DOCUMENT RESUME

ED 045 699 TM 000 260

AUTHOR Klein, Stephen

TITLE Evaluating Tests in Terms of the Information They Provide.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY Bureau of Elementary and Secondary Education (DHEW/OE), Washington, D.C.
REPORT NO Eval-Comm-2-2
PUB DATE Jun 70
NOTE 6p.

EDRS PRICE EDRS Price MF-$0.25 HC-$0.40
DESCRIPTORS *Achievement Tests, Criterion Referenced Tests, *Decision Making, Evaluation Techniques, Measurement Techniques, Norm Referenced Tests, *Objectives, Program Evaluation, Student Evaluation, *Test Construction, *Test Results

ABSTRACT

Despite their advantages over other assessment techniques, current achievement and ability tests are not especially efficient sources of information for the range of educational decisions for which they are used and relied upon. Two major types of tests, criterion-referenced and norm-referenced, and two types of use, student evaluation and program evaluation, are considered in this context. The strengths and weaknesses of criterion and norm-referenced measures are discussed in detail. A four-step approach to test construction is proposed, combining the better components of the criterion and norm-referenced approaches, which may overcome some of the information problems of current tests. This proposal entails the specification of measurement objectives, the development of test items for each objective, the development of test items to measure related objectives, and the provision of both a score and a score interpretation for each objective. Responses to potential criticisms of this approach, including its reliability and its usefulness to teachers, are made. It is concluded that tests should be evaluated in terms of the quantity, quality, and the cost of the information they provide. (PR)

June, 1970 Vol. 2, No. 2

UCLA Evaluation comment

CSE

Center for the Study of Evaluation

Statement of Intent

The Center for the Study of Evaluation, founded in June, 1966, is an educational research and development center sponsored by the U.S. Office of Education under Title IV of the Elementary and Secondary Education Act of 1965.

The Center, directed by Marvin C. Alkin, is a unique organization working exclusively on problems of educational evaluation and is devoted to three prime objectives: to develop a theory for the study of evaluation; to develop methods and instruments for measuring program effectiveness; and to provide a scientific basis for program and policy decisions in education. After an initial period of exploration, the Center's efforts have been increasingly focused on a relatively few research projects which fall within the scope of major program areas. This programmatic approach has resulted in combined efforts by specialists in various disciplines and in the development of a conceptual framework around which a comprehensive theory of evaluation can be built.

Evaluation Comment provides a forum for the discussion of significant ideas and issues in the study of evaluation of educational programs and systems. *Evaluation Comment* is especially interested in publishing creative or controversial ideas, concepts, and dialogue about evaluation of instructional programs that promise to improve knowledge about evaluation or, at least, to excite interest and comment from readers.

A copy of *Evaluation Comment* is distributed free of charge to each scholar, researcher, or practitioner on our mailing list. One to five additional copies may be obtained free of charge; however, where greater amounts are needed readers are encouraged to reproduce the Comment for their own purposes. To be included in our mailing list or to order, subject to availability, additional copies of *Evaluation Comment*, please write to:

James Burry, Managing Editor
Evaluation Comment
Center for the Study of Evaluation
145 Moore Hall
University of California, Los Angeles
Los Angeles, California 90024

Marvin C. Alkin, Editor

EVALUATING TESTS IN TERMS OF THE INFORMATION THEY PROVIDE[1]

Stephen Klein
University of California, Los Angeles

The decision to give a student a grade in a course, to admit him to college, to assign him to a particular educational curriculum, or to promote him is based almost entirely on his performance on ability and achievement tests. Such tests, also, are relied upon to provide information about the quality of educational programs and systems. For example, whether Project Headstart or a program for the gifted will be continued depends in large part on how well the students in the programs perform on the tests used in evaluating them. Thus, test results wield enormous power in educational decisions that determine the kinds of educational programs a student receives and the level and direction of his educational career. These, in turn, influence greatly his place in society.

Reliance upon and faith in the efficacy of testing have resulted mainly from the relative *efficiency* of tests as vehicles for providing *information for decisions* about

[1]The critical comments and reviews of Ralph Hoepfner, Ted Husek, Jason Millman, and James Popham were most appreciated in the development of the ideas presented in this paper.

students and the educational programs they receive. In other words, tests are almost always cheaper, quicker, fairer, and more valid and reliable information sources than are other assessment techniques such as interviews. By following this line of reasoning one step further, it becomes apparent that the value of a test is determined by the quantity, quality, and cost of the information it provides for educational decisions.

The two major points of this paper are: (a) current ability and achievement tests, whether constructed by test experts or teachers, are not especially *efficient* sources of information for the range of educational decisions for which they are relied upon; and (b) tests *can* be constructed that will be efficient for making such decisions. Before discussing these points, however, it is necessary to consider two aspects of tests — their *purpose*, i.e., *how* the information they provide will be used and, further, the *philosophy* underlying the manner in which they are constructed.

Any good educational measurement text describes the varied purposes of tests, such as selection, placement, etc. For the present discussion, however, we shall examine only two major types of uses. These are: (a) *student evaluation*, i.e., tests used in making decisions about individual students; and (b) *program evaluation*, i.e., tests used with groups or samples of students to provide information for decisions concerning educational programs that students might receive or are receiving.

Tests used in the first category, student evaluation, provide information about such things as whether a student has learned what he was supposed to have learned from a course or whether he has the prerequisite knowledge and ability needed for college. Tests used in program evaluation, on the other hand, are supposed to provide information about how well a program achieved or is achieving its objectives.

A second way of classifying tests is in terms of the *philosophy* underlying the manner in which they are constructed which, in turn, is reflected in how scores are reported. Here again, two major types of tests should be considered — norm-referenced versus criterion-referenced tests (Glaser, 1963). Popham and Husek (1969, p. 2) have noted the following differences between these kinds of measures:

"Norm-referenced measures are . . . used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device . . . Criterion-referenced measures . . . are used to ascertain an individual's status with respect to some criterion, i.e., performance standard." In the college selection situation, for example, the admissions officer has traditionally been concerned primarily with each applicant's *relative* likelihood of success. Norm-referenced data, such as high school grades and test scores on college admissions tests, have been the most successful predictors for this purpose. Criterion-referenced data, on the other hand, are very useful in determining whether an educational program achieved its objectives. The differential utility of these two kinds of data for various kinds of evaluation problems has led to developing norm-referenced measures for student evaluations and criterion-referenced tests for program evaluations.

This doubling of test development costs in order to handle different kinds of evaluation problems results in duplication of effort (such as developing essentially the same items twice where one really could serve for both purposes). It is also misleading to the typical test user who would expect that after such specialization, the final products would meet his student or program evaluation needs. The reason he is deceived is that neither the existing norm- nor criterion-referenced measures are adequate for either student or program evaluation. In other words, they do not provide all the necessary information for making either kind of evaluation decision.

Norm-referenced measures, for example, are very effective in ranking students (or groups of students) in terms of their ability, knowledge, or other salient characteristics. When constructed and standardized properly, they also provide a good basis of comparison between students or groups at different schools through the use of norm and percentile tables. Normative test data would be very useful for selection and promotion decisions and even for program evaluation decisions if one knew what a score on such measures really meant. Unfortunately, score conversions such as percentiles and stanines do not indicate either what the student has learned or at what level he will perform if he were promoted or admitted to college. This problem has led to applying criterion-referenced interpretations to norm-referenced data via the use of grade norms and predicted grade point averages. Despite the many difficulties associated with such "scores," they do contribute to the test user's understanding of the general level at which the student can perform. Predicted GPA's and grade norms fail, however, in describing precisely what the student has and has not learned. One implication of this situation is that some very different admissions decisions might be made if it were disclosed that even students with scores below the 25th percentile had the skills actually needed to be able to perform college work. A second type of problem with standardized norm-referenced tests is that they are likely to measure a somewhat different set of objectives than those stated for a specific educational program. For example, a score on a published science test may represent overall performance on 10 objectives; however, a given science program may be concerned with only four objectives, and just two of these may be included in the standardized test. Thus, the single total score on the science test, whether converted to grade norms or not, would not be an appropriate measure for evaluating the success of the science program. In brief, norm-referenced measures have been very useful in providing data about general performance levels needed for many student and program evaluation decisions, but very weak in contributing information regarding specific skill and knowledge development.

Criterion-referenced measures, on the other hand, complement their normative counterparts. They do this by adequately describing the specifics and what a test score

means. They do not, however, provide the often needed normative base for comparisons and interpretations. The relative strengths of criterion-referenced measures have led to their being relied upon for many program evaluation decisions. Some school districts even require independent educational firms supplying special training to specify in the contractual agreement the criterion levels at which students will perform.

The foregoing use of criterion-referenced measurement would be a laudable practice if one knew how to determine what criterion objectives to specify, or what level of performance constitutes their attainment, or how to interpret the results if the objectives are or are not achieved. To illustrate this point, let us suppose that a new course unit in 10th grade biology led to 30% of the students attaining all of the unit's 20 objectives, 50% of the students attaining 15 objectives, and only 20% of the students achieving less than 10 objectives. These results look very impressive and a school official might be very pleased with the effectiveness of the program. But would he still be happy if he discovered that most students could achieve 10 of these objectives before taking the unit, or that the criterion of attainment was 1 out of 5 items correct per objective, or that the items used to measure an objective were not truly representative of the range of items that might have been employed, or that 80% of the students at other schools (having students of comparable ability) attained all 20 objectives using a criterion of 4 out of 5 items correct per objective? One expects that the school official would make a rather different evaluative decision regarding the program's worth had this latter information been available to him. Clearly, grade norms or other kinds of normative based data would help clarify the actual utility and significance of the program in achieving its objectives.

Criterion-referenced measures, further, typically suffer from their being limited to the program's specific objectives. This may seem like a correct approach unless one asked such questions as: "If the student (or program) failed to meet an objective, did he (it) miss by an inch or a mile?" or "If two students achieved an objective, could one of them attain more advanced objectives?" The answers to these questions would certainly have a bearing on evaluation decisions dealing with the relative effectiveness of different programs and what subsequent educational treatments should be instituted (i.e., remedial or advanced). It should be noted, however, that these latter problems are not limitations of criterion-referenced measures per se, but of the way most of these measures are developed, scored, and interpreted.

To summarize, norm-referenced measures often provide useful information in evaluating the *relative* performance of students and programs with respect to general performance criteria. Their weakness mainly has been in failing to provide specific information about particular skill development and needs. Such information is necessary in making decisions regarding subsequent educational treatments and the effectiveness of a given program in achieving its limited set of objectives. Criterion-referenced measures, on the other hand, have the advantage of being able to provide the latter kinds of specific information. What they fail to do is provide a basis for interpreting fully what the attainment of an objective really means, i.e., whether it is significant and important or trivial and unnecessary.

The foregoing discussion is the basis for the first major point of this paper, namely: despite their comparative advantages over other assessment techniques, typical tests are still not especially efficient sources of information for the full range of educational decisions for which they are used and relied upon. Let us now turn to the second point, namely: tests can be constructed, and the results they provide can be reported in a way that will facilitate making such decisions. The basis for this new path is the obvious but generally unutilized technique of combining the better components of the norm- and criterion-referenced approaches. The essential characteristic of this approach is that it includes the concepts of item difficulty and normative score reporting in the development and interpretation of criterion based measures. This would entail the following steps:

1. *Specification of objectives.* The objective(s) each test is supposed to measure should be stated clearly. Popham (1965) and others have prepared excellent guides as to how objectives should be written. The decision as to which objective(s) to measure may be a difficult one, but literature reviews, research studies, professional judgments, and related sources of information should help clarify just what kinds and levels of performance should be assessed.

> Sample Objective: The student can add two numbers each of which is more than 9 but less than 100.

The level of generality at which an objective is stated is, of course, an important consideration. Some guidelines that may help in determining this level are (a) it usually is a good idea to have at least three items per objective; thus, one certainly should not have more objectives than the number of items on which he can collect adequate data. To achieve this end, one can either reduce the number of objectives measured during a testing session or broaden the statements of the objectives so that they include sub-objectives; (b) write the objectives at a level of generality that will be interpretable to the person who has to use the test results. It helps the test constructor to be specific, but too many specifics may make the data uninterpretable to the user unless he is at least provided with more general statements.

2. *Develop test items for each objective.* A clear statement of the objective will provide a very good guide to both the type of item and performance level(s) (i.e., item difficulties) needed to measure that objective. Guides for developing test items are readily available (e.g., Ebel, 1965; Wood, 1961) and should be followed to assure that items measure the objective, are appropriate for the students to be tested, and are feasible to administer in a cost effective manner. It is especially important, however, that the items selected for an objective be a good, rep-

resentative sample of the total population of items that might be used to measure that objective. This sampling should cover both the range of formats that might be used and the range of item difficulties.

Sample items: 10 + 20 =

38

+97

Thus, if the objective was that the student could add two numbers each of which was less than 100, it would not be a good idea just to use items involving the addition of one digit numbers. The primary reason for including items that span the difficulty levels within an objective is that differences between students and programs could be assessed more accurately. This would occur because a student's score on an objective would reflect the degree to which he attained it; and it is a well known fact that students do differ in their performance even with the same instruction because they differ in their ability to profit from it. Including items, then, that span the range of difficulties would eliminate the practice of setting arbitrary cutoff points to assess whether a student (or program) did or did not achieve an objective, since the percentage of items correct on the objective would provide an indication of the degree of attainment.

3. *Develop test items to measure related objectives.* In the case of a series of en route objectives, it is important to include items that measure performances that come before and after the one(s) being studied. By the same logic, it is equally important to assess performance both on objectives that are easier and more difficult to master than just the one(s) of major interest.

Sample items: 24 + 36 + 89 =

8

+25

The reasons for measuring these kinds of related objectives are that they (a) provide information about the unanticipated outcomes of educational programs, (b) indicate how close a program (or student) came to meeting or surpassing the objective(s), and (c) show the level at which subsequent educational treatments should be pitched. For example, the students' improvement in addition may have surpassed the stated objectives of an experimental mathematics program, but on further inspection, it might be revealed that this performance was obtained at the expense of proficiency in subtraction. Thus, even though the experimental program may not have been concerned with subtraction per se, it was important to assess it if one wished to evaluate fully the quality of this program.

One by-product of this approach is the information gained regarding the actual difficulty or learning sequence of various objectives. For example, if students perform better on items measuring a supposedly "advanced" or terminal objective than they do on the ones presumed to lead up to it, then the assumptions regarding the ordering of objectives might merit reappraisal.

4. *Provide a score and score interpretation for each objective.* This information should reflect both criterion- and norm-referenced performance on the items designed to measure the objective. A sample interpretation might read as follows: "Donald Jones (or Program #3) got four of the six items correct on objective number 7 (addition of whole numbers less than 100). Approximately 80% of the other students in Donald's class did this well. Students of equal ability in other classes (or programs) only got one-third of the items correct which is typical of the second graders in this state (i.e., the median score statewide on this objective is 33% correct)." This type of interpretation allows the reader to know what the student can and cannot do and also provides him with a frame of reference for interpreting this level of performance.

Before discussing this approach further, it is important to clarify what is meant by an educational "objective" and how its level of generality influences the way it is measured. An "objective" describes the type and level of performance a student might achieve. Very explicit statements of objectives, such as "the student can add two numbers each of which is less than 100," are termed "behavioral" (Popham, 1965) and refer to a relatively narrow range of performance levels. On the other hand, global objectives or goals, such as "the student can perform basic arithmetic computations," are less precise and encompass a wider range of performance levels (e.g., "1 + 1 = ?" vs. "39 ÷ 17 =?").

It is apparent, therefore, that the broader the objective to be assessed, the more items are needed to cover its full range of performances. The major implication of this situation for test construction is that the measurement of a global objective involves the assessment of several subobjectives. Since information about both types of objectives is often needed, scores should be reported for both. For example, "Joe's score in arithmetic computations was 18, which he obtained by scores of 6 in addition, 5 in subtraction . . ." The interpretation of these scores would, of course, require clear statements of the objectives along with the criterion- and norm-referenced information described in Step #4.

Now let us examine how the suggested four-step test construction approach differs from present practices. An inspection of current tests and manuals indicates that most publishers of standardized achievement tests usually claim to go through the first two steps of specifying objectives and writing items to measure them. However, they rarely provide scores on each of the objectives or content areas that their tests (or subtests) purport to measure. For example, the 55 items in the mathematics achievement test of the Cooperative Primary Tests (ETS, 1967) are supposed to assess the following eight concepts: Number, Symbolism, Operation, Function and Relation, Approximation, Measurement, Estimation, and Geometry.[2] However, only one score is provided for the 55 items. This problem is demonstrated in Table 1 where it can be seen

[2]The ETS test was chosen for analysis because it exemplified a common problem with most standardized achievement tests and the data were readily available in the test manual.

that two objectives (Number and Geometry) account for 42% of the test's items and have a mean item difficulty of .74; and another two objectives (Operation and Measurement) account for 38% of the items but have a mean item difficulty of .55. Dispersion of scores on the test, therefore, is obtained by having different difficulty levels for different objectives rather than by building in a broad range of difficulties within each objective. The implication of this test construction technique is that students who get high scores can achieve different kinds of objectives rather than just perform better than low scoring students on the same objectives. The test manual, on the other hand, implies that the student's score reflects his ability to master the eight objectives. Supplying the norm- and criterion-referenced information described in Step #4 above for each objective would indicate when tests are constructed in this manner. It would also be a major step towards helping to individualize instruction by showing the strengths and weaknesses of each student or program and clarifying what a test is really measuring and how it is doing it.

TABLE 1

Analysis of the Mathematics Achievement Test (Form 12a) of the Cooperative Tests of Primary Mental Abilities.

Content Area	Number of Items	Mean Item Difficulty
Number	17	.73
Symbolism	6	.66
Operation	13	.55
Function and Relation	3	.59
Approximation	1	.32
Measurement	8	.54
Estimation	1	.65
Geometry	6	.75

At this point, one wonders why the field of educational measurement has been so slow to incorporate the better characteristics of norm- and criterion-referenced tests and score reporting into a single package. One reason might be interdisciplinary rivalry ("lack of professional communication") between psychologists, who tend to develop norm-referenced tests and certain educationists, who prefer constructing criterion-referenced tests. A second reason may be the unwarranted fear of the usual criticisms of subscores on tests. In other words, many existing tests providing subscores fall victim to one or the other of the following two problems: high subtest score intercorrelation, i.e., the subtest scores are so highly related to each other as to make them indistinguisable; or unreliability due to the brevity of the subtest, i.e., the number of items in it.

The latter two criticisms can be dismissed by applying the principle that the utility of a test (or test score) should be evaluated in terms of the information it provides. For example, if the first situation occurred, i.e., high inter-score correlation, it would mean that either the items going into each score were providing essentially the same information because they were measuring the same thing (and, thus, should be combined into a single score) or performance on two objectives was similar because the students had equivalently good or poor training in both areas. In most instances, a simple experiment in which training is given on only one objective would clarify whether the subtest scores were really measuring different objectives. In other words, if the scores on all objectives improved equally after instruction, it is probably safe to assume that the subtests are measuring the same rather than different objectives (Husek, 1969). Thus, high inter-score correlations should disappear with differential learning if the subscores really provide information about what was and was not learned.

The second criticism, unreliability due to brevity of the subtests, is even less tenable. In the case of program evaluation, for example, a subtest can be lengthened easily by using item sampling techniques (Lord & Novick, 1968) thereby improving a test's reliability (the formula for this increase as a function of length may be found in any measurement text, e.g., Cronbach, 1960). In other words, a given student need only take a few items while other students receive a different set. Such item sampling procedures keep test length the same for a given student, but substantially increase the total number of items providing reliable information about how well a program is meeting its objectives. In the case of student evaluation, however, there is no substitute for highly reliable and valid information if one has to use that data in making a decision about a student's performance. But even relatively unreliable subscores are still valuable, since they could be used diagnostically to locate possible problem areas for further testing.

Another potential criticism of the proposed four-step approach is that teachers cannot use it in constructing their own tests. Teachers do not have the time or expertise to write clear, relevant objectives and/or good items to measure them (Thorndike, 1969). While this is true, it is also true that teachers can be relieved almost entirely of this chore by test experts. This idea may at first seem heretical to many educators, but on further reflection they will realize that experienced and trained item writers can do this job better than teachers. What is needed, therefore, is an atlas of objectives with sets of items (short tests) for each objective. This atlas should organize the objectives and their levels by such things as difficulty and the type of cognitive functioning required (e.g., Bloom, 1956; Guilford, 1967) and, where possible, include norm- and criterion-referenced interpretations of scores. With this tool, the teacher could select those objectives he wished to measure along with the necessary related objectives and the short tests needed to assess student performance on them. Teachers would, of course, have to help in this test development as well as construct items for those objectives not included in the atlas. In fact, frequent use of the atlas for course tests and quizzes may

even eliminate the need for the classical standardized achievement test since all the information (and more) would have been collected via the course examinations. Such atlases can be a reality and their development is already underway (e.g., PROBE, American Book Company, CTB/McGraw-Hill).

As noted in the beginning of this paper, tests should be evaluated in terms of the quantity, quality, and cost of the information they provide. It is the premise of this paper that the test construction and score reporting procedures outlined above will provide far more information than is being supplied by most currently available tests. The cost of developing these new procedures would be somewhat greater initially than current methods. There would be a savings, however, deriving from the use of a single set of tests for a variety of purposes and from reduced testing time and teacher involvement in test construction. It seems likely, therefore, that any additional development costs would be offset by the substantially greater quality and quantity of relevant information provided. As test publishers try this route, we can easily observe its merits by measuring the depth of the path beaten to their doors by people who have to use test data in making decisions.

REFERENCES

American Book Company. *Reading experience and development series tests.* New York: Author, 1969.

Bloom, B. S. (Ed.) *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York: David MacKay, 1956.

Cronbach, L. J. *Essentials of psychological testing.* (2nd ed.) New York: Harper Brothers, 1960.

CTB/McGraw-Hill. Del Monte Research Park, Monterey, California.

Ebel, R. L. *Measuring educational achievement.* Englewood Cliffs, New Jersey: Prentice Hall, 1965.

Educational Testing Service. *Handbook: Cooperative primary tests.* Princeton, New Jersey: Author, 1967.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist,* 1963, *18,* 514-521.

Guilford, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

Husek, T. Different kinds of evaluation and their implications for test development. *Evaluation Comment,* 1969, 2 (1), 8-10.

Husek, T., & Popham, W. J. Implications of criterion referenced measurement. *Journal of Educational Measurement,* 1969, *6,* 1-9.

Lord, F. M., & Novick, M. *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley, 1968.

Popham, W. J. *The teacher-empiricist.* Woodland Hills, California: Aegeus, 1965.

PROBE. *Instructional objectives and items exchange.* Center for the Study of Evaluation, University of California, Los Angeles, California.

Thorndike, R. L. Helping teachers use tests. *Measurement in Education,* 1969, 1 (1). National Council on Measurement in Education.

Wood, A. D. *Test construction.* Columbus, Ohio: Merrill, 1961.

IOX "Spin-off"

It was announced recently by Dr. Marvin C. Alkin, Director of CSE, that the Instructional Objectives Exchange (IOX) will "spin-off" from the Center as a separate, non-profit corporation effective May 31, 1970. IOX, which was conceived by the Center, has been a part of the Project for Research on Objective-Based Evaluation (PROBE). The new IOX corporation, under the leadership of Drs. W. James Popham, Eva Baker, and John McNeil, will be devoted to the collection, processing, and distribution of instructional objectives. PROBE, however, will continue as a Center project and will be devoted to the study of the conditions and form most appropriate to objective-based evaluation.

The decision for IOX to emerge as a separate entity was necessitated by two important considerations: (1) the services and activities of IOX are vital to the academic community and are necessarily of a continuing and "service" nature and (2) the Center must be able to continue to develop other implementation systems. The developmental and diffusion activities associated with the use of instructional objectives, therefore, will be conducted by the IOX corporation, while the research activities relevant to the use of instructional objectives in an evaluation setting will be conducted by PROBE under CSE auspices.

In this way, the "spin-off" of IOX will both allow its specific services to continue and enable the Center to devote more time and resources to the actual study of objective-based evaluation.