills—sentence making and reading comprehension and shutting up and refraining from bodily threat to the teacher, for example—will be maintained. When these student skills falter, the teacher

is likely to abandon stated objectives to attend to the "essentials." Also, there are unstated but expected conditions for the instruction. As conditions change, objectives are modified. In other words, no set of stated objectives is a fixed and final list of what the teacher will or should attend to. The list does reveal what its authors judged to be worth special attention.

A list of objectives is an expression of attitudes. It is a valuable, sometimes essential, component of evaluation. It can be obtained by many devices, including attitude scales. On any educational scene there will be more than one set of objectives. Different groups have different attitudes, different objectives. Objectives are judgment data better treated by the rules that govern mass subjective responses than by the honors bestowed upon "fundamental truths." Objectives, like attitudes in all their subjectivity, can be collected and scaled objectively.

Sometimes it will be useful to compare objectives to generalized value commitments. *Personal values* are judgment data. In chapter 3 in this issue of the *Review*, Westbury indicates the relevance of value analysis to evaluation. Objectives can take on different meanings depending on the values behind them. A classroom emphasis on contemporary rock music may be rooted in a teacher's desire to improve the social conscience or in a teacher's desire to share aesthetic experiences. A workshop on behavior modification may be the creation of someone who seeks a more rational approach to instruction but also of another who sees an opportunity "to make a fast buck." Two bases of values analysis suggest themselves to the evaluator: (1) a logical basis to check on the reasonableness of the selection of objectives, given a certain value position (see Jensen, 1950); and (2) an empirical basis to see how broadly (e.g., in the community, among the faculty; see Larkins and Shaver, 1969) certain value-positions are held and how desirable certain objectives are perceived to be. Understanding value positions may be a shortcut to understanding educational objectives, or it may not. The evaluator needs to know what kinds of knowledge his clientele and audiences can use.

A knowledge of values should facilitate the specification of objectives. The evaluator should be acquainted with various schemas for categorizing values (Whitehead, 1929; Vernon and Allport, 1931; Oliver and Shaver, 1966; Scriven, 1966).

In Table I is an illustration of 16 value-laden broad educational objectives (Downey, 1960). Whether it would be useful to relate specific program objectives to these 16 broad objectives (or any other values typology) is something for the evaluator to consider when drawing up plans for a study.

TABLE I
DIMENSIONS OF THE TASK OF PUBLIC EDUCATION:
A CONCEPTUAL FRAMEWORK

A. Intellectual Dimensions

1. *Possession of Knowledge*: A fund of information, concepts.
2. *Communication of Knowledge*: Skill to acquire and transmit.
3. *Creation of Knowledge*: Discrimination and imagination, a habit.
4. *Desire for Knowledge*: A love for learning.

B. Social Dimensions

5. *Man to Man*: Cooperation in day-to-day relations.
6. *Man to State*: Civic rights and duties.
7. *Man to Country*: Loyalty to one's own country.
8. *Man to World*: Inter-relationships of peoples.

C. Personal Dimensions

9. *Physical*: Bodily health and development.
10. *Emotional*: Mental health and stability.
11. *Ethical*: Moral integrity.
12. *Aesthetic*: Cultural and leisure pursuits.

D. Productive Dimensions

13. *Vocation-Selective*: Information and guidance.
14. *Vocation-Preparative*: Training and placement.
15. *Home and Family*: Housekeeping, do-it-yourself, family.
16. *Consumer*: Personal buying, sellings, and investment.

The task of measuring values is difficult, Boulding (1957) said. Values, he claimed, are heterogeneous aggregates. But the elements must have some similarities or they would not be recognized as values. The evaluator's task is to reduce the apparent heterogeneity to a manageable representation, to separate the things-people-want-most from the other things, in a simple yet valid way. If in a given school situation a peoples' wants can be gathered into large groups of wants, one can label each group an educational value.

Another form of judgment data is the priorities given to certain objectives. The literal meaning of *priority* indicates "what comes first in the sequence of events," but the meaning here concerns relative importance. A list of objectives implies priorities; those expressed objectives have been considered to be more important than certain other objectives, a crude dichotomy. Priorities can be solicited that make finer gradations of importance. Priorities can indicate what kind and amount of emphasis will be given each objective. If there were unlimited resources or if all objectives were attainable in the time available, it would not be so important to specify the priorities. In actuality, it is important not only to choose the objectives to be pursued but to allocate scarce resources to each of the

several objectives. Of course, professional people do both—but usually intuitively. That these selections are not explicit and conscious in even the most prescriptive training programs is partly a reflection of the difficulty of stating “specific priorities” but mostly a reflection that people at all levels of expertise make most decisions intuitively. It may be true that some things will always be done better intuitively. We researchers need to find out whether or not specific priorities aid the direction of school programs. It is my belief that excessive attention has been given to precise goal-specification and insufficient attention to statements of priorities.

It would follow that the evaluator and the educator should have some procedures by which they can translate objectives into priorities when given an array of needs, a ledger of resources, and some knowledge about what the results of various instructional procedures are likely to be. Since there is no formula for deriving these priorities, the alternative is to find out what people want. What do staff, lay leaders, students—whoever are the important people—say should be the allotments of time, the allocation of concern and incentive, and the extent of remediation in the face of failure? These priorities are also judgment data. Pfeiffer (1968) introduced a few projects in which systems analysis has been used to quantify judgments and attach priority numbers to alternate objectives.

A fourth kind of judgment data deserves emphasis in an evaluation study: *standards*. Used here the term *standard* means a desired level or quality of something as cited by an authority. Standards answer the question “How much is good?” Standards are another form of objective: those seen by outside authority-figures who know little or nothing about the specific program being evaluated but whose advice is relevant to programs in many places.

When a local educator sets up the equivalent of a standard (e.g., students should score at the national mean on a reading test, should type at 50 words per minute, or should stay out of jail during the course of study), that standard is usually called an objective. When an authority figure, unaware of the means by which the local teachers will pursue objectives or even of what the local objectives are, indicates a desired level or performance or a desired environment, it is more likely to be called a standard. When President Nixon’s Committee on National Goals reports, it is likely to speak of criteria and standards—criteria to tell Americans what goals they should pursue and standards to prescribe a successful pursuit.

Standards can be specific or broad. Spokesmen for mathematics teachers from time to time specify minimum competencies for all young people, and spokesmen for the American Library Association express their ideas about exemplary library facilities for certain kinds of schools. These

are standards. Standards are another part of that background of values needed for the definition of success and failure. They are data usually to be treated more as expert legal testimony or as historical documentation than as population statistics. Yet the difference between casual treatment and deliberate and orderly treatment of such judgment data may be the difference between a perfunctory description and a sound base for decision-making. Taylor and De Corte (1969) examined the distinction between norms and standards and summarized the literature on reaching empirical definitions of standards.

Two obvious forms of judgment data remain to be mentioned. Many educational activities are undertaken to change student attitudes. Affective domain *outcome data* are judgment data. These will not be discussed directly in this chapter though many of the techniques cited are potentially useful. Krathwohl et al. (1964), Davis (1964), and Mayhew (1965) made major contributions toward the understanding of affective outcomes.

The last form of judgment data is perhaps the most critical to good evaluation: the *summative judgments* people make about the overall program or some component of it. Many different viewpoints will be gathered in some evaluation studies. The evaluator needs a tool kit of techniques for gathering and presenting these data. Such techniques will not be reviewed in this chapter, although ways of discovering what the educator should do will often be appropriate for discovering the worth of what the educator did do.

Additional References: Kluckhohn (1951); Churchman and Ratoosh (1959); Kuhn (1963); Educational Policies Commission (1961); Taylor (1970).

Evaluation Theory and Procedure

The principal claim of this chapter is that the processing of judgment data is important in educational evaluation. Evaluation always includes some “processing” of subjective data. But most of the writings on evaluation methodology do not mention procedures for gathering or analyzing judgment data. Most writers do not include judging the worth of alternative objectives and identifying standards as one of the evaluator’s jobs. And, even worse, a few writers treat differences in objectives and perceived value as the by-product of mismanaged schools, a by-product that could be expected to disappear with a more prescriptive system of curriculum development and a more behavioristic system of quality-control. Lortie (1967) perceived the social system as enduringly and desirably pluralistic; he found little promise in rational and prescriptive evaluation plans designed for relatively value-free data.

The evaluation literature is not the place to look for procedures for analysis of judgmental data. Metfessel and Michael (1967, p. 935) identified the following as one of eight major phases of an evaluation study.

7. Interpret the data in terms of certain judgmental standards and values concerning what are considered desirable levels of performance on the totality of collated measures—the drawing of conclusions which furnish information about the direction of growth, the progress of students, and the effectiveness of the total program.

But in a detailed appendix of data-gathering techniques they did not identify ways of gathering those judgmental data, standards, and values. In perhaps the most thorough large-scale evaluation study ever conducted, Smith and Tyler (1942) discussed the contributions of measurement data to guidance and administrative decisions; but they did not rely on standardized procedures to obtain and analyze judgment data. They accepted whatever objectives the educators gave them. Considering only the formal evaluation procedure, they chose to examine the Eight-Year Study objectives neither against other objectives, nor as formal transformations of values, nor in terms of the specific priorities given to them. A large majority of contemporary evaluation plans do likewise. Manuals and guidelines (Grobman, 1968; Tyler, 1968; McIntyre et al., 1969; Southwest Educational Development Laboratory, 1969) for project evaluation typically call for gathering statements of objectives without reference to their value loadings. They require no formal attention to priorities or standards. Such de-emphasis of judgment data was challenged by Glass (1968). He objected to Bloom's proposal to subsume evaluation methodology under contemporary testing methodology. Scriven's criticism (1967) of Cronbach's (1963) paper was another effort to avoid capture of evaluation methodology by the psychometric research establishment and to emphasize the specific "goods" and "bads" of classroom instruction. Scriven did not, unfortunately, say that judgment data should be treated as the social psychologist would treat "preferences" or as the economist would treat "utilities." He did not even say that the educator should receive an objective report as to what values are placed on various things by various groups. He did say that curriculum and instructions should be subjected to value analyses such as the philosopher or historian might employ. Berlak, in chapter 4 of this *Review*, appears to agree with Scriven in preferring logical analysis to empirical. Stake (1967a) voted for empiricism, urging the development of a social-science-based technology to handle judgment data.

For his taxonomy of educational evaluation designs, Worthen (1968) did not acknowledge that the extent to which an evaluation gathers and processes judgment data (as here defined) might be an important basis for differentiation among designs. This seems surprising since at the time he wrote the paper his academic adviser (Stufflebeam) had given as much attention to a formal plan for evaluating judgment data as any designer.

In Stufflebeam's (1969) CIPP model,* the C stands for context, which is to be interpreted partly as the context of values, i.e., the standards and objectives of the program to be evaluated. Stake (1967b) likewise emphasized value alternatives, assigning much of the total evaluation data matrix to judgment data with categories called "rationale," "intents," "standards," and "judgments."

Professional judgment—implicitly demeaned in most evaluation designs—was given more appropriate honor in the "accreditation" model of evaluation (see Glass, in press). The accrediting associations and related agencies have been admirably thorough in considering the many parts of an educational system, resisting pressures that would force complex dimensions of value into an undimensional criterion of merit. But these professional associations have been as indifferent as the Tylerian evaluators and the systems analysts to the need for objective methods for handling judgment data. The following statement taken from *Evaluative Criteria* (National Study of Secondary School Evaluation, 1969, p. 1:9) illustrates the personalistic standards solicited by this evaluation approach:

The checklists and evaluations should be evaluated on the following four-point scale:

- 4 Excellent
- 3 Good
- 2 Fair
- 1 Poor or missing
- na Not applicable

Question will frequently arise about the basis for comparison of points on the scale. The answer is extremely difficult to give. In any entity as complex as a school, it is not easy to describe in detail what *excellent* or *poor* really means in the hundreds of items for which evaluations are required. The best answer seems to be that the evaluator should draw upon his total experience in schools and make the best judgment he can on the basis of that experience.

Most enthusiastic advocates of "experiential" standards would agree that steps can be taken to communicate more clearly what a particular rating means by anchoring meaning in illustration and by sharing meaning through programmatic training and use. The methods cited on the following pages could improve this communication.

Additional References: Hyman and Wright (1967); Thomas (1968); American Institutes for Research in the Behavioral Sciences (1969).

*C = Context; I = Input; P = Process; P = Product

Methods of Gathering Judgment Data

In this section are reviewed some of the instruments and procedures available to evaluators for gathering data on objectives, values, priorities, and standards. Identified are three situations in which (from time to time) the evaluator works. In one, he solicits judgmental responses with a standardized protocol from a group of individuals and then aggregates these responses to describe the value commitments of that group. In another, he employs one or more "experts"—with or without a checklist or other structuring device—to make personal observations of events or processes and then to describe the value commitments apparent in them. In the third situation the evaluator employs experts to analyze documents—e.g., laws, curriculum guides, textbooks, or critical reviews—and to reconstitute the value commitments made in creating them. In any one study, of course, he might do all three. In this section also are citations of the technical means used by pollsters, observers, and analysts to gather judgment data.

Instruments for Aggregate Data

The most common procedure for getting judgment data about a group is to persuade members to indicate their individual viewpoints and to aggregate these in some way. The description of the whole is the aggregate descriptions of the parts. In this subsection are considered four ways of getting such data: (1) surveys, (2) scaling, (3) Q-technique, and (4) the semantic differential. These four techniques could be used in a pretest-posttest design to study attitude change as an outcome of an educational program; but the purpose of the discussion in this chapter is to consider background attitude status to probe the complex of values held by various groups. Regardless of whether attitude change is an objective, community and staff values influence teaching and learning. Those are the values that define the success of the program, and those are the values the reader should keep in mind as methods of data collection are identified.

Surveys are undertaken to obtain categorical answers to specific questions from a particular group of people. They may involve personal interviews or paper-and-pencil questionnaires. To get acquainted with survey methods, the evaluator should read Hyman (1955), Stephan and McCarthy (1958), and Trow (1967, 1969). A list of currently used educational questionnaires is published periodically by Gleason for the American Council on Education. Major evaluation studies should rely on professional assistance. It is available commercially from such agencies as Gallup Associates of Princeton, New Jersey, and the National Opinion Research Center of Chicago (see NORC, 1944). Professional assistance is also available on campuses which have survey-research offices. Directors of lesser studies

can develop or adapt their own procedures from those reported in the literature.

The educational-research literature describes few major surveys. The perspectives of education held by Catholics were surveyed by Greeley and Rossi (1966) and Neuwein (1966). Much earlier, Sandifer (1943) looked at lay perceptions of Progressive Education. Many non-generalizable studies of local circumstances are commissioned by school administrators. Most often these have an emphasis on economic characteristics; occasionally they attend to personal-value positions. (See Mort and Furno, 1960, and James, 1963, for illustrative studies; see Furno, 1966, for a review.)

Surveys are valuable when good data can be obtained by direct questioning. Often, however, ideas are vague, a single question is ambiguous, or meanings are personal and obscure. A more redundant and probing method is needed. When an objective or point of view is important enough to justify a more costly search and elusive enough to defy direct questioning, the evaluator may switch to *rating scales*. A display of many rating scales and a good introductory presentation were made by Shaw and Wright (1967). Guilford's *Psychometric Methods* (1954) is useful for a review of such classical topics as pair-comparisons and the time-order error. More theoretical treatments of scaling were developed by Torgerson (1958) and Coombs (1964).

Coughlan (1969) used pair-comparisons to study teachers' work values. Sjogren, England, and Meltzer (1969) devised an instrument for assessing the personal value-orientation of administrators. Gorlow and Noll (1967) developed an instrument for use with college students. Its forced-choice items were based on eight previously researched value dimensions. A number of scales designed to measure attitude and self-concept were described by Dowd and West (1969).

Messick (1961) used multidimensional scaling to portray the political preferences of lay adults. The advantage of the multidimensional approach is that dimensions (in this case, value orientations) do not have to be hypothesized in advance; the disadvantage of a multidimensional study is that the resulting dimensions are usually difficult to interpret and label.

A special use of rating scales (including a special factor analysis) was developed by Stephenson (1953). He called it the Q-technique. Its most common component, the Q-sort, is briefly and nicely described by Nunnally (1959). Downey (1960) used a Q-sort, for which the respondent sorted 16 stimulus cards into a 1-2-3-4-3-2-1 distribution of frequencies, to obtain priority values for global educational objectives. Stephenson showed how such sortings could be factor analyzed ipsatively (correlating persons instead of correlating scales) to get profiles for the individual respondent.

In a series of outstanding studies of educational values, Kerlinger and his associates used the Q-technique and other factor analytic approaches. They found two basic independent attitudes: that toward progressivism and that toward traditionalism. They did not find these to be, as many would expect, opposite poles of a single continuum but independent factors. In other words, they found that knowledge of a person's support for experimental projects, reconstruction, and life adjustment curricula (progressivism) is not a sound basis for predicting his criticism of the school, his esteem for knowledge, or his educational conservatism (traditionalism). But things within one of these two clusters would predict others within that cluster (Kerlinger, 1967; Kerlinger and Pedhazur, 1968; Sontag, 1968).

The *semantic differential* is a special scaling procedure for fixing meaning to objects and ideas. Developed by Osgood, Suci, and Tannenbaum (1957), the procedure enables the researcher to explore what people perceive to be the meaning of a concept. In education an evaluator may examine such concepts as "compensatory education," "state aid to parochial schools," "new math," or even "my teacher." When the search is for judgmental data, those concepts will be delineated by such descriptive scales as good-bad, needed-not needed, and useful-useless. Osgood and many researchers have used the semantic differential more to explore the dimensions of meaning people utilize than to explore the meaning of single concepts (see Snider and Osgood, 1969, for an excellent bibliography). For that goal, the factor analysis factors are more important than the specific concept descriptions. Since the evaluator is responsible for describing the value-background of a particular educational program, the semantic differential often yields descriptions more vague than he can use. Much time can be wasted trying to interpret new scales and new factors. The evaluator can probably make better use of the semantic differential by using scales from previous studies and by interpreting the results in terms of findings of those studies.

Geis (1968) examined the usefulness of the semantic differential for evaluating a course-content-improvement project (Harvard Project Physics), but his interest was in measuring student understandings rather than perceptions of the curriculum. Wittrock, Wiley, and McNeil (1967) used the technique (in a way consistent with the aim of this chapter) to examine what the concept "public school teachers" means. Harvey et al. (1968) related semantic differential data on teacher beliefs to classroom climate and student behavior. Taylor and Maguire (1967) used the semantic differential to study high school biology objectives. This last work is described in the section to follow in "Putting Objectives to the Test."

Many project evaluators, e.g., Peckham (undated), are using the semantic differential and other preference scales in project evaluation.

Though their findings may not be generalized beyond the confines of their project, it is regrettable that their reports are not available for the guidance of other evaluators designing studies and selecting instruments. Some such reports can be found in *Research in Education*; but since the ERIC items are classified by what-was-studied and who-was-studied rather than by the instrumentation, the search is a taxing one for the methodologist.

Additional References: Oppenheim (1966); William and Roberson (1967); Roper (1950); Wehling and Charters (1969).

Observation and Expert Review

How important things are to people can be pretty obvious. What they do longest and with most enthusiasm and what they work hardest to repair are what is important to them. The evaluator's problem with many judgmental data is not in gathering them, but in gathering them in such a way that the report readers, those who are not there to observe, can understand what was seen. The readers need some basis for deciding what confidence to place in the report. The observer and the reviewer need protocols, guides, routines to alert themselves by predetermined schedule to important features so as to give their observations the replication upon which confidence can be based. An excellent example is the form Westphal and Boldt (1970) developed to record a college physics lecturer's priorities in an objectives-by-audiences matrix.

Helen Peak (1953) reviewed the problems of gathering observational data in the Festinger and Katz handbook, *Research Methods in the Behavioral Sciences*. She emphasized the need for thorough training of the observers. Samph (undated) gathered evidence that the observer is likely to influence what happens in the classroom. More indirect ways of gathering observational data were suggested by Webb et al. (1966).

A number of techniques (e.g., Flanders Interaction Analysis) were developed for observing the social and instructional interaction among teachers and students. These are discussed by Rosenshine in chapter 5 in this issue of the *Review*. By and large, the techniques do not spotlight the *educational* criteria and standards operating in the classroom; but the evaluator may find them illustrative for developing his own protocol.

The shortage of procedures for making systematic observations of educational activities is particularly dismaying because the *site visit* is a widely used evaluation method. When a large-scale program is under way at some distant place, the most common way to evaluate it is to appoint a small number of respected persons to go there and inspect it. This method receives a proper share of criticism. It is evident that the program staff works hard to make the operation atypically handsome during the visit

and the visitors grasp at the slimmest shred of evidence for something to report. Despite these defects, the method of site visits deserves its eminence because it is designed for the most sensitive instruments available: experienced and insightful men. Furthermore, it is capable of quick adaptation to local circumstances. Its failings could be remedied by heeding Helen Peak's advice: train (even briefly) the visitors and provide a set of standardized scales for directing visitor attention for describing what is seen.

Most site visitors for high school accreditation are at least indirectly guided by the National Study *Evaluative Criteria* (1969), a publication which probably directly guides the lengthy "Self-Study" which precedes many visits. This publication gives the visitor looking for judgmental data many important topics to consider but, as mentioned before, little basis for getting suitable data.

Most site visitors to colleges in midwestern states are guided by the North Central Association's *Guide for the Evaluations of Institutions of Higher Learning* (1965). This document draws attention to seven basic questions but leaves the scaling to the ingenuity of the visitor.

1. What is the educational task of the institution?
2. Are the necessary resources available for carrying out the task of the institution?
3. Is the institution well organized for carrying out its institutional task?
4. Are the programs of instruction adequate in kind and quality to serve the purposes of the institution?
5. Are the institution's policies and practices such as to foster high faculty morale?
6. Is student life on campus relevant to the institution's educational task?

Workers at the Educational Testing Service are developing some needed scales; one of them is called the Institutional Functioning Inventory (Peterson, undated).

Most site-visit evaluators dispatched by the U. S. Office of Education to examine national laboratories and R & D centers are directed (USOE, undated; Chase, 1969) by broad criterion questions to identify the extent to which these units have become well managed and productive organizations. Here, in the Tylerian tradition mentioned earlier, no one is encouraged to discover whether or not the objectives are locally controversial or misstated or inconsistent with institutional philosophy or standards. A special-purpose site-visit plan (DESDEG) that does give such encouragement was developed by Renzulli and Ward (1969) to guide the outside evaluator who is evaluating a program for gifted children.

For another type of scrutiny, documents and artifacts can be sent from the school or project to the experts. (See Welch and Walberg, 1968.) *Expert review* is an important evaluation technique for textbooks, achievement tests, audio-visual materials, instructional and audio- and video-tapes. Again the purposes of evaluation may be best served by some standard analytic device for highlighting objectives or standards. Perhaps the greatest criticism of the expert reviews of educational tests edited by Buros (1965) is that the reviewers are not guided by a checklist or common set of standards. This is not to say that reviewers are unaware of basic principles of testing as summarized in Lindquist (1951) or of such guides as the Bloom Taxonomy (1956) for identifying the purposes of test items. Knowledge of principles and possible purposes does little to assure that a reviewer will look at values and expectations of different groups of users. If value data are important, the reviewer must be coached to look at values. This is more likely to happen in the review of textbooks because Gordon (1967) and Morrisett and Stevens (1967) have provided outlines that direct attention to judgment data. For the review of tests and tapes the priorities-seeking evaluator is pretty much on his own.

The thorough evaluator is tempted to analyze the documents of the community, the newspapers, and the minutes of meetings to learn how ideas and values have fared across time. Researchers call the technique *content analysis*. Berelson (1952) identified three situations for using content analysis: (1) when the researcher is curious about the contents themselves, (2) when he seeks inferences about the producers of the content, and (3) when he seeks to understand the audience that would use the content. Research on values and objectives fits nicely into both categories (2) and (3). Cahen (1970) modified a discomfort-relief quotient developed by Dollard and Mowrer (1947) to analyze the program of a professional school. (See Pool, 1959, for other examples.) Rules for good content analysis were spelled out by Cartwright (1953). Methodologists and users agree that the categories need to be carefully planned in advance. The evaluator may find the anthology of rationales, the taped interviews, and the video-taped recitations more exciting in prospect than in retrospect when countless pages of unanalyzed pastings and transcripts lie before him.

Content analysis is partly historiographic. The evaluator would do well to seek help from Barzun and Graff (1957) or Wise, Nordberg, and Reitz (1967). In keeping with his purposes, however, the evaluator should recognize the differences between research methods and evaluation methods (Cartwright, 1953, 449-454; Stake, 1969). Unlike the researcher the evaluator is not obligated, nor does he have good opportunity, to generalize to other instructional programs or procedures. For describing the specific project, he may find some methods of the public-relations man appropriate; the writings of McCloskey (1959) and Kindred (1960) may be instructive.

The best place within the educational-research literature for the evaluator to find aids to analysis of casual writing is that dealing with essay grading. The outstanding work in this area has been done by Godshalk, Swineford, and Coffman (1966) and Gosling (1966). Here as with so many psychometric studies the locus of attention has been on writing abilities rather than on what is needed here—the identification of personal concern in discourse or narrative.

Page is one researcher in this area who has realized the potential of the computer for discerning writing ability, writing style, content familiarity, and even personal concern. In his Project Essay Grade (1966) he found that his computer program could identify characteristics of student essays that correlated as well with English teacher gradings as the teachers' marks correlated among themselves. The promise of this work, it seems, lies in giving educators indicators for monitoring routine teaching and learning, not in developing a basis for substituting for teacher judgments. The general potential of natural language processing was discussed by Stone et al. (1966). The day when computers could aid the values analysis of curricular materials seems to be neither here nor far away.

Additional References: Payne (1969); Rosenshine (1969); Malinowski (1961).

Using Judgment Data in Evaluation Projects

In this section the uses of judgment data are examined. In the previous sections it was pointed out that judgment data are part of the context of education, that what is taught and what is learned are partly determined by personal preferences. An evaluator should realize that his audience can only poorly understand a program if they have no information on what is seen to be worth doing by those who are doing it. It is important to know what motivates Johnny, the learner; but for the program evaluator it is at least as important to know what motivates those who want Johnny to learn. Judgment data help show that the design of the program makes sense, or that it does not. Without judgment data an evaluator cannot show that a program is succeeding.

The evaluator should consider not only how educational objectives manifest themselves in teaching and learning, but also how those objectives embody the aspirations and discontents of the people involved. It is supposed that such motives were taken into account when objectives were established, but it is also supposed that students learn what they are taught. Evaluation is twice needed.

The development of sets of objectives has been thoroughly described elsewhere (Mager, 1962; Suchman, 1967; Eisner, 1969; Baker, 1969). Herzog (1959) romanticized the definitional problem with the words:

Big criteria have little criteria
upon their backs to bite 'em.
The small ones have still smaller,
and so on *ad infinitum*.

Krathwohl (1965) made a valuable contribution; he identified the stages of refinement of objectives and argued for performance objectives as the final product of the transformation. In the public health field Suchman (1967) illustrated a chain of objectives progressing from the most immediate practical objectives to the ultimate ideal goal. Taylor and Maguire (1966) presented the same ideas in a formal linear model, with societal press as the origin of objectives and criterion behavior as the outcome. In contrast to most followers of Tyler and Mager, Taylor and Maguire said that the origin of objectives should be the values of the people involved, not just the aims of the professional educators. The teacher serves an important function—as do the principal, board of education, textbook writer, and others—in translating national purposes and community needs into lesson plans. The evaluator should be able to give a clearer and more valid representation of community needs and generalized values. In the summative-evaluation sense and as an aid to planning and carrying out his study, the evaluator should display objectives against community value data to show what is congruent and what is not.

Putting Objectives to the Test

Few investigators examine empirically the relationship between values, objectives, and priorities. Maguire and Taylor have. Maguire (1968) obtained teacher value-ratings of a heterogeneous set of objectives, then from the same teachers got different expressions of the priorities that should be given to these objectives. He found that the teachers perceived the objectives in at least four dimensions; different in Subject-Matter Value, Motivational Qualities, Ease of Implementation, and Statement (Semantic) Properties. In another study, Taylor and Maguire (1967) obtained value-ratings of high school biology objectives from three important groups: subject-matter experts, curriculum writers, and biology teachers. The investigators found substantial agreement in group viewpoint, with experts and teachers least alike. In an earlier study Taylor (1966) used various scaling techniques to find specific points of disagreement about topic priorities. He worked with such specific topical assignments as "the study of structure versus function" and "the study of the

biological roots of behavior." These studies are valuable, I believe, much more because of the direct attack on the problems of measuring priority than for the guidance they give to the science educator.

According to the linear model visualized by Taylor and Maguire (1966), educational objectives are an intermediate product of a logical operation. Although the model represented societal press as generalized personal behaviors, it is easier for most curriculum analysts to represent it as social values. Whichever way, such judgmental data can be presented in matrix form. The input to the process is a matrix of values, the output is a matrix of objectives. For each personal point of view there is an input values profile, a row of the input matrix. For each point of view there would be a profile of intended outcomes, a row of the output matrix. The entire data collection on values thus would be represented by a persons-by-value-position matrix, and the data collection on objectives would be represented by a persons-by-objectives matrix. The elements within the matrices would be numerical ratings or priorities. The similarity of the values matrix to the objectives matrix can be examined, by eye or mathematically.

This matrix representation provides a way of displaying judgment data but also a platform for considering the question "By what transformation (e.g., matrix algebra operations) do objectives derive from values?" An educator probably responds to the value patterns of his people in ways that reflect his awareness of (1) the needs of those who are to benefit from education, (2) the resources available for education, and (3) the probability that any given way of teaching would alleviate a need. A researcher seeking to explain educator behavior may need to include these three domains in his theory. And an evaluator seeking to aid an educator may gather data from these three domains.*

It is reasonable to assume that the educator sets lower program priorities on those things held in high value but for which he sees no current need and for which he sees a need but no sufficiently inexpensive or potentially successful educational strategy. This suggests that what an educator or any other person knows about student need, potentially successful pedagogy, and instructional resources may help in the analysis of the objectives he would emphasize for a school program. The evaluator who thinks along these lines quickly becomes reminded of the need for research findings on the functional relationships between various amounts

*If an evaluator desires to find the main ways his judges or supervisors are reacting to such things as objectives or standards, he may use regression analyses. Maguire and Glass (1968) and Schenck and Naylor (1968) described how regression can be used to categorize rater behavior in terms of the characteristics of the things he is rating.

of input and output as bases for educational decisions, a topic to be discussed later. Cronbach and Gleser (1965) established a useful methodological precedent in the field of personnel decisions.*

In the administration of Title III programs for educational innovation and supplementary services, the U. S. Office of Education has paid increasing attention to "Needs Assessment." It has mandated state-by-state studies. Some efforts to satisfy this requirement are being made as if *need* were a non-reactive characteristic (such as age, geographic location, or loudness) and not dependent on who is being asked to describe it. To describe need is not only to describe the something but to describe the persons asked. People see need differently. The evaluator has no obligation to find consensus. In fact, he is acting improperly if he does not report the diversity of viewpoints of need (McLure, 1968). The developers of a state plan do have to find a compromise. Pennsylvania now has its plan (Educational Testing Service, 1965). A reader specifically interested in Title III might use the needs-assessment plan for Florida as a model. (Florida Educational Research and Development Council, 1968.) Popham (1969) recognized the important need in *Needs Assessment* for empirical data on preferences. He equated critical needs with discrepancies between (1) preference of items from the UCLA Objectives Exchange and (2) performance levels on still-to-be-constructed-or-selected National-Assessment-like criterion-reference achievement items. (See Popham and Skager, 1968; Department of Elementary School Principals, 1967.)

Dorothy Fraser (1963, p. 105,) speaking for the National Committee of the NEA Project on Instruction, recognized the criticality of *priorities* but avoided the educational-technologist procedural question saying, "There is no set of specifications for a balanced curriculum which can be applied to every school in the United States, just as there is no uniform prescription to determine what should be included and excluded from the school program." The idea that *how-to-translate-local-objectives-into-local-program* is only a local matter is unacceptable. The National Committee did review the important subject-matter areas; it should also have addressed itself to the questions of "How can the profession help the community see and adjust the 'balance' in its curriculum?"

Such techniques as PERT (Cook, 1966) or the curriculum-development models of Gagné (1967) or Hively, Patterson, and Page (1968) are helpful in ordering what to do first and what to do second; but only in the hands of a skilled manager will they be helpful to the educator trying to give specific priorities to different objectives.

*I do not imply that the best way to generate objectives is to consider values and needs first. I mean only that data on objectives may be better understood if data on values and needs are also at hand.

