DOCUMENT RESUME

ED 060 134 TM 001 400

AUTHOR Davis, Frederick B.
TITLE 1971 AERA Conference Summaries: II. Criterion Referenced Measurement.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
REPORT NO TM-R-12
PUB DATE Mar 72
NOTE 20p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Conference Reports; *Criterion Referenced Tests; *Literature Reviews; Meetings; Norm Referenced Tests; Research Reviews (Publications); Resource Materials; *Test Construction; Testing; Test Interpretation; Test Reliability; *Tests; Test Validity
IDENTIFIERS AERA; American Educational Research Association

ABSTRACT

Thematic summaries of those papers from the 1971 annual meeting of the AERA dealing specifically with criterion-referenced measurement are presented. An evaluative commentary by the writer is also included. (AG)

TM REPORTS

NUMBER 12

1971 AERA Conference Summaries

II. Criterion Referenced Measurement

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

The Clearinghouse operates under contract with the U. S. Department of Health, Education and Welfare, Office of Education. Contractors are encouraged to express freely their judgment in professional and technical matters. Points of view expressed within do not necessarily, therefore, represent the opinions or policy of any agency of the United States Government.

March 1972

CRITERION-REFERENCED MEASUREMENT

Frederick B. Davis

ERIC Clearinghouse on Tests, Measurement, and Evaluation

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
   TM 000 987 (ED number not yet available)

6. State Educational Assessment Programs: An Overview
   TM 001 024 (ED number not yet available)

7. Criterion Referenced Measurement: A Bibliography
   TM 001 046 (ED number not yet available)

INTRODUCTION

About 575 of the 700 papers presented at the 1971 AERA Annual Meeting in New York City 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) 175 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 by our Clearinghouse. Five topics were chosen for the series: Criterion referenced Measurement, Evaluation, Innovation in Measurement, Statistics, and Test Construction.

Individual 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 O
Bethesda, Maryland 20014

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

Editor, ERIC/TM

Among papers read at the Annual Meeting of the American Educational Research Association held in February 1971, five dealt specifically with criterion-referenced measurement. Brief summaries of these are provided together with an evaluative commentary by the writer.

Nitko, in a paper entitled "A Model for Criterion-Referenced Tests Based on Use" discussed the design and development of criterion-referenced tests in the light of the purposes that such tests are intended to fulfill. The primary purpose of criterion-referenced tests is "to yield scores that are directly interpretable in terms of specified performance standards (Glaser and Nitko, 1970, p. 653)." Or, as Nitko (1970, p. 38) wrote: "A criterion-referenced test is one that is deliberately constructed to give scores that tell what kinds of behaviors individuals with those scores can demonstrate." Specifically, Nitko points out, the term "criterion" in "criterion-referenced test" does not mean a series of scores used as a criterion for obtaining predictive validity coefficients; neither does it mean a cutting score used for determining whether any given examinee has "passed" or "failed" or "has attained mastery" or "has not attained mastery." The criterion as defined by Glaser (1963, p. 519), "is the behavior which defines each point along the achievement continuum." In short, the term "criterion" in "criterion-referenced tests" refers to content, as in the familiar term, "content validity."

In constructing a criterion-referenced test, the behavior categories that are to be measured must be clearly specified in a test outline. Items are then devised to test these behaviors. A systematic plan must be devised to make sure that each try-out form of the test includes a representative sample of items in the behavior categories. In making this plan, the domains of items must be carefully examined and, if necessary, stratified so that the sampling will be truly representative. Item analysis data should then be used as a basis for refining the items through insightful editing, but the use of item-test correlation coefficients or difficulty indexes must not be allowed to affect the validity of the test by distorting the proper representation of behavior categories or achievement levels, as specified by the test outline.

Criterion-referenced tests have special applicability in instructional contexts, and especially with the use of individualized instructional procedures. For such procedures, the desired outcomes, or terminal objectives, are specified; and learning exercises, or tasks, are provided that enable the learners to perform the behaviors that constitute the desired objectives. To facilitate this process, a sequence of intermediate goals is set up and several different instructional methods may be (but, in practice, rarely are) provided for use with learners who differ with respect to the method, or approach, that permits them to learn most efficiently. For example, it is believed that some children learn to read most efficiently by a predominantly visual-perceptual approach while others learn most efficiently by a predominantly aural-phonic approach.

When individualized instructional plans are used, decisions need to be made with regard to the *placement* of the pupil in the sequence of units, the *diagnosis* of which instructional approach is likely to be most effective, and the *attainment* of the desired outcomes. Preliminary decisions about *placement* can best be made with a broad-range test that comprises a battery of subtests, one for each major topic in the domain to be taught. If the content of each domain is hierarchical (one skill or facility depending on previous skills or facilities, as in mathematics), each subtest should indicate the highest-level unit at which a pupil can perform satisfactorily. If the content is not hierarchical (as is often the case in the social studies), the subtest scores should indicate which topics the pupil has already learned satisfactorily. Nitko recommends that, in constructing this broad-range test, traditional item analysis procedures be used to maximize predictive validity, but he does not specify what the criterion to be predicted should be.

The second state of the placement decision apparently makes use of a pretest based on the content of the first unit that the broad-range test has indicated should be studied. Presumably, this pretest would be the same criterion-referenced test (or an equivalent form of it) to be used at the end of the unit for measuring degree of attainment of the desired objectives.

The meaning of the hierarchical, or psychological, structure of content to be taught is discussed and illustrated. Nitko concludes by

writing, "In short, it is the use to which test results are put that determines their nature and the construction methodology. In instruction, various procedures cannot be considered independently of the instructional context in which they will be used. Particularly important is the integration of test design with instructional design."

In a paper entitled "Empirical Data on Criterion-Referenced Tests," Hsu has described techniques of item analysis and reliability estimation used for criterion-referenced tests. He does not recommend item-discrimination indices that depend on differences in pupil performance brought about by instruction. Instead, he suggests three other possibilities:

1. The difference (Dp) between the proportions of pupils who mark an item correctly in the group whose total scores were *above* the point designated as indicating "mastery" and in the group whose total scores were *below* that point. In practice, use of this index would favor items with high internal consistency and items that were marked correctly by percentages of the two groups that were close to 50.

2. The phi coefficient (Ø) obtained in the fourfold table for each item where the dichotomies consist of marking the item correctly or incorrectly and of obtaining a total score above or below the point designated as indicating mastery. To interpret the coefficient as a product-moment r, the dichotomies would have to be natural; in this usage, neither of them is likely to be. In practice, use of this index would favor items with high internal consistency and items that were marked correctly by about the same percentage of examinees as obtained scores above the point designated as indicating "mastery." As Hsu points out, the index is not usable when all examinees fall in one category of either dichotomy or of both dichotomies.

3. The point-biserial correlation coefficient (r_{pbis}), where the distribution of total scores is not dichotomized and the item scores (pass or fail) are considered to be a natural dichotomy. In practice, use of this index would favor items toward median difficulty. It is not usable when all examinees fall in one category of the item-score dichotomy or obtain the same total score.

Empirical studies were made to determine the relationships among Dp, Ø, and r_{pbis} and their consistency from sample to sample under various conditions. Broadly speaking, these studies show that:

1. In samples where test scores are widely and symmetrically distributed, the three indices are highly correlated.

2. In samples where test scores are narrowly distributed and skewed, the three indices have low correlations.

3. In samples having similar test-score distributions, the three indices are relatively consistent from one sample to another.

4. When items are grouped by difficulty level, the three indices are most closely correlated among groups of items of middle difficulty.

Hsu's brief discussion of the reliability of criterion-referenced tests does not warrant summary here, but his estimations of the reliability coefficients of 4- and 5-item tests in Tables 2, 3, and 6 for small samples are interesting. Using Kuder-Richardson equation 20, he obtained a median coefficient of .83. Even though he deliberately chose small groups of very homogeneous items, it is somewhat surprising that the coefficients are as high as they are. Apparently, when items sample a very restricted homogeneous universe of skills they are marked with high consistency by well-motivated examinees.

Roudabush and Green is a paper entitled "Some Reliability Problems in a Criterion-Referenced Test" described the design and construction of a criterion-referenced test intended for use in Grades 4-8 throughout the United States. Consequently, specific objectives were culled from the texts and materials most widely used in schools, collated, and classified into broader objectives. The categorical and hierarchical structure of these was then ascertained.

With a larger number of objectives (about 400) to be measured, the test would become impossibly long if achievement in every objective were measured with enough items to make the achievement scores for each objective adequately reliable. If objectives are combined to reduce the number of achievement scores, and thus provide enough items to yield highly reliable scores, the test loses its diagnostic value.

Presumably, errors of measurement (the source of unreliability) arise mainly from two situations:

1. A pupil who has not mastered the objective tested by an item can mark the item correctly by guessing among a finite number of choices;

2. A pupil who has mastered the objectives tested by an item may mark the item incorrectly by clerical error, a lapse of attention, a predetermined mind-set, etc.

To reduce the first cause of errors of measurement, the authors recommend increasing the numbers of choices in the items. For arithmetic items, they suggest a means of coding free-response items for machine scoring. To reduce the second cause of errors of measurement, the authors suggest use of multiple-regression equations for predicting pass or fail on any one item (measuring a specific behavioral skill) from scores on all other items in the test (each of which measures a sensibly different, though related behavioral skill) or from total scores on groups of, say, 6 items judged homogeneous and found to be so empirically. Clearly, development of stable cross-validated regression equations would require the use of thousands of cases, and scoring would be slow (and thus expensive) by computer-scoring standards. In any event, the upper limit of the multiple correlation coefficients might be low since any one of them could not exceed the square root of the product of the reliability coefficient of the single-item dichotomous scores used as the dependent variables and the reliability coefficient of the weighted composite scores used as predictors. In some ways, this ingenious suggestion makes one think of using an elephant gun to kill a gnat.

A paper by Brennan and Stolurow entitled "An Elementary Decision Process for the Formative Evaluation of an Instructional System" provides a set of objective rules, based on item performance data, for identifying test items and sections of an instructional procedure that may require revision. The rules, however, will not necessarily tell the evaluator *how* to make these revisions. Let us consider at this point the terminology used by Brennan and Stolurow so that their materials can be summarized compactly.

Pretest: a test given prior to instruction on its content;

Terminal test: a test given almost immediately after instruction on its content;

Posttest: a test given "some time" after instruction on its content;

BER: Base Error Rate (the observed proportion of examinees who mark a pretest item incorrectly);

TER: Theoretical Error Rate (the proportion of examinees most likely to mark a pretest item incorrectly by chance alone);

BDI: Base Discrimination Index (correlation of item scores with total scores on pretest);

PER: Posttest Error Rate (the proportion of examinees who mark a posttest item incorrectly);

PDI: Posttest Discrimination Index (correlation of item scores with total scores on posttest);

IER: Instructional Error Rate (the proportion of examinees who mark incorrectly a terminal test item for a specified single objective);

DER = TER - BER;

RER = PER - IER;

PMPG = (BER - IER)/BER

C1 is a score level above which, in the evaluator's judgment, forgetting is great enough to warrant revision of the instruction;

C2 is a score level above which, in the evaluator's judgement, the instruction should be questioned;

C3 is a score level above which, in the evaluator's judgment, the instruction needs revision.

Brennan and Stolurow state a number of rules regarding the revision of test items. They assume that items measuring the same objective in the pretest, terminal test, or posttest are either identical or, in the same sample would measure the same functions and yield equal means, variances, and intercorrelations. Then:

1. If TER = BER, no item revision is needed;
2. If TER < BER, no item revision is needed, especially if the difference is great;
3. If TER > BER, item revision is probably needed, especially if the difference is great;
4. If BDI is negative, item revision is probably needed;
   If BDI is zero or positive, no item revision is needed;
5. If BDI is either negative or positive, the prerequisites for the objective tested should be questioned;

6. If PER is low and PDI is zero, no revision of item or instruction is needed;
7. If PER is low and PDI is either positive or negative, both the item and instruction may need revision;
8. If PER is high and PDI is negative, both the item and instruction need revision;
9. If PER is high and PDI is either positive or negative, the instruction needs revision and the item may need revision;
10. If PDI is positive or negative, the prerequisites for the objective tested by the item should be questioned;
11. If BDI and PDI are both negative, the item needs revision;
12. If IER and PER are low, no revision of instruction is needed;
13. If IER is low and PER is high, the instruction needs revision;
14. If IER is high and PER is low, the instruction should be questioned;
15. If IER and PER are both positive, the instruction needs revision;
16. If DER$<\emptyset$ by a significant amount, the item needs revision;
17. If RER$>$CI, the instruction needs revision;
18. If RER$<$-C2, the instruction should be questioned;
19. If PMPG$<$C3, the instruction needs revision.

The authors discuss statistical tests for the equality of means, variances and intercorrelations for dichotomously scored test items and suggest the use of Cochran's Q Test, which is appropriate when the population distributions of scores on the items are not assumed to be normal. They offer no exact test for the equivalence of phi coefficients among a set of items. The reviewer suggests the possibility of using an exact test given by Hotelling (1940) of the difference between two product-moment coefficients (and phi coefficients are such). This could be applied to each pair of coefficients in a matrix, thus identifying a difference between any two coefficients that showed statistical significance at a preselected level.

"The Effect of Criterion-Referenced Testing Upon the Use of Remedial Exam Opportunities" by Blumenfeld, Bostow, and Waugh is an article dealing with the practical use of criterion-referenced tests in undergraduate instruction in psychology. The results of an experiment involving several groups of students suggest that a larger proportion of students who could

profit from studying for and taking a second criterion-referenced unit test take such a test when the passing mark on the first unit test is set high than when that passing mark is set low. The data also suggest that, if only minimal credit is given for low scores, the students will tend to attain or exceed a high passing mark. It may be inferred that, when criterion-referenced tests are used to test achievement in successive units of a course, they should be used with relatively high passing marks.

The first four papers summarized above should serve to correct some of the misconceptions that have arisen during the past five years about criterion-referenced tests. The latter should be viewed in perspective as achievement tests constructed, as good tests of this type always have been, to provide evidence of level of attainment in a carefully defined body of content. In fact, Nitko's article may lead one to conclude that the real nature of criterion-referenced tests might have been clearer to prospective users had they been called "content-referenced tests" or if it had been pointed out that, although all achievement tests are subject to content-referenced and to norm-referenced interpretations, the most valid content-referenced interpretations can be made only when the test itself has been systematically built to detailed specifications in terms of behaviorally defined objectives that constitute the content to be measured. In short, it is not tests themselves but the interpretations of test scores that may legitimately be dichotomized by the descriptors "content-referenced" and "norm-referenced."

The papers by Nitko, Hsu, and Brennan and Stolurow may dispel earlier misconceptions about the use of item analysis data in the development of tests designed to provide maximally useful content-referenced score interpretations. All of these papers indicate that appropriate types of such data can be very helpful if they are used insightfully, especially for detecting unnoticed faults in items. Although insightful revision or elimination of items is the most important outcome of item analysis, the technique has often been used mechanically to select items solely on the basis of item-test correlation coefficients of one sort or another. As Davis (1952) pointed out years ago,

> For achievement tests, great care must be exercised that items judged unacceptable by subject-matter experts be excluded and that the final form preserve the balance among topics specified in the test outline. Then, too, proper regard for the shape of the distribution of item difficulties must be observed, as noted earlier in this article. The value of item-discrimination indices must always be considered in the light of adequacy of the criterion variable, the purpose for which the test is to be used, and the way it serves that purpose...the usefulness of item-discrimination indices is often smaller than is commonly supposed (pp. 116-118).

Like discrimination indices, difficulty indices have often been misused. For example, items close to 50 percent difficulty have frequently been selected for a test in the belief that such items are perfectly pitched in difficulty. But for tests made up of more than one item, this is true only when it is desired to maximize the number of differentiations that can be made among all of the examinees when the product-moment intercorrelations of the items average .33 or lower (as they ordinarily do) or when it is desired to maximize the number of differentiations that can be made between examinees below and examinees above the raw-score median regardless of the level of item intercorrelation. Since neither of these objectives is likely to be relevant in the development of achievement tests designed to provide content-referenced interpretations, classical test theory suggests that items should *not* be selected on this basis. Item difficulty should, perhaps, come about simply as a by-product of efforts to make the items elicit behaviors that constitute overt manifestations of the feelings, skills, and knowledge that made up the objectives of instruction and of the effectiveness of the procedures used to teach these objectives.

The paper by Roudabush and Green discusses some problem of reliability in connection with scores from what might be described as a criterion-referenced survey test of mathematics. For this type of test, at least two legitimate types of interpretations can be made:

1. We may estimate the percent of the behaviors in the domain that the examinees have shown that they can perform correctly. If multiple-choice items are used, scores on the test that have been corrected for chance success will ordinarily allow making a better estimate of this percent than will number-right scores. It should be noted, however, that this

type of content-referenced interpretation does not indicate the particular behaviors that have or have not been demonstrated by each examinee. Therefore, it does not fulfill the purpose of criterion-referenced tests stated by Nitko (1970, p. 38).

2. We can determine whether any examinee did or did not correctly demonstrate the specific behavior tested by each separate item. But it is dangerous to infer that the examinee's performance would be at the same level of competence on each of a large number of equivalent (though not identical) items testing the same behavior. Although the best estimate of his true level of competence with respect to a specific behavior is his score on the one item testing it that he has tried, this estimate is subject to error, possibly to a far greater degree of error than we ordinarily tolerate in test interpretation. Unless satisfactory evidence to the contrary is provided, diagnosis of individual strengths and weaknesses on the basis of one-item tests should be regarded as highly tentative.

Theoretically, the accuracy of measurement of a one-item test for any given examinee could be estimated by obtaining the standard deviation of scores on a large number of equivalent (though not identical) items administered to him under specified conditions. The standard deviation of these scores would be the standard error of measurement of that individual's obtained scores. In practice, we are unable to administer a sufficiently large number of equivalent items to any one individual, so we may administer two equivalent items to a large number of examinees and compute the over-all standard error of measurement as an estimate of the standard error of measurement of the obtained single-item score of any examinee drawn at random from the sample. The required equation for an item scored 1 for a correct response and 0 for an incorrect response or an omission is:

$$s_{meas\ i} = \sqrt{p_i q_i \quad (1 - r_{iI})} \ ,$$

where p_i = the proportion of the sample that marked the item correctly; $q_i = 1 - p_i$; and r_{iI} = the product-moment correlation coefficient between scores on the two equivalent items. Clearly, $p_i q_i$ is largest for items of 50 percent difficulty in a sample (.50 x .50 = .25) and becomes small for difficult or easy items; for example, when $p_i = .90$, $p_i q_i = .09$. Since criterion-referenced tests are often administered immediately after a unit

of material has been taught to find out what behaviors have or have not been learned by each pupil, the items of which they are made up are usually found to be easy. Ordinarily, the reliability coefficients of single items are very small, ranging from, say, .10 to .20. The writer found in a sample of 800 airmen that the median reliability coefficients of very homogeneous perceptual items ranged from about .15 to .18. Yet Scandura and Durnin (1971) report data indicating that the reliability coefficients of single items testing highly specific behaviors (pertaining to the use of rules in solving arithmetic problems) that have been taught and practiced just prior to the testing were as high as .60 to .90 in very small samples. It may be that under certain special circumstances single test items have higher reliability coefficients than would be expected. Data in the papers by Hsu and by Roudabush and Green support this conjecture.

Scored 1 for a correct response and 0 for an incorrect response or an omission, a single item that was answered correctly by 68 percent of a sample and that had a reliability coefficient of .75 would have a standard error of measurement of about .23. Therefore, an examinee who obtained a score of 1 would be unlikely by chance alone to obtain a score of 0 on an equivalent item. If this item displayed a reliability coefficient of .15, however, it would have a standard error of measurement of .43. Under these circumstances, an examinee who obtained a score of 1 could fairly readily obtain a score of 0 by chance alone on an equivalent item. Additional experimental evidence is needed to determine the standard errors of measurement of short diagnostic tests administered directly after the content measured by the tests has been taught.

From this discussion it is apparent that, although the second type of content-referenced interpretation does indicate the particular behaviors that any examinee has demonstrated, such data may be so unreliable as to make them of doubtful value. Twenty or thirty years ago some test-scoring services reported results in such a way that pupils and teachers could see exactly which items in achievement tests had been marked correct or incorrect. But these data have not become widely used, partly because they publicized the scoring keys for the tests and partly because they were unreliable. The diagnostic matrix described by Roudabush and Green may overcome objections made to the older procedures.

However, it may be that a criterion-referenced test covering a wide domain is not likely to provide data that satisfactorily fulfills the basic purpose of such tests. What, perhaps, should be available for any given domain is a coordinated set of diagnostic subtests, each of which is made up of items that are homogeneous in the sense that they test performance on one specific behavior or on a cluster of behaviors that are taught as a unit. The experimental justification for obtaining a total score for each subtest from items measuring a cluster of behaviors would consist of evidence that the tetrachoric intercorrelations of single items in the cluster were as high, or nearly as high, as their reliability coefficients would permit and lower than their correlations with single items in other subtests in the coordinated set covering the domain being measured. Each subtest would comprise enough items so that a perfect score on it would not be likely to be obtained by chance alone (at some designated level of probability) by an examinee who had not truly mastered the behavior being tested. Ordinarily, a pupil would need to take only one subtest at any one time, and preliminary data provided by Hsu and by Roudabush suggest that acceptably reliable scores could be obtained from as few as 10-12 homogeneous items.

At this point, it should be noted that when a teacher or counselor interprets a test score made by an individual pupil, knowledge of the reliability coefficient of the scores on that test in a large sample of his peers is not of direct value. The test interpreter needs the standard error of measurement at or very close to the score made by the pupil under consideration. With this information available, a confidence interval (at any designated level) can be constructed and inferences about the proximity of the pupil's obtained score to his true score can be drawn. Likewise, the statistical significance (at any desired level) of the difference between the pupil's obtained score and the passing mark (if one has been established) can be estimated. In making this estimate, the criterion-referenced reliability coefficient presented by Livingston (1972) must not be used in computing the standard error of measurement, as Harris (1972) has pointed out. If the equation $s_{meas\ T} = s_T\sqrt{(1 - r_{tT})}$ is employed, r_{tT} must be the conventional reliability coefficent.

Livingston's criterion-referenced reliability coefficient is of value for indicating consistency of measurement in placing examinees above or below any designated dichotomic point (a passing mark, perhaps). The conventional reliability coefficient, on the other hand, indicates consistency of measurement in placing examinees across the range of obtained scores. Livingston (1972) has also noted that criterion-referenced correlation coefficients may be more meaningful than product-moment correlation coefficients for expressing the relationships among criterion-referenced tests when scores are expressed only in terms of a dichotomy such as "passing" or "failing."

In conclusion, we see that the interaction of measurement and instruction, which underlies the development of criterion-referenced measurements, has already brought new applications and extensions of classical test theory. The prospect of continuing changes in the field is what makes life interesting for psychometricians.

Papers Reviewed

Some additional valuable references furnished by the author are grouped separately following this list of the 1971 AERA papers reviewed in this summary.

Blumenfeld, G. J., & Others. Effect of criterion referenced testing upon the use of remedial exam opportunities. 10p. (ED 049 310; MF and HC available from EDRS).

Brennan, R. L., & Stolurow, L. M. An elementary decision process for the formative evaluation of an instructional system. 41p. (ED 048 343; MF and HC available from EDRS).

Hsu, T. C. Empirical data on criterion-referenced tests. 17p. (ED 050 139; MF and HC available from EDRS).

Nitko, A. J. A model for criterion-referenced tests based on use. 17p. (ED 049 318; MF and HC available from EDRS).

Roudabush, G. E., & Green, D. R. Some reliability problems in a criterion-referenced test. 13p. (ED 050 144; MF and HC available from EDRS).

References

Davis, F. B. Item analysis in relation to educational and psychological testing. *Psychological Bulletin*, 1952, 49, 97-121.

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

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington: American Council on Education, 1970. Pp. 625-670.

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

Hotelling, H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, 1940, 11, 271-283.

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

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.

Scandura, J. M., & Durnin, J. H. *Assessing behavior potential: Adequacy of basic theoretical assumptions*. Philadelphia: University of Pennsylvania, 1971.