DOCUMENT RESUME

ED 073 143 TM 002 389

AUTHOR	Davis, Frederick B.
TITLE	Criterion-Referenced Measurement.
INSTITUTION	Educational Testing Service, Princeton, N.J.; ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	ERIC-TM-17
PUB DATE	Feb 73
NOTE	8p.; 1972 Aera Conference Summaries
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bibliographies; Conference Reports; *Criterion Referenced Tests; *Educational Research; *Evaluation Methods; *Measurement Instruments; Speeches; Tests

ABSTRACT

Eight papers on various aspects of criterion-referenced measurement presented at the 1972 AERA Conference are reviewed. A list of the papers reviewed, their authors, and, when applicable, the ED numbers is provided. In addition, 10 references considered as valuable are provided. (DB)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

TM Report 17

February 1973

1972 AERA Conference Summaries

CRITERION-REFERENCED MEASUREMENT

Frederick B. Davis
University of Pennsylvania

This publication was prepared pursuant to a contract with the National Institute of Education, U. S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy

PREVIOUS TITLES IN THIS SERIES

1. Developing Criterion-Referenced Tests
   ED 041 052

2. Test Bias. A Bibliography
   ED 051 312

3. Ability Grouping: Status, Impact, and Alternatives
   ED 052 260

4. Developing Performance Tests for Classroom Evaluation
   ED 052 259

5. Tests of Basic Learning for Adults: An Annotated Bibliography
   ED 058 274

6. State Educational Assessment Programs: An Overview
   ED 058 309

7. Criterion Referenced Measurement: A Bibliography
   ED 060 041

8. Measures Pertaining to Health Education: I. Smoking
   ED 060 042

9. Measures Pertaining to Health Education: II. Drugs. An Annotated Bibliography
   TM 002 078 (ED Number not yet available)

10. Measures Pertaining to Health Education: III Alcohol. An Annotated Bibliography
    TM 002 079 (ED Number not yet available)

11. 1971 AERA Conference Summary
    I. Evaluation: State of the Art
    ED 060 043

12. 1971 AERA Conference Summary
    II. Criterion Referenced Measurement
    ED 060 134

13. 1971 AERA Conference Summary
    III. Educational Statistics
    ED 060 133

14. 1971 AERA Conference Summary
    IV. Test Development, Interpretation, and Use
    ED 060 135

15. 1971 AERA Conference Summary
    V. Innovations in Measurement
    ED 060 044

INTRODUCTION

About 700 of the 1,000 papers presented at the 1972 AERA Annual Meeting in Chicago, Illinois were collected by the ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TM). ERIC/TM indexed and abstracted for announcement in *Research in Education (RIE)* 200 papers which fell within our area of interest–testing, measurement, and evaluation. The remaining papers were distributed to the other Clearinghouses in the ERIC system for processing.

Because of an interest in thematic summaries of AERA papers on the part of a large segment of ERIC/TM users, we decided to invite a group of authors to assist us in producing such a series based on the materials processed for RIE. Four topics were chosen for the series: Criterion Referenced Measurement, Evaluation, Statistics, and Test Construction.

Most papers referred to in this summary may be obtained in either hard copy or microfiche form from.

ERIC Document Reproduction Service (EDRS)
P.O. Drawer 0
Bethesda, Maryland 20014

Prices and ordering information for these documents may be found in any current issue of *Research in Education.*

CRITERION-REFERENCED MEASUREMENT

Frederick B. Davis

This review covers eight papers on various aspects of criterion-referenced measurement presented at the annual meeting of the American Educational Research Association held in Chicago in April 1972. Because of great concern for accountability in education and renewed attention to the individualization of instruction, wide interest has been aroused in criterion-referenced measurement. Research workers have, therefore, devoted increasing efforts to this topic. As a result, many fine theoretical papers and reports of experimental studies are beginning to appear in the journals.

In a paper entitled "A Classical-Test-Theory Approach to Criterion-Referenced Tests," Samuel A. Livingston has discussed extensions of classical procedures to estimate intercorrelations, reliability coefficients, and validity coefficients for scores that are used only to separate examinees into two categories: those scoring at or above a designated "passing mark" and those scoring below that point. To Livingston, all tests that yield scores used in this way are criterion-referenced tests. He says, "... a criterion-referenced test is any test for which the user wants to compare each student's score not with the mean of some group, but with a specified criterion score, which does not depend on the scores the students actually obtain on the test .... This definition implies that all of the items on the test must measure the same thing—otherwise, it makes no sense to specify a single criterion score."

Whether one accepts or rejects Livingston's definition of criterion-referenced tests does not mean that one may not recognize the value of Livingston's procedures for estimating the intercorrelations, reliability coefficients, and validity coefficients of sets of dichotomous (pass-fail) scores. For the necessary computing equations and their rationales, the reader should consult previous articles by Livingston (1972a, 1972b).

Intuitively, it makes sense that the reliability of pass-fail scores should be higher than the reliability of raw scores derived from the same multi-item test because variations in the rank order of examinees *within* each of two dichotomous groups from one equivalent form of the test to any other equivalent form (while the true scores of the examinees remain the same) have no effect on the pass-fail status of any examinee. In fact, "perfect" dichotomous-group reliability would be reached as long as score variations from form to form took place *within* groups and did not cause any examinees to shift from one dichotomous group to another. The price of gaining higher criterion-referenced reliability (as Livingston calls it) is the acceptance of exceedingly crude categorization—pass or fail. Clearly, if one's objective in using any test is to dichotomize the examinees around a preselected passing mark (or, more generally, a preselected cutting score), he may prefer to use Livingston's coefficient as an estimate of the reliability of his dichotomized scores (Livingston, 1972b).

A note of caution should be introduced at this point for those who make practical use of content-referenced interpretations (called criterion-referenced interpretations by many people) of achievement tests properly constructed to measure a representative sample of a carefully defined domain of behaviors. If it is desired to determine whether one examinee's score is significantly above or below the passing mark, a standard error of measurement of an obtained score should be used for this purpose. Both Livingston (1972b) and Harris (1972) agree on this point. The present writer goes a bit farther in recommending use of the standard error of measurement appropriate for each score level. Livingston's criterion-referenced reliability coefficient is not used in estimating standard errors of measurement.

Livingston's paper shows that the relationship between the criterion-referenced reliability coefficient and the norm-referenced (that is, conventional) reliability coefficient is linear when the difference between the mean of scores in the sample and the passing mark is held constant regardless of score level. For any given test, as the difference between the mean score in a sample and the passing mark increases, the most rapid increase in the criterion-referenced reliability coefficient will occur when that difference is about half a standard deviation of obtained scores. In general, the greater the difference, the more the criterion-referenced reliability coefficient will exceed the conventional reliability coefficient.

Livingston's discussion of the validity of dichotomized scores should be of special interest to research workers in both education and industry. Applications of dichotomized mastery-test or selection-test scores are made in both fields. The correlation between a set of dichotomized test scores and a set of dichotomized criterion scores may differ in both magnitude and sign from the correlation of continuous variables consisting of obtained scores in the same variables. As the dichotomic points are changed (by altering the passing mark), the criterion-

referenced validity coefficient of the test scores changes.

The second paper to be reviewed was entitled "An Index of Efficiency for Fixed-Length Mastery Tests" by Chester W. Harris. Limiting his field to fixed length mastery tests, Harris outlines a way of estimating an index of efficiency to indicate the extent to which number-right scores on a mastery test can discriminate two groups defined as having all members scoring at or above a designated cutting score or all members scoring below that cutting score. His approach is suggested by Fisher's linear discrimination function for two groups (Fisher, 1936; Tatsuoka, 1971, Chapter 6), and the resulting index turns out to be the squared point-biserial correlation coefficient. Harris points out some interesting properties of the index and notes that, for symmetric distributions, it takes a maximum value for any given test when the proportion of all examinees in the group at or above the cutting score is .50. As Harris has stated, this result agrees with that obtained by Richardson (1936) by different methods. The present writer expresses one meaning of efficiency (in the sense meant by Harris) as the proportion of all possible differentiations between examinees in the two dichotomous groups that actually is made. This proportion is maximized when the dichotomizing score is the median score. In practice, a higher score is usually employed as a cutting score with a mastery test; therefore, the efficiency index will probably not often be found to be close to its maximum value. It is interesting to note that there is likely to be a tendency for the index of efficiency suggested by Harris to decrease as the criterion-referenced reliability coefficient suggested by Livingston increases. This situation is not paradoxical; it results from the fact that the two indexes provide different kinds of information about test scores and from the lack of symmetry of the distributions of scores ordinarily obtained when mastery tests are administered as pretests or posttests.

The paper presented by Dr. Thomas E. Kriewall on "Aspects and Applications of Criterion-Referenced Tests" is organized under three main topics: "The Instructional Context for Criterion-Referenced Testing"; "The Fundamentals of the Theory of Criterion-Referenced Testing", and "The Implementation of the Criterion-Referenced-Test Model."

Kriewall points out that criterion-referenced measurement is intended to yield for individual pupils content-specific estimates of proficiency that are useful in instructional management systems. Such estimates may be used to place pupils in temporary learning groups for instructional treatment; to determine which pupils have attained minimal preselected standards of proficiency and which ones require additional instruction; to assess the relative effectiveness of specified instructional procedures, and to identify hierarchical relationships within a given content area. Kriewall defines criterion-referenced tests as instruments constructed to provide proficiency estimates of the types described.

In the context of measurement procedures woven into the very fabric of the learning-teaching process, Kriewall points out, the usual index of item difficulty (an estimate of the proportion of examinees in a defined population who will mark an item correctly) is not usually of value. Instead, there is need to estimate the proportion of items that make up a defined population of tasks (which define an objective of instruction) that each individual examinee can answer correctly. It should be remembered, however, that if only one item is drawn from such a population to estimate the population proportion, the errors of measurement may be large even if the items are so homogeneous (with respect to the function measured) that errors of sampling are not troublesome.

Kriewall deplores the use of subjective judgment by test constructors and suggests that *prima facie* content validity for a test be obtained by defining each objective to be measured in terms of a specified population of items, of which a *random* sample is to be used in measuring the objective. However, this procedure might cause the validity of a test to rest on the industry and the subjective insight and skill of the item writers and editors; it would eliminate the subjective judgment ordinarily exercised by the test assembler in trying to make sure that each objective is measured by a *representative* sample, or stratified representative sample, of the editorially satisfactory items available.

With respect to generating parallel forms of a criterion-referenced test, Kriewall recommends use of a random-number generator to sample the specified item population. Unless the true interitem correlations in the population of items are all unity (a situation not likely to be experimentally demonstrable), it might be safer in dealing with the usual small groups of items to stratify the latter subjectively in terms of the elements of the objective that they were intended to measure and then assign items from each stratum (or cell) at random to the forms. This procedure has been widely employed in test construction for some years.

The fact that criterion-referenced tests are ordinarily used to measure proficiency with respect to very specific content or skills or to evaluate the effect of instruction in that content or those skills means that distributions of scores derived from criterion-referenced pretests or posttests are markedly skewed because the latent trait or traits being measured are, at those particular times, markedly skewed. Kriewall has noted this situation and states, "The data of interest to the teacher are not the class mean and relative ranking of class members but rather data which permit the correct classification of students into subsequent instructional groups, together with estimates of absolute levels of proficiency within each group."

If the items in a criterion-referenced test are, in fact, such that their true intercorrelations are unity (as recommended by Livingston in the first paper reviewed), the teacher would be interested in forming instructional groups on the basis of total scores on the test. However, if the items measured separate elements that have true intercorrelations of less than unity in the objective that the test was constructed to measure, the teacher would be interested in forming groups for remedial instruction in those elements which had not been mastered. This is, in fact, often a major goal for teachers who use criterion-referenced tests. Unfortunately, they may not fully recognize the danger of misassigning pupils to instructional groups on the basis of individual scores on one or on very few items.

Kriewall defines test-score reliability in the classical sense as the squared product-moment correlation between true and observed scores. This definition leads to the concept of the reliability coefficient as an index of the consistency of obtained scores derived from administration of equivalent forms of a test to the same examinees under specified conditions. As Kriewall indicates, a test constructor whose goal is to maximize the variance of a test that he is building should, insofar as practical circumstances permit, use items of .50 difficulty if number-right scoring is to be used in samples of examinees like those used to estimate the item difficulties. However, this goal is only one of many recognized by classical test theory and for which techniques of test construction can be specified.

The major goal in developing a criterion-referenced test is to provide items that will elicit examinee behaviors that literally constitute overt manifestations of the feelings, skills, and knowledge (facts and understandings) that make up the objectives of measurement. Under these circumstances, classical test theory suggests that item difficulty level should not enter into the procedure used to build the test. Item difficulty should come about simply as a by-product of efforts to make items elicit behaviors of the types specified. This matter was explicitly discussed by Davis (1951, p. 315).

Applications of classical test theory to estimating the reliability of criterion-referenced tests have been presented by Livingston and Harris in the first two papers reviewed. Because individual scores on such tests are used mainly for guiding instruction (prescribing reteaching or progression to additional material) and for assessing achievement level at a given point in time, the standard error of measurement of each individual's obtained score (not the over-all standard error of measurement) may be of a greater value to teachers than the reliability coefficient.

In the second part of his paper, Kriewall presents the fundamentals of his theory of criterion-referenced tests.

1. He stresses the need for defining a learning objective in terms of specified performance *tasks* and then selecting a random sample of *items* for a given test form. However, in practice, items constructed to measure performance of these tasks cannot usually be obtained without involving human item writers. Consequently, the extent to which the items measure the specified tasks depends on their insight, ingenuity, and industry. This is especially true if the performance tasks that are to be measured go beyond simple and very specific tasks like those involved in the fundamentals of arithmetic. Thus, to generalize from a sample of *items* to a population of *tasks* may not always be legitimate whereas to generalize from a sample of items to a population of items "like those observed" is always legitimate.

2. He points out that if test scores are to have absolute meaning, the items must truly define the learning objective. If, in addition, the items are of the free-response type, a zero true score means that the examinee has no knowledge of the objective and a perfect true score means that the examinee has "completely" mastered the objective tested. If items were selected according to difficulty to conform to some predetermined level, the absolute value of the observed scores would not necessarily be meaningful.

3. He defines the proficiency of a given student (a) with respect to a defined learning objective (k) at any given point in time as ζ_{ak}. This may be regarded as the mean difficulty level of an item population for a given individual. The complement of this parameter is the error rate.

4. He points out that, for a test constructed to meet the specifications described in his paper, the standard error of measurement for any given examinee is determined by proficiency (as defined in the preceding paragraph) and test length, as first shown by Lord (1955).

In the third part of his paper, Kriewall discusses problems in setting a cutting score, or "passing mark," used to separate those who have mastered the content tested from those who have not. Like Livingston, he refers to this cutting score as a "criterion," though Glaser (1963, p. 519) used the term to mean "the behavior which defines each point along the achievement continuum." Nitko (1970, p. 38) specifically stated that the term "criterion" does not mean a cutting score. This point is mentioned only because the terminology used in discussions of criterion-referenced measurement may sometimes be confusing to students.

Kriewall shows how a band of obtained scores can be defined to set limits of acceptable error in classifying

examinees as having mastered or as not having mastered the content measured provided that the examiner has specified the maximum number of items that an examinee can mark incorrectly and still be regarded as having "mastered" the content. Thus, the major factor in setting a cutting score is the subjective judgment of the examiner with respect to proficiency level. With regard to the length of "mastery" tests. Kriewall suggests that 20-25 items per test are required if the difference between proficiency limits for mastery or nonmastery is small. In general, he believes that "tests of greater length are likely to be wasteful for most practical instructional decision situations."

The fourth paper under review is 'Student Evaluation: Toward the Setting of Mastery Performance Standards" by James H. Block. It is directed toward establishing a method for setting mastery-test performance standards in a rational way. To do so requires that an objective criterion be used with which to judge the outcome of requiring learning to various criteria. For example, a teacher might require that students mark correctly 95 per cent of the items on a criterion-referenced mastery test before proceeding to the next unit of work. Would it have been best for her to have set the "passing mark" at 65 per cent, 75 per cent, 85 per cent, or 95 percent?

As a criterion for deciding which of these would be best, Block suggests a weighted composite of performance levels at the end of a course of instruction. For an actual experiment, he used performance levels in the following areas: achievement level, retention, transfer, and rate of learning of the subject matter taught as well as level of short-term and long-term interest in the subject matter. Ninety-one eighth-grade pupils were taught a three-unit sequence in elementary matrix algebra during one school week. The students were randomly assigned to five treatment groups. A control group learned with no requirement of meeting a specified performance level; but pupils in each of the four experimental groups were required to demonstrate learning of a preselected per cent (65, 75, 85, or 95) of the content taught.

The results showed that the maintenance of specified performance levels did have effects significant at the .05 level on different criteria Broadly speaking, learning to the 95-per-cent performance level was optimal for the criteria of achievement level, retention, transfer, and rate of learning whereas learning to the 85-per-cent performance level was optimal for the criteria of short-term and long-term interest.

With these data it would be possible to specify a cutting score required for "passing" on one composite measure used at the end of each unit. A weight representing the teacher's judgment of the importance of each of the five criteria would be applied to scores on each of the elements of the composite, the weighted average would represent the level of performance that should be represented by the passing mark. Additional experimental data to supplement these findings are now being gathered.

The fifth paper under review was written by Eugenia M. Koos and James Y. Chan on "Criterion Referenced Tests in Biology." This paper provides a detailed description of the development of a series of single-topic tests built to measure attainment of 14 objectives consisting of inquiry skills in biology taught to high-school sophomores. Scoring was accomplished with weights for each item based on the pooled judgment of the item writers. The authors hope to use the tests for research work on the relationship of inquiry skills to elements of Guilford's structure-of-intellect model and on the possibilities of measuring complex cognitive skills by means of criterion-referenced tests.

The sixth paper reviewed is entitled "Adapting Criterion-Referenced Measurement of Individualization of Instruction for Handicapped Children: Some Issues and a First Attempt" and was written by Barton P. Proger, Lester Mann, Robert M. Burger, and Lawrence H. Cross. It defines distinguishing features of handicapped children that require specialized measurement systems and describes a criterion-referenced measurement system designed especially for the handicapped. The authors argue that teaching and testing procedures used with normal pupils cannot simply be adapted for use with handicapped pupils. Consequently, they devised a criterion-referenced system called the Individual Achievement Monitoring System which is described in detail in the paper.

The seventh paper to be reviewed is "A Pragmatic Approach to Criterion-Referenced Measures" by Stephen H. Ivens. The author mentions some of the difficulties encountered in using traditional methods for evaluating item quality and test-score reliability. Some procedures for overcoming these difficulties have already been discussed in the preceding papers that have been reviewed here. Ivens suggests that any one or a group of homogeneous items that displays a difficulty index markedly different from the others should be carefully scrutinized for possible ambiguity or other fault.

He recommends that items for the evaluation of instruction be selected on the basis of the difference between the proportion of examinees marking an item correctly on a pretest and on a posttest. While there is no doubt that this technique would identify items sensitive to the effects of instruction, there remains the danger that such a test might no longer constitute a valid measure of the elements of the objectives of instruction. In other words, it might violate the basic requirement (mentioned by Ivens earlier in his paper) that a criterion-referenced test must be comprised of test items keyed to a set of behavioral objectives so that the behaviors measured literally constitute a representative sample of those in the objectives.

Ivens recommends that item reliability be estimated by calculating the proportion of examinees whose item scores (pass or fail) are the same on two administrations of the item as a posttest to the same examinees or on two items judged to measure the same specific point and administered in a posttest. He also suggests that a measure of test reliability can be obtained by calculating mean item reliability.

Another index of item quality may be obtained by using the proportions defined below:

$$\text{Index} = (P_B - P_A)(1 - |P_C - P_B|),$$

where P_A = proportion of examinees who mark the item correctly in a pretest;

P_B = proportion of the same examinees who mark the item correctly in a posttest;

P_C = proportion of the same examinees who mark the item correctly on a retest (after the posttest).

Ivens contends that this index reflects item validity in the first term and item reliability in the second term. Readers of this review will note that alternative procedures for obtaining some of the evaluative statistics desired by Ivens have been presented by Livingston, Harris, and Kriewall in preceding papers.

The eighth and last paper reviewed is "Implementing A Mixed Program of Criterion- and Noncriterion-Referenced Measurement" by Harold F. Rahmlow. This paper sketches the nature of a transition from norm-referenced measurement to criterion-referenced measurement in a continuing program designed to certify candidates as chartered life underwriters.

In each of two projects, a criterion-referenced test was developed and used to improve an instructional program. Yet the value of the courses is to be determined by the success of their graduates in passing certifying examinations which are keyed to general objectives rather than to the specific objectives of the courses. A problem to be tackled is that of developing some realistic measures of on-the-job performance in serving the public to substitute for or to supplement the certifying examinations.

In conclusion, the reviewer is happy to report that these papers and others published during the past year on criterion-referenced measurement are showing increasing perception of the basic problems of this field and its relationship to classical test theory.

PAPERS REVIEWED

Some additional valuable references furnished by the author are grouped separately following this list of the 1972 AERA papers reviewed in this summary.

Block, J.H. Student evaluation: Toward the setting of mastery performance standards. 28p. (ED 065 605, MF and HC available from EDRS.)

Harris, C.W. An index of efficiency for fixed-length mastery tests. 10p. (ED 064 349, MF and HC available from EDRS.)

Ivens, S.H. A pragmatic approach to criterion-referenced measures. 9p. (ED 064 406, MF and HC available from EDRS.)

Koos, E.M., & Chan, J.Y. Criterion referenced tests in biology. 14p. (ED 062 399, MF and HC available from EDRS.)

Kriewall, T.E. Aspects and applications of criterion-referenced tests. 17p. (ED 063 333, MF and HC available from EDRS.)

Livingston, S.A. A classical-test-theory approach to criterion-referenced tests. 12p. (ED number not yet available from EDRS.)

Proger, B.P., & Others. Adapting criterion-referenced measurement to individualization of instruction for handicapped children: Some issues and a first attempt. 24p. (ED 064 354, MF and HC available from EDRS.)

Rahmlow, H.F. Implementing a mixed program of criterion- and noncriterion-referenced measurement. 7p. (ED 063 331, MF and HC available from EDRS.)

REFERENCES

Davis, F.B. Item selection techniques. In E.F. Lindquist. (Ed.), *Educational Measurement*. Washington. American Council on Education, 1951.

Fisher, R.A. The use of multiple measurement in taxonomic problems. *Annals of Eugenics*, 1936, 7, 179-188.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, 18, 519-521.

Harris, C.W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, 9, 27-29.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, 9, 13-26. (a)

Livingston, S.A. A reply to Harris's "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, 9, 27. (b)

Lord, F.M. Estimating test reliability. *Educational and Psychological Measurement*, 1955, 15, 325-336.

Nitko, A.J. Criterion-referenced testing in the context of instruction. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970. Pp. 37-40.

Richardson, M.W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, 1, 33-49.

Tatsuoka, M.M. *Multivariate analysis*. New York: Wiley, 1971.