DOCUMENT RESUME

ED 064 406 TM 001 654

AUTHOR Ivens, Stephen H.
TITLE A Pragmatic Approach to Criterion-Referenced Measures.
PUB DATE 72
NOTE 9p.; Paper presented at a symposium at a joint session of the annual meetings of the AERA and the National Council on Measurement in Education (Chicago, Ill., April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Behavioral Objectives; *Criterion Referenced Tests; *Item Analysis; Performance Criteria; *Test Construction; *Test Reliability

ABSTRACT

A discussion of criterion-referenced measures is presented. Two characteristics define the criterion-referenced measure: the presence of a performance criterion, and test items keyed to a set of behavioral objectives. The performance criterion, in an educational setting, is usually a relative standard of performance. There are two ways of constructing items for a criterion-referenced test: the item-form approach and the specification of objectives. Item reliability can be assessed by calculating the proportion of subjects whose items scores (pass or fail) are the same on a posttest and a retest, or on a posttest and a parallel form. A measure of score reliability can be obtained by calculating the mean item reliability; it may also be assessed using the concept of within-subject equivalence of total scores. Another index that can be used to assess item and test quality combines the concepts of reliability and validity. The most important part of a criterion-referenced measure is the set of behavioral objectives the measure is based on. These objectives set the stage for judging the effectiveness of the teacher's instruction, and evaluating the student's learning. (CK)

A PRAGMATIC APPROACH TO CRITERION-REFERENCED MEASURES[1]

Stephen H. Ivens
College Entrance Examination Board
Atlanta, Georgia

Measurement theory has traditionally concerned itself with the accurate estimation and interpretation of an individual's score in relation to the scores of other individuals. Measures yielding such scores have been known as norm-referenced. In contrast to norm-referenced measures are criterion-referenced measures that yield scores for which the interpretation is not dependent on their position in relation to other scores. The interpretation is, however, dependent on the specific content of the items in the measure and the degree to which the individual has attained criterion performance. Two characteristics, then, define a criterion-referenced measure: the presence of a performance criterion, and test items keyed to a set of behavioral objectives.

The performance criterion, in an educational setting, is usually a relative standard of performance. It is based on one's expectations, and revised when these expectations appear unrealistic. Although a criterion-referenced measure could be scored dichotomously, i.e., pass or fail, there is no reason why it cannot be scored as a norm-referenced measure.

There are essentially two different approaches to the construction of items for criterion-referenced measurers (see Popham, 1970). The first of these

---

[1] A Symposium Presentation at a joint session of the Annual Meetings of American Educational Research Association and the National Council on Measurement in Education, Chicago, Illinois, April, 1972.

approaches uses an item form to generate a population of items, all of which measure the same objective. The second approach is to generate the items by whatever means are available and, on an empirical basis, to revise or delete those items that do not perform as desired.

Regardless of the procedure used to construct criterion-referenced measures, traditional methods of evaluating norm-referenced measures may at times be inappropriate for criterion-referenced measures. Traditional methods depend on variability and criterion-referenced measures, in the ideal case, yield score distributions with zero variance. Even in less than ideal situations, criterion-referenced measures yield skewed distributions, with numerous identical scores, thus vitiating the application of traditional indices of item and test quality.

As mentioned earlier, there are two ways of constructing items for a criterion-referenced test and one's choice of these methods is primarily determined by the nature of the behavioral objectives. The item-form approach works well in areas like mathematics where the objectives can be very narrowly defined (e.g., Kriewall, 1969). In less structured content areas, however, the specification of objectives in such detail may not be feasible (e.g., Hills, 1970). In deference to the classroom teacher, it may not be practical to ask for such specificity for the pool of objectives would be much to large to handle easily.

If a pool of items keyed to an objective is generated by whatever means are available to the item writer, then item difficulty is an important concept. Within a pool of items on a given objective, it is certainly conceivable that the difficulty of some items may be more appropriate than

others, and that revisions or deletions may be advantageous. Such information can be obtained from pretest and posttest difficulty values for the items. Within each item pool, those items with difficulty values that are perceptibly different from the remaining items in the pool would be suspect. By using the remaining items in the pool as a control group, rival hypotheses such as prior knowledge or faulty instruction can be eliminated as being the determiners of such aberrant values.

In the optimal case, an item used in a criterion-referenced measure would have a zero or chance-level difficulty value on a pretest and a 1.00 value on the posttest. For such an item, it would be clear that instruction was needed, and that instruct was effective. A high difficulty value on the pretest would cause one to examine the item for specific determiners or some other clues which pointed to the answer. In the absence of these, one might conclude that instruction on the topic would be wasteful. A low difficulty value on the posttest would suggest that there were ambiguities in the item, that distractors were more similar than the distinctions that the student had been taught to make, or that there was a flaw in the instruction. An index as simple as the difference between the two difficulty values may be used as an item selection index for criterion-referenced test items. From a pool of six items on each of ten objectives, this author (1970) constructed two criterion-referenced tests using this difference index to select items. For each objective, the two items with the larger values went in the first form of the test and the two items with the lower values went in the second form of the test. Marked differences in the quality of the tests were

apparent when these tests were administered to a new sample of students.

Two separate studies (Cox and Vargas, 1966; Popham, 1970) have compared this difference between the upper and lower 27 percent who passed the item on the posttest. The findings indicate that the pretest-posttest difference index selects different items than traditional item-analysis indices based on an item's discriminating ability on a posttest only.

In norm-referenced testing, item-total correlations are computed to ask directly a question about the homogeneity of the items, and indirectly a question about the validity of each item. In criterion-referenced testing, item homogeneity is of primary concern when we are examining the items written for a given objective.

If the criterion-referenced measure is constructed without the use of the item-form, item homogeneity and content validity can be assessed through the pretest and posttest difficulty values. Once again, the pool of items on a given objective is used as the control against which each item is evaluated. For a given objective, similarly low difficulty values on the pretest and similarly high difficulty values on the posttest imply that the set of items is homogeneous.

Item homogeneity across objectives would be of concern if the objectives, for some reason, could be considered dependent on each other. The logic underlying such a dependency would necessitate the homogeneity of the items. A lack of homogeneity would point to one of two possible conclusions. Either the items were not adequately reflecting the objectives, or the objectives were sufficiently independent of each other to vitiate the

assumed dependency.

Item reliability can be assessed by calculating the proportion of subjects whose item scores (pass or fail) are the same on a posttest and a retest, or on a posttest and a parallel form. In the first instance, the index is a measure of item stability, and in the second, the index is a measure of item equivalence. In both cases, however, the maximum value of one would reflect perfect agreement across all subjects.

A measure of score reliability can be obtained by calculating the mean item reliability. An advantage to this method of calculating score reliability is that one is able to identify the particular items that are causing an undesirably low score reliability, thus allowing one to delete or revise those items.

Score reliability may also be assessed using the concept of within-subject equivalence of total scores. For each subject, the raw scores from two test administrations, either test-retest or parallel forms, would be converted into percent-correct scores. For each examinee, the absolute difference between the percent correct on the two administrations would be obtained. It is hoped, of course, that these percent-difference scores would be small--an indication of high reliability. The actual reliability index would consist of reporting the percent of subjects with percent-difference scores of a given size or less, e.g., a difference of 5 percent points or less. To compare reliabilities across tests, one might report two kinds of information for each method: the percent of scores agreeing within say 5 percent, and the percentage interval within which say 90 percent

of the scores agree. For example, it may be reported that for a given test, stability is reflected in that for 84 percent of the examinees, scores upon retesting after one week with no intervening instruction agree with scores on the earlier test within 5 percent, and that for 90 percent of the examinees, the retest score is within 8 percent of the score attained by that examinee on the earlier test.

The discussion so far has been concerned with possible analogues to the traditional concepts of item difficulty, item selection, and reliability. Another index that can be used to assess item and test quality combines the concepts of reliability and validity. This index requires three administrations of the same test to the same subjects; once as a pretest, once as a posttest, and once as a retest. If the test is functioning as expected, scores would be near the chance level on the pretest, and near mastery on the posttest and the retest. Thus, for each item and for each subject, we would expect a maximum change in performance from pretest to posttest, and a minimum change from posttest to retest.

The index consists of calculating for each item the value of the expression

$$(\underline{p}\,\text{post} - \underline{p}\,\text{pre})\ (1 - |\underline{p}\,\text{retest} - \underline{p}\,\text{post}|)$$

where $\underline{p}$ represents the proportion of subjects passing the given item on the particular administration. This index can range from a maximum value of one to a minimum values of minus one. Values less than zero can only be attained if the proportion of subjects passing the item on the pretest is greater than the proportion passing on the posttest--clearly an undesirable occurance in criterion-referenced testing.

As stated earlier, this index is a combination reflecting both reliability and validity. The first term in the expression is an index of validity in that it reflects performance between the pretest and the posttest. The second term reflects reliability (stability) in that it reflects performance from the posttest to the retest.

Although the previous discussion was concerned with the use of this index to assess item quality, it can be used to assess overall test quality. The test index is obtained by averaging the index values across all items of the test.

A similar measure of instructional effectiveness based on the ratio of actual gain to maximum possible gain from pretest to posttest has been suggested (see McGuigan and Peters, 1965; Brennan, 1970). Although this index may be useful, it appears to suffer from a lack of a theoretical basis for judging test effectiveness. This can be illustrated by the following hypothetical example. Assume that two tests, A and B, with maximum possible scores of 20 were administered as pretests and posttests to the same subjects. The pretest and posttest means for test A were 4 and 12, respectively, and the corresponding values for B were 12 and 16. This yields index values of .50 for both tests because test A showed a gain of 8 out of 16 possible points and test B showed a gain of 4 out of 8 possible points. Although the index values indicate the two tests were equally effective, there appears to be no rationale for such a decision. Further investigation is needed to determine what magnitude of gains in what part of the score scale constitute equal effectiveness.

The most important part of a criterion-referenced measure is the set of behavioral objectives the measure is based on. These objectives set the

stage for judging the effectiveness of the teacher's instruction, and evaluating the student's learning. Without carefully written objectives, the task of constructing a criterion-referenced test is self-defeating. Although the ideas presented in this paper may serve as an aid in assessing item and test quality for criterion-referenced measures, they cannot replace the creative artistry of the item writer.

REFERENCES

Brennan, R. L. Some statistical problems in the evaluation of self-instructional programs. Research Memorandum No. 1, June, 1970, Harvard University, Cambridge, Massachusetts

Cox, R. C., & Vargas, Julie S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Hills, J. R. Experience in small graduate classes and approaches to evaluating criterion-related measures. In Christine McGuire (Chm.), Criterion-related measures: Bane or boon? Symposium presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Ivens, S. H. An Investigation of Item Analysis Reliability and Validity in Relation to Criterion-Referenced Tests. (Doctoral dissertation, Florida State University) Tallahassee, Florida: University Microfilms, 1970, No.

Kriewall, T. E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report No. 103, October, 1969, Wisconsin Research and Development Center, Madison.

McGuigan, F. J., & Peters, J. Assessing the effectiveness of programmed tests--methodology--some findings. Journal of Programmed Instruction, 1965, 3 (1), 23-34.

Popham, W. J. Indices of adequacy for criterion-referenced tests items. In R. Glaser (Chm.), Criterion-referenced measurement: Emergining issues. Symposium presented at a joint session of the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Minneapolis, March, 1970.