DOCUMENT RESUME

ED 050 154 TM 000 544

AUTHOR Davis, Frederick B.
TITLE Criterion-Referenced Tests.
PUB DATE Feb 71
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Academic Performance, *Criterion Referenced Tests, *Diagnostic Tests, Individual Differences, *Individualized Instruction, Instructional Materials, Norms, *Predictive Ability (Testing), Reliability, Scores, Standardized Tests, *Test Construction, Testing, Tests

ABSTRACT

Confusion has arisen because tests are described as criterion-referenced or norm-referenced. Generally, these terms should apply to scores and not to tests since either type of score may be obtained for any test. Various terms such as absolute scores, fixed-standard scores and mastery-test scores may be more appropriate substitutes for criterion-referenced scores. Mastery-test scores grow out of the historical development of instructional tests allowing the student to demonstrate that certain prescribed skills and practices had been learned. With the advent of individualized instruction in the 1920's, diagnostic tests were developed to determine the already established level of accomplishment. Because instructional materials and accompanying diagnostic and mastery tests were not made generally available, individualized instruction was abandoned in the schools till the 1950's. Today, modern test theory can provide many guidelines to the content validity, length, item format, and scoring of mastery tests. In conclusion, mastery and diagnostic tests should supplement standardized survey tests in educational evaluation; there need be no problem of choosing between them. (CK)

CRITERION-REFERENCED TESTS

Frederick B. Davis

University of Pennsylvania

A criterion-referenced test has been defined as "a measuring instrument deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards."[1] The interpretation of an examinee's score is wholly independent of the performance of other examinees in a "norm group" representative of some defined population. Ordinarily, scores are expressed as the number of items correct or the percentage of items correct.

[1] Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In Thorndike, R. L., ed. Educational measurement. Washington: American Council on Education, 1971. (In press)

At this point, it is important to consider whether a test properly constructed and scored in the manner described could be administered to samples of pupils representative of populations in which its use would be appropriate and whether percentile ranks could be assigned to each raw score in each of the populations sampled. Obviously, this could be done and norm-referenced score interpretations could be made. Clearly, then, it is not the test itself that determines whether scores from it may be norm-referenced. Consequently, it might be wise to avoid describing tests as "criterion-referenced" or "norm-referenced." If we are to use these terms at all, they should be applied to scores, not to tests. The fact is that either type of score may be obtained for any test. Certain principles of test theory indicate when either type is appropriate for a given test.

Although the term "norm-referenced scores" described reasonably well what

it is intended to describe, there are persuasive reasons why the term "criterion-referenced scores" should be abandoned. First, the terms "criterion-referenced scores" and "norm-referenced scores" dichotomize all scores; hence, their use implies strongly that a test from which the former are derived has been carefully constructed to measure some defined criterion variable while a test from which the latter are derived has not been. In other words, educators and laymen are likely to infer that tests yielding criterion-referenced scores have higher "content validity" than tests yielding norm-referenced scores. This inference is categorically unjustified since any test can yield either type of score and since the content validity of a test is dependent mainly on the care and skill employed in designing and writing items for it and by the nature of the variable measured by it.

Second, as Glaser and Nitko have pointed out, many people confuse criterion-referenced tests with tests yielding scores that have been correlated with an external criterion or with several such criteria in order to estimate the predictive validity coefficient or coefficients of such scores.[2]

---

[2]Glaser, R. and Nitko, A. J., loc. cit.

Among the terms that come to mind to replace "criterion-referenced scores" are "fixed-standard scores," "absolute scores," and "mastery-test scores." Of these, "fixed-standard scores" might be commonly confused with standard scores or normalized standard scores (like T-scores). The term "absolute scores" suggests that a true zero point has been established for the variable being measured, which is an unlikely accomplishment in educational measurement. "Mastery-test scores" is a phrase that grows out of the historical development of instructional tests used informally in the classroom and coincides with what

Glaser and Nitko appear to mean by criterion-referenced scores. They have stated that "the instructional process requires information about the details of the performance of the learner in order to know how instruction should proceed.... When this performance has been attained by an individual learner to the degree required by the design of the instructional system, then the learner is said to have attained mastery of the instructional goal."[3]

---

[3] Glaser, R. and Nitko, A. J., loc. cit.

Henceforth in this paper I'll use the term "mastery-test scores" in place of "criterion-referenced scores."

Norm-referenced scores are used primarily to compare the performance of one examinee with that of others in a representative sample of some defined relevant population. They are less frequently used to differentiate among examinees in a sample; consequently, terms like "differentiation scores" or "differential scores" are not maximally appropriate. Instead, I'll use the phrase "comparison scores" in place of "norm-referenced scores."

Since time immemorial, teachers have, with varying degrees of success, measured the level of performance of their pupils on material or processes that have just been taught by means of tests that meet Glaser and Nitko's definition of what the latter call criterion-referenced tests. In 1864, for example, Chadwick wrote that the Reverend George Fisher had prepared a book called the Scale Book, "which contains the numbers assigned to each degree of proficiency in the various subjects of examination.... The numerical values for spelling...are made to depend upon the percentage of mistakes in writing from dictation sentences from works selected for the purpose, examples of which are contained in the 'Scale Book' in order to preserve the same standard of

difficulty."[4] By the 1920's, the logic of individualizing instruction to

[4] Chadwick, E. Statistics of educational results. The Museum, a Quarterly Magazine of Education, Literature, and Science, 1864, 3, 480-4.

give every pupil the time and instruction needed to bring him to a predetermined level of accomplishment led to the development and use of diagnostic tests to guide instruction and of mastery tests to permit demonstration that certain prescribed skills and principles had been learned. The Winnetka Plan, the Morrison Unit-Mastery Plan, and the Dalton Plan made provision for frequent testing to make sure that pupils mastered the performance of specified skills or tasks at a predetermined level. In the Dalton Plan, you will recall, each pupil signed a contract to reach certain specified competencies in a given unit and was allowed to go on to the next unit only after he had demonstrated this level of competence on a mastery test.

Because instructional materials and accompanying diagnostic and mastery tests were not made generally available, these plans for individualizing instruction were generally abandoned in most schools. The majority of teachers simply lack the skill and the time required to formulate performance standards and to construct the hundreds of short diagnostic or mastery tests needed to guide individualized instruction in fairly large groups and to evaluate each pupil's performance with respect to these standards. Fortunately, as programed courses of study became available during the 1950's that were made up of learning exercises revised experimentally to teach efficiently the competencies that constitute their behavioral objectives and subobjectives, short diagnostic and mastery tests were keyed to each step in the instructional process. These yield raw scores (usually number of items answered correctly) that are linked directly

to performance standards determined in advance. Teaching, learning, and evaluation are woven together in such a way as to maximize the effectiveness of instruction for each individual pupil. Fears that these developments will stifle teacher initiative and professional development have been expressed. But these need not be justified. On the contrary, the teacher's role as a guide to individual learning activities, as a motivating agent, and as a classroom manager to engender an atmosphere conducive to learning can become more rewarding and more challenging than before.

Properly planned programs of evaluation should combine the frequent use of short diagnostic and mastery tests with the occasional use of standardized achievement tests, interest inventories, and specialized aptitude tests. Each type of test supplements the others. For what it may be worth, it is my opinion that many schools now use too few short diagnostic and mastery tests for instructional purposes and too many standardized tests. The reason for this is simply that most teachers do not have access to a supply of diagnostic and mastery tests keyed to the specific objectives of their instruction. I can see no practical solution to this problem short of creating and making available complete packages of behavioral objectives, instructional materials and procedures, and short diagnostic and mastery tests keyed to the objectives and prefiled in convenient, long-lasting cabinets. One part of this package without the others is nearly useless. Furthermore, as the introduction of Project PLAN has already shown, teachers must be tactfully and consistently guided in the use of such packages in their classrooms.

I should point out, however, that use of these packages for individualizing instruction and guiding learning will not prevent comparisons of the school achievement of different pupils. Say, for example, that the arithmetic curriculum in City A is organized for the first six years of schooling into

carefully planned units of work leading to the attainment of 1,000 behavioral objectives. No pupil ever "fails" in arithmetic; every one spends as much time as he needs to attain each objective as it comes in the ordered sequence. At the end of two years a few pupils would have attained 400 or more objectives; others would have attained only 100 or fewer objectives. Parents are kept informed from time to time about the progress of their children in arithmetic by reports indicating, among other things, the number of objectives covered. If this information is not provided by the school officially, parents will compare notes and makeestimates of their own. Naturally, they will ask teachers questions like, "Why has Sally Brown covered 200 objectives in arithmetic whereas my son has covered only 70 objectives in arithmetic? How many objectives should he have covered?" Inevitably, in one way or another differences in the number of objectives covered take on normative significance to parents and pupils alike.

The more instruction is individualized and made efficient, the more noticeable individual differences in rate and capacity for learning will become. Educators must accept this fact and deal with it. One solution would be the sort of thing that some labor unions have adopted. A skilled man who works rapidly and efficiently is simply informed in one way or another to get back into line and conform to an acceptable display of ability. Another solution is to encourage diversity and the display of talent by providing a wide range of ways in which pupils can distinguish themselves and gain self-esteem.

This paper may perhaps best be concluded by discussing briefly the guidance that modern test theory can provide with respect to evaluation instruments like mastery tests. Specifically, what does test theory have to say about:

1. How to maximize the content validity of mastery tests;
2. How to make mastery-test scores legitimately interpretable in terms of specified performance standards;

3. How reliability coefficients and accuracy of measurement can be estimated for mastery-test scores;

4. How to evaluate the likelihood and seriousness of errors in determining whether a pupil has truly met predetermined standards of performance for any given instructional objective;

5. How long mastery tests need to be;

6. What considerations influence the format of mastery-test items and how they should be scored.

First, the content validity of mastery-test scores can be maximized by conscientiously carrying out the conventional first step in the design of any achievement test. A detailed test outline must be prepared listing the specific objectives and subobjectives of the instructional unit to be evaluated. These must be expressed in terms of observable behaviors, to each of which one or more test exercises can be keyed. The display of substantive knowledge, skills and processes, attitudes, and feelings should be included, as required, in the populations of behaviors to be sampled by items.

Sampling the population of possible items for testing a specific objective may in practice, be carried out by approximation procedures. For example, Glaser and Nitko mention the fact that the population of problems in the addition of 3, 4, and 5 addends with the restriction that each addend shall be a single-digit integer from 0 through 9 consists of 111,000 different problems. Proposals for rules to be followed in creating the desired number of items from a huge population have been discussed by several investigators. In evaluating these proposals, item writers should recognize that the true tetrachoric intercorrelations of item scores (usually pass or fail) of items drawn from the population of items covering any narrowly delimited objective will be close to unity. Therefore, minor deviations from a perfectly random sample of items are

not likely to affect seriously a test's content validity.[5] It is important,

[5] Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

however, for the test outline to specify the extent to which the direct efforts of instruction and its transfer to analogous materials are to influence the test variance. For example, if a spelling rule is taught, its application to the words used in the instructional process is likely to be displayed better than its application to other words to which the rule also applies.

To make legitimate the interpretation of number-right scores, corrected raw scores, or per-cent-correct scores on any test, the content of the test must be homogeneous; that is, all of the items must measure the same variable (plus chance, of course). Such a test is said to be univocal. If a test is made up of a weighted composite of different skills, its raw scores do not properly represent successive levels of performance in any single objective. Consequently, when a pupil obtains less than a perfect score, the teacher cannot, on the basis of that score alone, determine what specific content or process he has not learned adequately. This situation and the uses to which mastery-test scores are put lead to the conclusion that such tests should be univocal. These considerations also indicate that many separate mastery tests are needed, and that for practical reasons they should be as short as possible. Since their reliability coefficients depend largely on their length, it is apparent that efficiency of measurement (i.e., reliability per unit of time) is at a premium in such tests.

Whenever decisions are made wholly or partly on the basis of test scores, the frequency with which these decisions are in error becomes a matter of concern.

This is partly because we want to be fair to the pupil and partly because errors lead to inefficiency in the instructional process. The errors can take two forms when we are using mastery-test scores to determine whether to advance a pupil to the next unit or to reteach the unit on which he has been tested: First, we can advance him when he should be held back; Second, we can hold him back when he should be advanced. The incidence of such errors depends partly on the reliability coefficient of the determinations. Consider the reliability coefficient of scores on a 5-item test of skill in getting the main thought of five reading paragraphs that I administered to 421 college freshmen in 1940. Every examinee answered every item. The mean score was 2.97 items answered correctly; the variance of these scores was 1.21; the reliability coefficient was .18, and the standard error of measurement for any single score drawn at random from the 421 obtained was 1.00. Thus, an examinee who scored 3 points could easily have a true score anywhere between 2-4 points. The data show the caution with which scores from short tests have to be interpreted. If we are interested only in separating the examinees into two groups: (1) those who obtained scores of 0-4, inclusive; and (2) those who obtained scores of 5 and are judged to have reached the predetermined level regarded as adequate for advancement to the next unit of instruction, the reliability coefficient for determining into which of the two groups each pupil belongs is .66, the cut-off score being 4.5. The procedure used to estimate this reliability coefficient for the "advance-no advance" determinations was recently provided by Livingston.[6] The result is in harmony with classical test theory. In general,

[6] Livingston, S. A. The reliability of criterion-referenced measures. Baltimore: Center for the Study of Social Organization of the Schools, The Johns Hopkins University, Report No. 73, July 1970

the greater the difference between the cut-off score and the mean of the entire group, the more the reliability coefficient of the "advance-no advance" determinations will exceed the conventional reliability coefficient of the scores. Since Livingston has also shown that reliability coefficients for dichotomic determinations (made by whole-number cut-off scores) vary with test length as predicted by the Spearman-Brown formula, we can estimate the number of items like those in the 5-item test that would be required to produce determinations of any desired reliability.

If such determinations were the only basis for irrevocable placements of long-term importance to the pupils, we should insist on a reliability coefficient of the determinations that would be above .90. But the penalty for misplacing a pupil at the end of a unit of instruction is not great because the decision can soon be changed by a teacher who observes his performance and each unit is likely to be short. Nevertheless, any errors of placement lower the over-all efficiency of the instructional process so we want to hold their incidence to some acceptably low percentage, such as five out of every hundred decisions. Procedures for accomplishing this are well known. On the basis of the illustrative data that I have cited and other data of this kind that are available to me, I would hazard a guess that the majority of mastery tests would yield dichotomic classifications with acceptable accuracy if the tests were made up of 20-30 items.

If provisions can be made to score mastery tests by hand by qualified professional personnel (such as the classroom teachers themselves), the task of item writing is greatly simplified because a variety of item formats, including free-response questions, can be used. This freedom is especially helpful for making tests for use in the elementary school with children below the age of 11.

Since examinees ordinarily have a chance to try every item in classroom tests, the conventional correction for chance success will not alter the rank order of number-right scores. However, when true-false items or multiple-choice questions with as few as 2-4 choices are used, corrected scores ordinarily provide considerably better estimates of the per cent of the population of items sampled that is actually known by a pupil than are provided by number-right scores. It would be of interest to investigate the extent to which partial knowledge and misinformation balance each other in the conventional correction formula when it is used with mastery tests of the type we have been discussing. Very little information is available about this matter and analytic formulations are not helpful.

In conclusion, it seems safe to say that mastery and diagnostic tests supplement standardized survey tests in educational evaluation. Each type serves an important educational need better than other types. Educators, therefore, are not faced with the problem of choosing between them but should concentrate their efforts in using all evaluation instruments to maximum advantage as needs for them appear.