THE COUTENANCE OF EDUCATIONAL EVALUATION

Robert E. Stake
Center for Instructional Research and Curriculum Evaluation
University of Illinois

President Johnson, President Conant, Mrs. Hull (Sara's teacher) and Mr. Tykociner (the man next door) are quite alike in the faith they have in education. But they have quite different ideas of what education is. The value they put on education does not reveal their way of evaluating education.

Educators differ among themselves as to both the essence and worth of an educational program. The wide range of evaluation purposes and methods allows, each to keep his own perspective. Few see their own programs "in the round", partly because of a parochial approach to evaluation. To understand better his own teaching and to contribute more to the science of teaching, each educator should examine the full countenance of evaluation.

Educational evaluation has its formal and informal sides. Informal evaluation is recognized by its dependence on casual observation, implicit goals, intuitive norms, and subjective judgment. Perhaps because these are also characteristic of day-td-day, personal styles of living, informal evaluation results in perspectives which are seldom questioned. Careful study reveals informal evaluation of education to be of variable quality - sometimes penetrating and insightful, sometimes superficial and distorted.

Formal evaluation of education is recognized by its dependence on checklists, structured visitation by peers, controlled comparisons, and standardized testing of students. Some of these techniques have long histories of successful use. Unfortunately, when planning an evaluation, few educators consider even these four. The more common notion is to evaluate informally: to ask the opinion of the instructor, to ponder the logic of the program, or to consider the reputation of the advocates. Seldom do we find a search for relevant research reports or for behavioral data pertinent to the ultimate curricular decisions.

Dissatisfaction with the formal approach is not without cause. Few highly relevant, readable, research studies can be found. The professional journals are not disposed to publish evaluation studies. Behavioral data are costly, and often do not provide the answers. Too many accreditation -type visitation teams lack special training or even experience in evaluation. Many checklists are ambiguous; some focus too much attention on the physical attributes of a school. Psychometric tests have been developed primarily to differentiate among students at the same point in training rather than to assess the effect of instruction on acquisition of skill and understanding. Today's educator may rely little on formal evaluation because its answers have seldom been answers to questions he is asking.

The educator's disdain of formal evaluation is due also to his sensitivity to criticism - and his is a critical clientele.. It is not uncommon for him to draw before him such curtains as "national norm comparisons, " "innovation phase, " and "'academic freedom" to avoid exposure through evaluation. The ' politics" of evaluation is an interesting issue in itself, but it is not the issue 'here. The issue here is the potential contribution to education of formal evaluation. Today, educators fail to perceive what formal evaluation could do for them. They should be imploring measurement specialists to develop a methodology that reflects the fullness, the complexity, and the importance of their programs. They are not.

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). Little attempt has been made to measure the match between what an educator intends to do and what he does do. The traditional concern of educational-measurement specialists for reliability of individual-student scores and predictive validity (thoroughly and competently stated in the American Council on Education's 1950 edition of Educational Measurement) is a questionable resource. For evaluation of curricula, attention to individual differences among students should give way to attention to the contingencies among background conditions, classroom activities, and scholastic outcomes.

This paper is not about what should be measured or how to measure. It is background for developing an evaluation plan. What and how are decided later. My orientation here is around educational programs rather than educational products. I presume that the value of a product depends on its program of use. The evaluation of a program includes the evaluation of its materials.

The countenance of educational evaluation appears to be changing. On the pages that follow, I will indicate what the countenance can, and perhaps, should be. My attempt here is to introduce a conceptualization of evaluation oriented to the complex and dynamic nature of education, one which gives proper attention to the diverse purposes and judgments of the practitioner.

Much recent concern about curriculum evaluation is attributable to contemporary large-scale curriculum -innovation activities, but the statements in this paper pertain to traditional and new curricula alike. They pertain, for example, to Title I and Title III projects funded under the Elementary and Secondary Act of 1966. Statements here are relevant to any curriculum, whether oriented to subject-matter content or to student process, and without regard to whether curriculum is general-purpose, remedial, accelerated, compensatory, or special in any other way.

The purposes and procedures of educational evaluation will vary from instance to instance. What is quite appropriate for one school may be less appropriate for another. Standardized achievement tests here but not there. A great concern for expense there but not over there. How do evaluation purposes and procedures vary? What are the basic characteristics of evaluation activities? They are identified in these pages as the evaluation acts, the data sources, the congruence and contingencies, the standards, and the uses of evaluation. The first distinction to be made will be between description and judgment in evaluation.

Description and Judgment

The countenance of evaluation beheld by the educator is not the same one beheld by the specialist in evaluation. The specialist sees himself as a "describer", one who describes aptitudes and environments and accomplishments. The teacher and school *administrator, on* the other hand, expect an evaluator to grade something or someone as to merit. Moreover, they expect that he will judge things against external standards, on criteria perhaps little related to the local school's resources *and goals*.

Neither sees evaluation broadly enough. Both 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.

The specialist in evaluation seems to be increasing his emphasis on fullness of description. For many years he evaluated primarily by measuring student progress toward academic objectives. These objectives usually were identified with the traditional disciplines, e.g. mathematics, English, and social studies. Achievement tests - standardized or "teacher-made" - were found to be useful in describing the degree to which some curricular objectives are attained by individual students in a particular course. To the early evaluators, and to many others, the countenance of evaluation has been nothing more than the administration and normative interpretation of achievement tests.

In recent years a few evaluators have attempted, in addition, to assess progress of individuals toward certain "inter -disciplinary" and "extracurricular" objectives. In their objectives, emphasis has been given to the integration of behavior within an individual; or to the perception of interrelationships among scholastic disciplines; or the development of habits, skills, and attitudes which permit the individual to be a craftsman or scholar, in or out of school. For the descriptive evaluation of such outcomes, the Eight-Year Study (Smith and Tyler, 1942) has served as one model. The proposed National Assessment Program may be another - this statement appeared in one interim report:

> . . . all committees worked within the following broad definition of 'national assessment:' 1. In order to reflect fairly the aims of education in the U.S., the assessment should consider both traditional and modern curricula, and take into account <u>all the aspira</u>tions schools have for developing attitudes and motivations as well as knowledge and skills ... " [Italics added]. (Educational Testing Service, 1965).

In his 1964 paper, "Course Improvement through Evaluation, " Lee Cronbach urged another step: a most generous inclusion of behavioral - science variables in order to examine the possible causes and effects of quality teaching He proposed that the main objective for evaluation is to uncover durable relationships -those appropriate for guiding future educational programs. To the traditional description of pupil achievement, we add the description of instruction and the description of relationships between them. Like the instructional researcher, the evaluator - as so defined - seeks generalizations about**,** educational practices. Many curriculum project evaluators are adopting this definition of evaluation.

Description is one thing, judgment is another. Most evaluation specialists have chosen not to judge. But in his recent Methodology of Evaluation Michael Scriven has charged evaluators with responsibility for passing upon the merit of an educational practice. (Note that he has urged the evaluator to do what the educator has expected the evaluator to be doing.) Scriven's position is that there is no evaluation until judgment has been passed, and by his reckoning the evaluator is best qualified to judge.

By being well experienced and by becoming well-informed in the case at hand in matters of research and educational practice the evaluator does become at least partially qualified to judge. But is it wise for him to accept this responsibility? Even now when few evaluators expect to judge, educators are reluctant to initiate a formal evaluation. If evaluators were more frequently identified with the passing of judgment, with the discrimination among poorer and better programs, and with the awarding of support and censure, their access to data would probably diminish. Evaluators collaborate with other social scientists and behavioral research workers. Those who do not want to judge deplore the acceptance of such responsibility by their associates. They believe that in the eyes of many practitioners, social science and behavioral researchwill become more suspect that it already is.

Many evaluators feel that they are not capable of perceiving, as they think a judge should, the unidimensional value of alternative programs. They anticipate a dilemma such as Curriculum I resulting in three skills and ten understandings and Curriculum II resulting in four skills and eight understandings. They are reluctant to judge that gaining one skill is worth losing two understandings. And, whether through timidity, disinterest, or as a rational choice, the evaluator usually supports "local option, " a community's privilege to set its own standards and to be its own judge of the worth of its educational system. He expects that what is good for one community will not necessarily be good for another community, and he does not trust himself to discern what is best for a briefly-known community.

Scriven reminds them that there are precious few who can judge complex programs, and fewer still who will. Different decisions must be made - or Harvard Physics? - and they should not be made on trivial criteria, e.g. mere precedent mention in the popular press, salesman personality, administrative convenience, or pedagogical myth. Who should judge? The answer comes easily to Scriven partly because he expects little interaction between treatment and learner, i.e., what works best for one learner will work best for others, at least within broad categories. He also expects that where the local good is at odds with the common good, the local good can be shown to be detrimental to the common good, to the end that the doctrine of local option is invalidated. According to Scriven the evaluator must judge.

Whether or not evaluation specialists will accept Scriven's challenge remains to be seen. In any case, it is likely that judgments will become an increasing part of the evaluation report. Evaluators will seek out and record the opinions of persons of special qualification. These opinions, though subjective, can be very useful and can be gathered objectively, independent of the solicitor' opinions. A responsibility for processing judgments is much more acceptable t the evaluation specialist than one for rendering judgments himself.

Taylor and Maguire (1965) have pointed to five groups having important opinions on education: spokesmen for society at large, subject-matter experts, teachers, parents, and the students themselves. Members of these and other groups are judges who should be heard. Superficial polls, letters to the editor, and other incidental judgments are insufficient. An evaluation of a school program should portray the merit and fault perceived by well-identified groups, systematically gathered and processed. Thus, judgment data and description data are both essential to the evaluation of educational programs.

Data Matrices

In order to evaluate, an educator will gather together certain data. The data are likely to be from several quite different sources, gathered in several quite different ways. Whether the immediate purpose is description or judgment, three bodies of information should be tapped. In the evaluation report it can be helpful to distinguish between antecedent, transaction and outcome data.

An antecedent is any condition existing prior to teaching and learning which may relate to outcomes. The status of a student prior to his lesson, e.g. his aptitude, previous experience, interest, and willingness, is a complex antecedent. The programmed -instruction specialist calls some antecedents "entry behaviors. " The state accrediting agency emphasizes the investment of community resources. All of these are examples of the antecedents which an evaluator will describe.

Transactions are the countless encounters of students with teacher, student with student, author with reader, parent with counselor - the succession of engagements which comprise the process of education. Examples are the presentation of a film, a class discussion, the working of a homework problem, an explanation on the margin of a term paper, and the administration of a test. Smith and Meux studied such transactions in detail and have provided an 18 – category classification system. One very visible emphasis on a particular class of transactions was the National Defense Education Act support of audiovisual media.

Transactions are dynamic whereas antecedents and outcomes are relatively static. The boundaries between them are not clear, e. g. during a transaction we can identify certain outcomes which are feedback antecedents for subsequent learning. These boundaries do not need to be distinct. The categories serve to remind us to be exhaustive in our data collection.

Traditionally, most attention in formal evaluation has been given to outcomes - outcomes such as the abilities, achievements, attitudes, and aspirations of students resulting from an educational experience. Outcomes, as a body of information, would include measurements of the impact of instruction on teachers, administrators, counselors, and others. Here too would be data on wear and tear of equipment, effects of the learning environment, cost incurred. Outcomes to be considered in evaluation include not only those that are evident, or even existent, as learning sessions end, but include applications, transfer, and relearning effects which may not be available for measurement until long after. The description of the outcomes of driver training, for example, could well include reports of accident -avoidance over a lifetime. In short, outcomes are the consequences of educating - immediate and long-range, cognitive and conative, personal and community-wide.

Antecedents, transactions, and outcomes, the elements of evaluation statements, are shown in Figure I to have a place in both description and judgment. To fill in these matrices the evaluator will collect judgments (e. g. of community prejudice, of problem solving styles, and of teacher personality) as well as descriptions. In Figure 1 it is also indicated that judgmental statements are classified either as general standards of quality or as judgments specific to the given program. Descriptive data are classified as intents and observations. The evaluator can organize his data-gathering to conform to the format shown in Figure 1.

The evaluator can prepare a record of what educators intend, of what observers perceive, of what patrons generally expect, and of what judges value the immediate program to be. The record may treat antecedents, transactions, and outcomes separately within the four classes identified as. Intents, Observation, Standards, and Judgments, as in Figure 1. The following is an illustration of 12 data, one of which could be recorded in each of the 12 cells, starting with an intended antecedent, and moving down each column until an outcome has been indicated.

> Knowing that (1) Chapter XI has been assigned and that he intends (2) to lecture on the topic Wednesday, a professor indicates (3) what the students should be able to do by Friday, partly by writing a quiz on the topic. He observes that (4) some students were absent on Wednesday, that (5) he did not quite complete the lecture because of a lengthy discussion and that (6) on the quiz only about 2/3 of the class seemed to understand a certain major concept. In general, he expects (7) some absences but that the work will be made up by quiz-time; he expects (8) his lectures to be clear enough for perhaps 90 percent of a class to follow him without difficulty; and he knows that (9) his colleagues expect only about one student in ten to understand thoroughly each major concept in such lessons as these. By his own judgment (10) the reading assignment was not a sufficient background for his lecture; the students commented that (11) the lecture was provocative; and the graduate assistant who read the quiz papers said that (12) a discouragingly large number of students seemed to confuse one major concept for another.

Evaluators and educators do not expect data to be recorded in such detail, even in the distant future. My purpose here was to give twelve examples of data that could be handled by separate cells in the matrices. Next I would like to consider the description data matrix in detail.

For many years instructional technologists, test specialists, and others have pleaded for more explicit statement of educational goals. I consider "goals," “objectives," and "intents" to be synonymous. I use the category title Intents because many educators now equate "goals" and "objectives" with "intended student outcomes. " In this paper Intents includes the planned-for environmental conditions, the planned-for demonstrations, the planned-for coverage of certain subject matter, etc., as well as the planned-for student behavior. To be included in this three-cell column are effects which are desired, those which are hoped for, those which are anticipated, and even those which are feared. This class of data includes goals and plans that others have, especially the students. (It should be noted that it is not the educator's privilege to rule out the study of a variable a variable by saying, "that is not one of our- objectives. " The evaluator

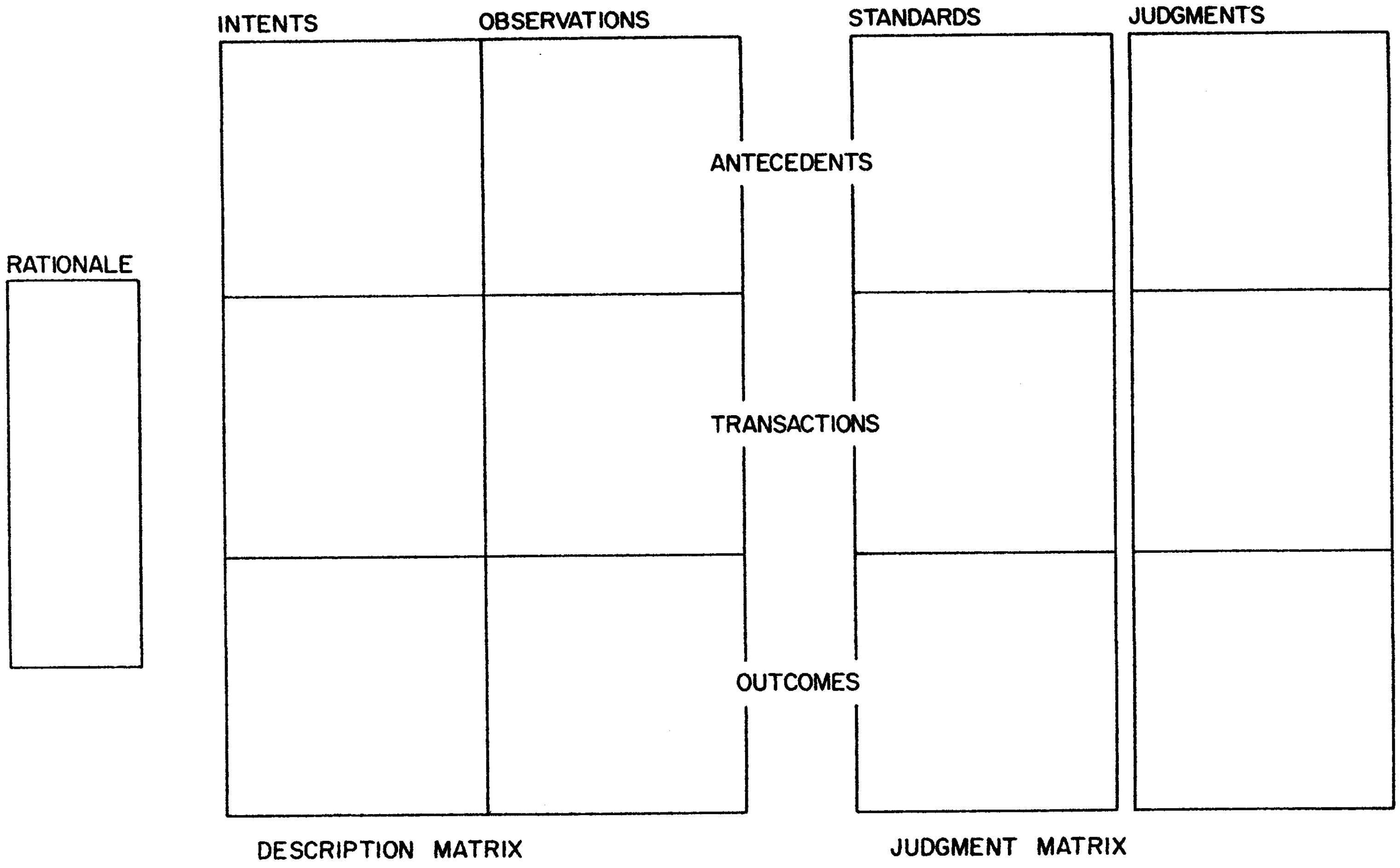

Figure 1. A layout of statements and data to be collected by the evaluator of an educational program.

should include both the variable and the negotiation.) The resulting collection of Intents is a priority listing of all that may happen.

The fact that many educators now equate "goals" with "intended student outcomes" is to the credit of the behaviorists, particularly the advocates of programmed instruction. They have brought about a small reform in teaching by emphasizing those specific classroom acts and work exercises which or contribute to the refinement of student responses. The A.A.A.S. Science Project, for example, has been successful in developing its curriculum around behavioristic goals (Gagne, 1966). Some curriculum-innovation projects, however, have found the emphasis on behavioral outcomes an obstacle to creative teaching (Atkin, 1,963). The educational evaluator should not list goals only in terms of anticipated student behavior. To evaluate an educational program, we must examine what teaching, as well as what learning, is intended. (Many antecedent conditions and teaching transactions can be worded behavioristically, if desired.) How intentions are worded is not a criterion for inclusion. Intents can be the global goals of the programmer (Mager, 1962). Taxonomic, mechanistic, humanistic, even scriptural - any mixture of goal statements are acceptable as part of the evaluation picture.

Many a contemporary evaluator expects trouble when he sets out to record the educator's objectives. Early in the work he urged the educator to declare his objectives so that outcome-testing devices could be built. He finds the educator either reluctant or unable to verbalize objectives. With diligence, if not with pleasure, the evaluator assists with what he presumes to be the educator’s job: writing behavioral goals. His presumption is wrong. As Scriven (1965) has said, the responsibility for describing curricular objectives is the responsibility of the evaluator. He is the one who is experienced with the language of behaviors, traits, and habits. Just as it is his responsibility to transform the behaviors of a teacher and the responses of a student into data, it is his responsibility to transform the intentions and expectations of an educator into "data. It is necessary for him to continue to "Is this an instance? " It is not wrong for an evaluator to teach a willing educator about behavioral objectives -they may facilitate the work. It is wrong for him to insist that every educator should use them.

Obtaining authentic statements of intent is a new challenge for the evaluator. The methodology remains to be developed. Let us now shift attention to the second column of the data cells.

Most of the descriptive data cited early in the previous section are classified as Observations. In Figure I when he described surroundings and events and the subsequent consequences, the evaluator[1] is telling of his observations. Sometimes the evaluator observes these characteristics in a direct and personal way. Sometimes he uses instruments. His instruments include inventory schedules, biographical data sheets, interview routines, checklists, opinionnaires, and all kinds of psychometric tests. The experienced evaluator gives special attention to the measurement of student outcomes, but he does not fail to observe the other outcomes, nor the antecedent conditions and instructional transactions.

[1] Here *and elsewhere* in this paper, for simplicity of presentation, the evaluator and the educator are referred to as two different persons. The educator will often be his own evaluator or a member of the evaluation team.

Many educators fear that the outside evaluator will not be attentive to the characteristics that the school staff has deemed most important. This sometimes does happen, but evaluators often pay too much attention to what they have been urged to look at, and too little attention to other facets. In the matter of selection of variables for evaluation, the evaluator must make a subjective decision. Obviously, he must limit the elements to be studied. He cannot look at all of them. The ones he rules out will be those that he assumes would not contribute to an understanding of the educational activity. He should give primary attention to the variables specifically indicated by the educator's objectives, but he must designate additional variables to be observed. He must search for unwanted side effects and incidental gains. The selection of measuring techniques is an obvious responsibility, but the choice of characteristics to be observed is an equally important and unique contribution of the evaluator.

An evaluation is not complete without a statement of the rationale of the program. It needs to be considered separately, as indicated in Figure 1. Every program has its rationale, though often it is only implicit. The rationale indicates the philosophic background and basic purposes of the program. Its importance to evaluation has been indicated by Berlak (1966). The rational should provide one basis for evaluating Intents. The evaluator asks himself or other judges whether the plan developed by the educator constitutes a logical step in the implementation of the basic purposes. The rationale also is of value in choosing the reference groups, e.g. merchants, mathematicians, and mathematics educators, which later are to pass judgment on various aspects of the program.

A statement of rationale may be difficult to obtain. Many an effective instructor is less than effective at presenting an educational rationale. If pressed, he may only succeed in saying something the listener wanted said. It is important that the rationale be in his language, a language he is the master of. Suggestions by the evaluator may be an obstacle, becoming accepted because they are attractive rather than because they designate the grounds for what the educator is trying to do.

The judgment matrix needs further explanation, but I am postponing that until after a consideration of the bases for processing descriptive data.

Contingency and Congruence

For any one educational program there are two principal ways of processing descriptive evaluation data: finding the contingencies among antecedents, transactions, and outcomes and finding the congruence between Intents and Observations. The processing of judgments follows a different model. The first two main columns of the data matrix in Figure 1 contain the descriptive data. The format for processing these data is represented in Figure 2.

The data for a curriculum are congruent if what was intended actually happens. To be fully congruent the intended antecedents, transactions, and out- - comes would have to come to pass. (This seldom happens -and often should not.)

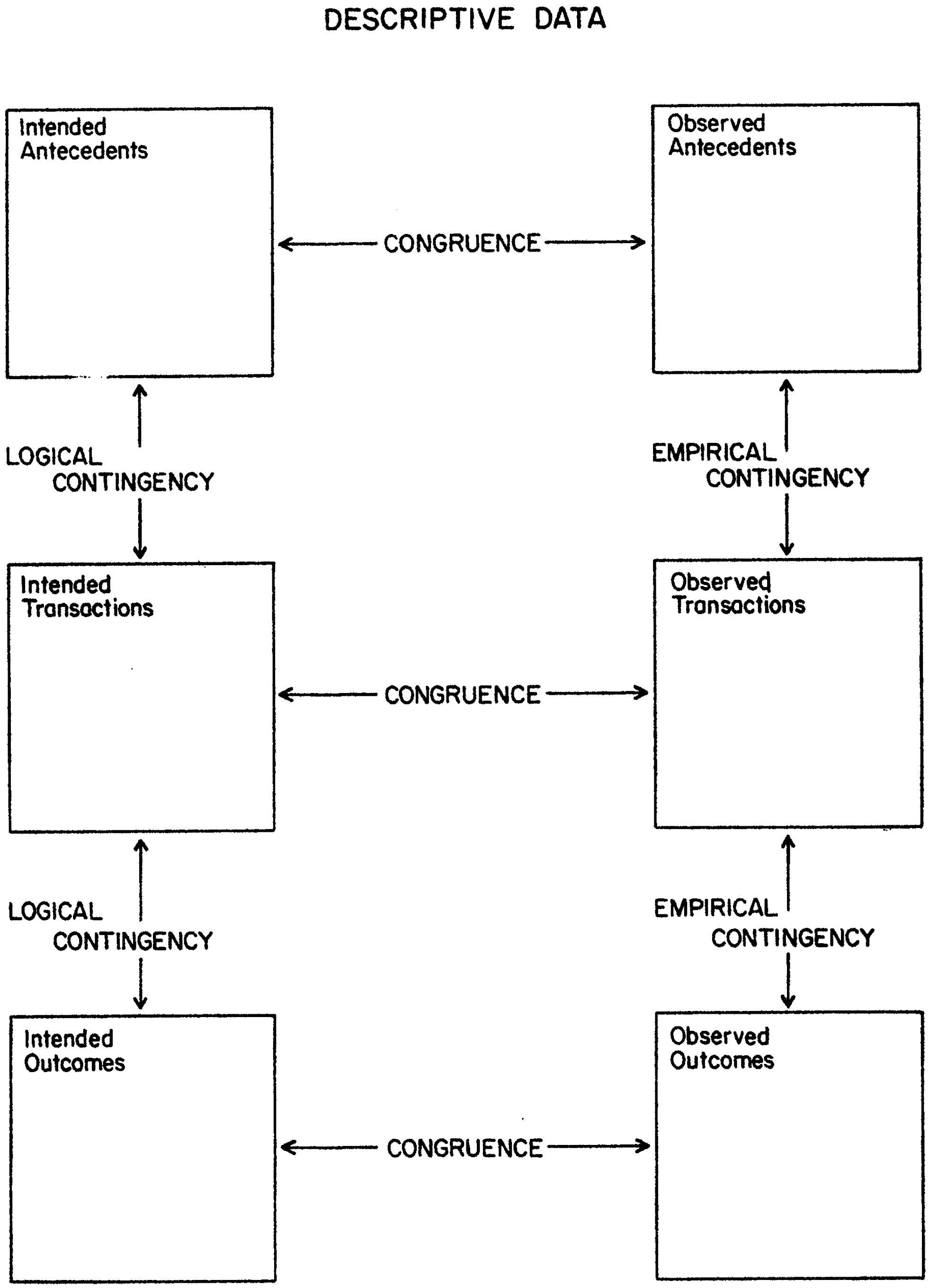

Figure 2. A representation of the processing of descriptive data.

Within one row of the data matrix the evaluator should be able to compare the cells containing Intents and Observations, to note the discrepancies, and to describe the amount of congruence for that row. (Congruence of outcomes has been emphasized in the evaluation model proposed by Taylor and Maguire, 1965).) Congruence does not indicate that outcomes are reliable or valid, but that what was intended did occur.

Just as the Gestaltist found more to the whole than the sum of its parts, the evaluator studying variables from any two of the three cells in a column of the data matrix finds more to describe than the variables themselves. The relationships or contingencies among the variables deserve additional attention. In the sense that evaluation is the search for relationships that permit the improvement of education, the evaluator's task is one of identifying outcomes that are contingent upon particular antecedent conditions and instructional transactions.

Lesson planning and curriculum revision through the years has been built upon faith in certain contingencies. Day to day, the master teacher arranges his presentation and selects his input materials to fit his instructional goals. For him the contingencies, in the main, are logical, intuitive, and supported by a history of satisfactions and endorsements. Even the master teacher and certainly less -experienced teachers need to bring their intuited contingencies under the scrutiny of appropriate juries.

As a first step in evaluation it is important just to record them. A film of floodwaters may be scheduled (intended transaction) to expose students to a background to conservation legislation (intended outcome). Of those who know both subject matter and pedagogy, we ask, "Is there a logical connection between this event and this purpose? " If so, a logical contingency exists between these two Intents. The record should show it.

Whenever Intents are evaluated the contingency criterion is one of logic. To test the logic of an educational contingency the evaluators rely on previous experience, perhaps on research experience, with similar observables. No immediate observation of these variables, however, is necessary to test the strength of the contingencies among Intents.

Evaluation of Observation contingencies depends on empirical evidence. To say, "this arithmetic class progressed rapidly because the teacher was somewhat but not too sophisticated in mathematics" demands empirical data, either from within the evaluation or from the research literature (see Bassham, 1960). The usual evaluation of a single program will not alone provide the data necessary for contingency statements. Here too, then, previous experience with similar observables is a basic qualification of the evaluator (Ausubel, 19).

The contingencies and congruences identified by evaluators are subject to judgment by experts and participants just as more unitary descriptive data are, The importance of non-congruence will vary with different view-points. The school superintendent and the school counselor may disagree as to the importance of a cancellation of the scheduled lessons on sex hygiene in the health class. As an example of judging contingencies, the degree to which teacher morale is contingent on the length of the school day may be deemed cause enough to abandon an early morning class by one judge and not another. Perceptions of importance of congruence and contingency deserve the evaluator's careful attention.

Standards and Judgments

There is general agreement that the goal of education is excellence - but how schools and students should excel, and at what sacrifice, will always be debated. Whether goals are local or national, the measurement of excellence requires explicit rather than implicit standards.

Today's educational programs are not subjected to "standard -oriented" evaluation. This is not to say that schools lack in aspiration or accomplishment. It is to say that standards - benchmarks of performance having widespread reference value - are not in common use. Schools across the nation may use the same evaluation checklist [2] but the interpretations of the checklisted data are couched in inexplicit, personal terms. Even in an informal way, no school can evaluate the impact of its program without knowledge of what other schools are doing in pursuit of similar objectives. Unfortunately, many educators are loathe to accumulate that knowledge systematically (Hand, 1965, Tyler, 1965).

There is little knowledge anywhere today of the quality of a student's education. School grades are based on the private criteria and standards of the individual teacher. Most "standardized" test scores tell where an examinee performing "psychometrically useful" tasks stands with regard to a reference group, rather than the level of competence at which he performs essential scholastic tasks. Although most teachers are competent to teach their subject matter and to spot learning difficulties, few have the ability to describe a student's command over his intellectual environment. Neither school grades nor standardized test scores nor the candid opinions of teachers are very informative as to the excellence of students.

Even when measurements are effectively interpreted, evaluation is complicated by a multiplicity of standards. Standards vary from student to student, from instructor to instructor, and from reference group to reference group. This is not wrong. In a healthy society, different parties have different standards. Part of the responsibility of evaluation is to make known which standards are held by whom.

It was implied much earlier that it is reasonable to expect change in an educator's Intents over a period of time. This is to say that he will change both his criteria and his standards during instruction. While a curriculum is being developed and disseminated, even the major classes of criteria vary. In their analysis of nationwide assimilation of new educational programs, Clark and Guba (1965) identified eight stages through which new programs go. For each stage they identified special criteria (each with its own standards) on which the program should be

[2] One contemporary checklist is Evaluative Criteria a document published by the National Study of Secondary School Evaluation (1960). It is a commendably thorough list of antecedents and possible transactions, organized mostly by subject-matter offerings. Surely it is valuable as a checklist, identifying neglected areas. Its great value may be a catalyst, hastening the maturity of a developing curriculum. However, it can be of only limited value in evaluating, for it guides neither the measurement nor the interpretation of measurement. By intent, it deals with criteria (what variables to consider) and leaves the matter of standards (what ratings to consider as meritorious) to the conjecture of the individual observer.

evaluated before it advances to another stage. Each of their criteria deserves elaboration, but here it is merely noted that there are quite different criteria at each successive curriculum-development stage.

Informal evaluation tends to leave criteria unspecified. Formal evaluation is more specific. But it seems the more careful the evaluation, the fewer the criteria; and the more carefully the criteria are specified, the less the concern given to standards of acceptability. It is a great misfortune that the best trained evaluators have been looking at education with a microscope rather than with a panoramic view finder.

There is no clear picture of what any school or any curriculum project is accomplishing today partly because the methodology of processing judgments is inadequate. What little formal evaluation there is is attentive to too few criteria, overly tolerant of implicit standards, and ignores the advantage of relative comparisons. More needs to be said about relative and absolute standards.

Comparing and judging

There are two bases of judging the characteristics of a program, (1) with respect to absolute standards as reflected by personal judgments and (2) with respect to relative standards as reflected by characteristics of alternate programs. One can evaluate SMSG mathematics with respect to opinions of what a mathematics curriculum should be or with regard to what other mathematics curricula are. The evaluator's comparisons and judgments are symbolized in Figure 3. The upper left matrix represents the data matrix from Figure 2. At the upper right are sets of standards by which a program can be judged in an absolute sense. There are multiple sets because there may be numerous reference groups or points of view. The several matrices at the lower left represent several alternate programs to which the one being evaluated can be compared.

Each set of absolute standards, if formalized, would indicate acceptable and meritorious levels for antecedents, transactions, and outcomes. So far we have been talking about setting standards, not about judging. Before making a judgment the evaluator determines whether or not each standard is met. Unavailable standards must be estimated. The judging act itself is deciding which set of standards to heed. More precisely, judging is assigning a weight, an importance, to each set of standards. Rational judgment in educational evaluation is a decision as to how much to pay attention to the standards of each reference group (point of view) in deciding whether or not to take some administrative action.[3]

[3] Deciding which variables to study and deciding which standards to employ are two essentially subjective commitments in evaluation. Other acts are capable of objective treatment; only these two are beyond the reach of social science methodology.

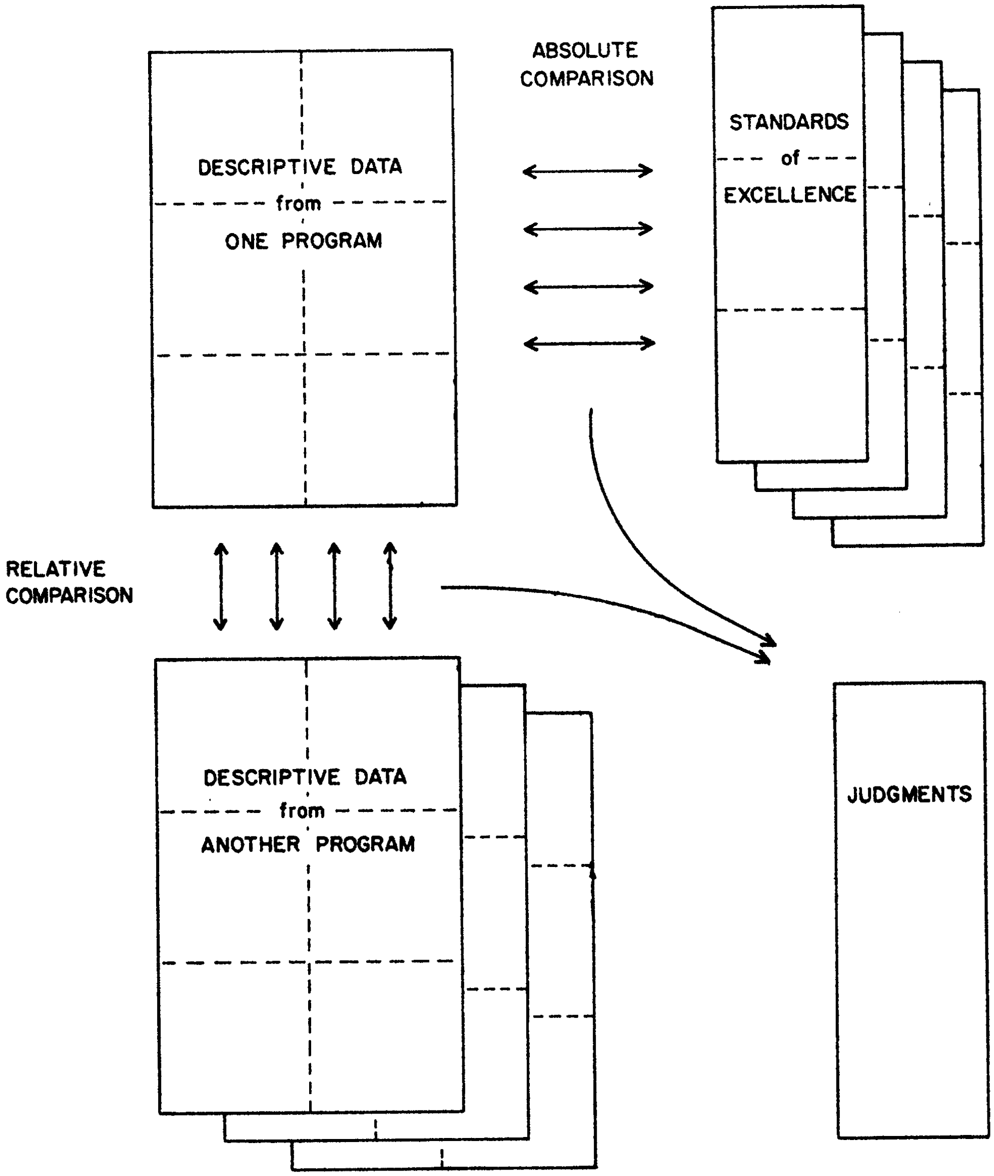

Figure 3. A representation of the process of judging the merit of an educational program.

Relative comparison is accomplished in similar fashion except that the standards are taken from descriptions of other programs. It is hardly a judgmental matter to determine whether one program betters another with regard to a single characteristic, but there are many characteristics and the characteristics are not equally important. The evaluator selects which characteristics to attend to and which reference programs to compare to.

From relative judgment of a program, as well as from absolute judgment we can obtain an overall or composite rating of merit (perhaps with certain qualifying statements), a rating to be used in making an educational decision. From this final act of judgment a recommendation can be composed.

As to which kind of evaluation - absolute or relative - to encourage, Scriven and Cronbach have disagreed. Cronbach (1963) suggests that generalizations to the local-school situation from curriculum -comparing studies are sufficiently hazardous (even when the studies are massive, well-designed, and properly controlled) to make them poor research investments. Moreover, the difference in purpose of the programs being compared is likely to be sufficiently great to render uninterpretable any outcome other than across -the -board superiority of one of them. Expecting that rarely, Cronbach urges fewer comparisons, more intensive process studies, and more curriculum "case studies" with extensive measurement and thorough description.

Scriven, on the other hand, indicates that what the educator wants to know is whether or not one program is better than another, and that the best way to answer his question is by direct comparison. He points to the difficulty of describing the outcomes of complex learning in explicit terms and with respect to absolute standards, and to the ease of observing relative outcomes from two programs. Whether or not Scriven's prescription is satisfying will probably depend on the client. An educator faced with an adoption decision is more likely to be satisfied, the curriculum innovator and instructional technologist less likely.

One of the major distinctions in evaluation is that which Scriven identifies as formative versus summative evaluation. His use of the terms relates primarily to the stage of development of curricular material. If material is not yet ready for distribution to classroom teachers, then its evaluation is formative; otherwise it is summative. It is probably more useful to distinguish between evaluation oriented to developer -author -publisher criteria and standards and evaluation oriented to consumer -administrator -teacher criteria and standards. The formative -summative distinction could be so defined, and I will use the terms in that way. The faculty committee facing an adoption choice asks, "which is best? Which will do the job best? " The course developer, following Cronbach's advice (1963), asks, "How can we teach it better? " (Note that neither are now concerned about the individual student differences). The evaluator looks at different data and invokes different standards to answer these questions

The evaluator who assumes responsibility for summative evaluation - than formative evaluation -accepts the responsibility of informing consumers as to the merit of the program. The judgments of Figure 3 are his target. It is likely that he will attempt to describe the school situations in which the procedures or materials may be used, He may see his task as one of indicating the goodness-of -fit of an available curriculum to an existing school program. He must learn whether or not the intended antecedents, transactions, and outcomes for the curriculum are consistent with the resources, standards, and goals of the school. This may require as much attention to the school as to the new curriculum.

The formative evaluator, on the other hand, is more interested in the contingencies indicated in Figure 2. He will look for covariations within the evaluation study, and across studies, as a basis for guiding the development of present or future programs.

For major evaluation activities it is obvious that an individual evaluator will not have the many competencies required. A team of social scientists is needed for many assignments. It is reasonable to suppose that such teams will include specialists in instructional technology, specialists in psychometric testing and scaling, specialists in research design and analysis, and specialists in dissemination of information. Curricular innovation is sure to have deep and widespread effect on our society, and we may include the social anthropologist on some evaluation teams. The economist and philosopher have something to offer. Experts will be needed for the study of values, population surveys, and contentoriented data-reduction techniques.

The educator who has looked disconsolate when scheduled for evaluation will look aghast at the prospect of a team of evaluators invading his school. How can these evaluators observe or describe the natural state of education when their very presence influences that state? His concern is justified. Measurement activity - just the presence of evaluators - does have a reactive effect on education, sometimes beneficial and sometimes not - but in either case contributing to the atypicality of the sessions. There are specialists, however, who anticipate that evaluation will one day be so skilled that it properly will be considered 66unobtrusive measurement" (Webb, 1966).

In conclusion I would remind the reader that one of the largest investments being made in U.S. education today is in the development of new programs. School officials cannot yet buy or borrow programs on a rational basis, and the needed evaluation is not under way. What is to be gained from the enormous
effort of the innovators of the 1960's if in the 1970's there are no evaluation records? Both the new innovator and the new teacher need to know. Folklore is not a sufficient repository. In our data banks we should document the causes and effects, the congruence of intent and accomplishment, and the panorama of judgments of those concerned. Such records should be kept to promote educational action, not obstruct it. The countenance of evaluation should be one of decision -making, not one of trouble-making.

Educators should be making their own evaluations more deliberate, more formal. Those who will -whether in their classrooms or on national panels - hope to clarify their- responsibility by answering each of the following questions: (1) Is this evaluation to be primarily descriptive, primarily judgmental, or both descriptive and judgmental? (2) Is this evaluation to emphasize the antecedent conditions, the transactions, or the outcomes alone, or a combination of these, or their functional contingencies? (3) Is this evaluation to indicate the congruence between what is intended and what occurs? (4) Is this evaluation to be undertaken within a single program or as a comparison between two or more curricular programs? (5) Is this evaluation intended more to further the development of curricula or to help choose among available curricula? With these questions answered, the restrictive effects of incomplete guidelines and inappropriate countenances are more easily avoided.

REFERENCES

American Council on Education. Educational Measurement. E. F. Lindquist (Ed.), Washington, D. C., 1951.

Atkin, J. M. Some evaluation problems in a course content improvement project. Journal of Research in Science Teaching, 1, 1963, 129-132.

Bassham, H. Teacher understanding and pupil efficiency in mathematics - study of relationship. Arithmetic Teacher 9: 383-87, 1962.

Berlak, H. Concepts and structure in the new social science curricula. Social Science Education Consortium Conference,' Purdue University, January, 1966. (Oral comment included in proceedings)

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

Dewey, J. Theory of Valuation. Chicago: University of Chicago Press, 1939.

Educational Policies Commission. The central purpose of American education. Washington, 1961.

Educational Testing Service. A long hot summer of committee work on National Assessment of Education. ETS Developments, Vol. XIII, November, 1965.

Gagne, R. M. Elementary science: a new scheme of instructions. Science, Vol. 151, No. 3706, 49-53.

Guba E. G. and Clark, D. L. An examination of potential change roles in education. Columbus, Ohio, 1965. (multilith)

Hand, H. C. National assessment viewed as the camel's nose. Phi Delta Kappa 47, September, 1965, 8-12.

Hastings, J. T. The why of the outcomes. Symposium paper on Measurement Problems in Curriculum Evaluation, AERA, February, 1965.

Mager, R. F. Preparing objectives for programmed instruction. San Francisco Fearson Publishers, 1962.

National Study of Secondary School Evaluation. Evaluative Criteria. Washington D. C., 1960.

Oliver, D. W. and Shaver, J. "Teaching students to analyze public **controversy:** a curriculum project report. " Authors, 1962.

Sriven, M. The methodology of evaluation. Bloomington: Indiana University, 1965. (mimeograph)

Smith, B. 0. and Meux, M. 0. A Study of the Logic of Teaching. Urbana: Bureau of Educational Research, University of Illinois, no date- (Trial Edition)

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

Taylor, P. A. and Maguire, T. 0. A theoretical evaluation model. University of Illinois, Department of Educational Psychology, 1965. (Mimeographed)

Tyler, R. W. Assessing the progress of education. Phi Delta Kappa , 47, September, 1965, 13-16.