DOCUMENT RESUME

ED 065 589 TM 001 858

AUTHOR	Oakland, Thomas
TITLE	An Evaluation of Available Models for Estimating the Reliability and Validity of Criterion Referenced Measures.
PUB DATE	Apr 72
NOTE	7p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Criterion Referenced Tests; *Evaluation Criteria; *Evaluation Methods; Item Analysis; Models; Norm Referenced Tests; *Test Reliability; *Test Validity

ABSTRACT

New strategies for evaluation criterion referenced measures (CRM) are discussed. These strategies examine the following issues: (1) the use of normed referenced measures (NRM) as CRM and then estimating the reliability and validity of such measures in terms of variance from an arbitrarily specified criterion score, (2) estimation of the reliability of single item tests, and (3) the modified use of item analyses as a means of judging the efficacy of instruction. (DB)

An Evaluation of Available Models for Estimating the Reliability and Validity of Criterion Referenced Measures[1]

Thomas Oakland
Department of Educational Psychology
The University of Texas

Various theoretical models and statistical methods have been proposed in an attempt to develop suitable strategies for evaluation criterion referenced measures (CRM). These proposals are an outgrowth of an awareness that CRM often have psychometric properties which are not directly amenable to classical methods for test validation (Randall, 1972). The new strategies discussed here examine the following issues: (1) the use of normed referenced measures (NRM) as CRM and then estimating the reliability and validity of such measures in terms of variance from an arbitrarily specified criterion score, (2) estimation of the reliability of single item tests, and (3) the modified use of item analyses as a means of judging the efficacy of instruction.

Livingston (1970, 1971, 1972) proposes that a NRM can be used as a CRM when the test user wants to compare each student's scores with an arbitrarily specified criterion score instead of comparing it in reference to the test's mean. A criterion score may be established above, at, or below the test's mean. Having established a specific criterion score, classical test theory techniques are employed for purposes of evaluation. Variance, covariance, and correlations are determined, based on deviations from the newly established criterion score, not on deviations from the test's mean.[2]

[1] Presented to the Annual Convention of the American Educational Research Association, Chicago, 1972.
[2] See Livingston (1972) for a redefinition of basic statistical concepts for CRM.

While the test validation procedures as outlined by Livingston appear to be viable under certain conditions,[1] they were judged to be inappropriate for our use for the following reasons: (1) an inability to find NRM that are suitable measures of our criterion behaviors and (2) Livingston's procedures are applicable for tests composed of many items rather than for single item tests.

During the last four years the test development section at the Laboratory has developed approximately 200 criterion referenced measures. It was necessary to develop these tests because relatively few existing NRM suitably measured our criterion behaviors (Kennedy, 1972). Probably few persons involved in research or evaluation activities have found existing NRM which adequately assess a specific criterion behavior. Due to this difficulty, the approach outlined by Livingston is limited.

CRM developed to assist in evaluation procedures must be precise enough to provide information regarding what persons can and cannot do in reference to specific criterion behaviors. The criterion behaviors which comprise our curricula are quite diverse. Therefore it is necessary to write one or more criterion referenced items for each criterion behavior. While we typically combine 10 to 15 criterion referenced items together in one test booklet, for economic reasons, each individual item within a test booklet is seen as a mini-test designed to measure a single criterion behavior. While NRM typically are composed of a number of items which measure a common ability or trait, the criterion referenced items within our test booklet are not necessarily a measure of a common ability or trait. Procedures based on classical test theory (such as Livingston's) are not

[1] See Harris (1972) for a critical discussion of Livingston's procedures.

readily amenable to tests in which items measure different objectives. Therefore, the procedures suggested by Livingston were not used.

Roudabush and Green (1971) discuss procedures for examining the reliability of CRM which consist of a series of items measuring separate behavioral objectives. Their discussion is directly related to the Prescriptive Mathematics Inventory (PMI), a CRM designed to assess approximately 400 mathematics objectives relevant to grades 4-8.

Their discussion pertains most directly to ways of detecting "false positives" (persons who answer an item correctly but have not mastered the criterion behavior) and "misses" (persons who answer an item incorrectly but have mastered the criterion behavior). False positives occur rarely on the PMI in that the probability of marking the correct answer by guessing is about one in a thousand. However, these cases may be detected by observing the processes students utilize in arriving at their answer. Detection of misses is more difficulty. Three methods are proposed. The first involves grouping similar items so as to have a larger sample of behavior pertaining to common objectives. For example, one may group 10 to 30 items measuring one's ability to add fractions; these would be grouped in a hierarchal order so that within this group the items become progressively more difficult. If a student misses an easier item while passing a series of more difficult items, the authors propose that one would conclude that the easier item was missed for some irrelevant reason and that the student actually posses the ability to answer the incorrect item.

We have found it exceedingly difficulty to establish a similar hierarchal order for our items. While a number of our tests contain items assessing related objectives which were designed in terms of the Taxonomy of Education Objectives: Cognitive Domain (Bloom, 1956), a significant number of items are not amenable to the classification system proposed by Roudabush and Green. They acknowledge the fact that "...A thorough going analyses of all the apriori relationships among all the items in the PMI would be voluminous and impossible to use."

The second approach suggested by Roudabush and Green utilizes point-biserial correlations to examine the relationship between items contained on essentially alternate form tests. It is difficult to draw any firm conclusion from their activities relative to this approach because of the poor quality and inadequate quantity of data. However, they report relatively low correlations, averaging in the middle 40s to 50s. Their results were sufficiently pessimistic so as to confirm our position that it was unwise for us to go through the expensive procedures of developing alternate form tests, in part, as a means of evaluating our CRM.

The third method they proposed involves the use of regression equations to predict item criterion scores. However, as yet they have not fully explicated this process.

Modified uses of item analysis have been suggested as a means of judging the extent to which a CRM assesses the effects of instruction (Popham & Husek, 1969) or as Cox and Vargas (1966) state, "...to identify items which discriminate between those needing training and those not needing training on the skill covered by each item." Cox and Vargas compared the results obtained from two item analysis procedures. Using both pretest and posttest scores, a Difference Index (DI) was obtained in two ways. A posttest minus pretest DI was obtained by subtracting the percent of students who passed an item on the pretest from the percent who passed the same item on the posttest. Also, a DI was obtained in the more conventional manner. After computing a total posttest score for each student, the distribution of scores was divided so as to identify the upper one-third and the lower one-third. Then the percent of students in the lower third was subtracted from the percent of students in the upper third on each posttest item. Spearman rank order correlations between the two sets of DIs were .37(N = 50, 31 items) and .40(N = 25, 40 items).[1] The authors conclude that their modified method of item analysis produces results sufficiently different from traditional methods to warrant its consideration for use with CRM.

While their proposed methods were considered, we felt that use of their methods was of limited use for the following reasons. Their methods are more appropriately used to determine the extent to which students may profit from instruction rather than to determine the reliability estimates which apply to a particular CRM. Also, the authors appear to be using their procedures to select the best items from a pool of available items; this selection process uses statistical procedures which are questionable for CRM (Popham & Husek, 1969; Randall, 1972). Also, the items finally selected may not assess adequately the full range of objectives which the test originally was designed to assess.

[1] Separate analyses were performed on two separate tests on which DIs were obtained using the procedure described above.

References

Bloom, B. (ed.), Taxonomy of Educational Objectives Handbook I: Cognitive Domain. New York: David McKay, 1956.

Cox, R. & Vargas, J. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Learning Research and Development Center, Reprint 7, University of Pittsburgh, 1966.

Harris, C. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, (1972, in press).

Kennedy, B. The role of criterion referenced measures within the total evaluation process. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Livingston, S. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, (1972, in press).

Livingston, S. A classical test-theory approach to criterion-referenced tests. Paper presented to the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Livingston, S. The reliability of criterion-referenced measures. Paper presented to the Annual Meeting of the American Educational Research Association, New York, 1971.

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

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

Randall, R. Contrasting norm referenced and criterion referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Roudabush, G. & Green, D. Some reliability problems in a criterion referenced test. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1971.