DOCUMENT RESUME

ED 052 253 TM 000 659

AUTHOR Dahl, Theodore
TITLE Toward an Evaluative Methodology for Criterion-Referenced Measures: Objective-Item Congruence.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
REPORT NO CSE-WP-15
PUB DATE Apr 71
NOTE 16p.; Paper presented at the Annual Meeting of the California Educational Research Association, San Diego, California 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests, *Decision Making, *Educational Objectives, Educational Testing, *Evaluation Techniques, Reliability, *Standardized Tests, Testing Problems, Testing Programs, Test Selection, Validity
IDENTIFIERS Monotonicity Analysis, *Objective Item Congruence

ABSTRACT

Consideration is given to the degree of congruence between objectives and test items in the evaluation of objective-based testing. An analytic technique that measures congruence is described, and a bibliography is provided. (CK)

TOWARD AN EVALUATIVE METHODOLOGY FOR CRITERION-REFERENCED MEASURES: OBJECTIVE-ITEM CONGRUENCE*

Theodore Dahl

Center for the Study of Evaluation

University of California, Los Angeles

CSE Working Paper No. 15

April, 1971

*Paper delivered at the Annual C.E.R.A. Convention, San Diego, California April 29 and 30, 1971.

Educational testing is a vast enterprise in the United States. Every year standardized tests are taken by millions of students across the country. In addition to the multi-million dollar expenditures necessary to accomplish this testing, vast amounts of other resources--both human and monetary--are allocated, reallocated, shifted, withdrawn, proposed or in some other way directly or indirectly influenced by the results of the tests.

The pervasiveness of testing in American education demands a thorough understanding of the meaning of the test results, which is necessarily contingent on the use of a pertinent test. Traditionally, school testing programs have been bound to standardized tests, which measure goals that are generally considered important over a large geographic area. Such tests are seldom completely relevant to local goals and should not be relied upon for strictly local use (Thorndike & Hagen, 1961, p. 289). This paper will first briefly summarize the reasons why standardized tests do not always yield the proper information to assist decision makers in planning a reasonable course of action. Secondly, the creation of a testing program which will alleviate some of the shortcomings of the standardized test will be suggested. Finally, some of the basic requirements of a testing program for supplying meaningful information to the educational decision maker will be considered.

STANDARDIZED TESTS

As the name implies, "standardization" is the key process in the construction of a standardized test. The first concern in constructing these tests is to determine what kind of achievement is to be measured within a content area such as mathematics or reading. More specifically, the test constructor must determine what should be measured for certain groups of students. The most uniform or standard method of defining groups is by grade level. The job, therefore, is reduced to determining what skills and what content are normally taught in a particular grade in schools across the country. Testing companies make considerable effort to ensure that their blueprint for constructing the test truly reflects the major skills and content areas taught in a majority of the schools which are potential users.

After the content of the test has been determined, test items are written to measure student acquisition of the skills, and expert opinion is sought to ensure that the items measure what they are supposed to measure. The test is then administered to an appropriate sample of students and an item analysis is performed. Items which are too easy or too hard, and hence do not yield any information relative to the comparison of students, are modified or discarded. In addition, the relationship of individual items to total test score is investigated. Items which are easier for students with a high total score than for students with a low total score are generally accepted, while items with the converse relationship are eliminated. An additional characteristic

desired in an item is that it should be easier for students in successive grades. For example, fourth graders should do better than third graders, and more poorly than fifth graders (Lindvall, 1967, p. 118).

After completion of the item analysis, another representative sample of students is selected and the modified test is administered. This group provides the norming sample, and their scores are used to determine the test score norms with which future examinees will be compared. Comparative scores may be reported several ways; percentiles and stanines are two widely used methods.

But what does this mean for a school district using a standardized test to assess the reading achievement of its students? The resultant scores indicate how well the students perform on the test relative to the performance of the norming group. There are at least three factors, however, which must be considered since they can greatly influence the meaning of the scores.

One factor is the degree of similarity between the students in the norming population and the students in the school district. Similarity is determined by such variables as socioeconomic status, ethnic heritage, and cultural background (Thorndike & Hagen, 1961, p. 154). Variables such as residence in urban, suburban, or rural neighborhoods are also important. Cox (1965) has also found sex to be an important variable. Obviously, the greater the discrepancy between the groups, the more caution one must use in comparing them.

A second factor concerns the degree of commonality between the skills and content included in the test, and the skills and content included in

the educational goals of the school system (Hopkins & Wilkerson, 1965). When local goals differ greatly from the goals represented by the standardized test, little information may be gained which will allow one to make statements about the success or failure of the local educational program.

Finally, the extent to which the items in the test measure what they are intended to measure is very important. Even though the items are reviewed for appropriateness by content experts, it must be remembered that an important criterion for inclusion of an item in a standardized test is often its difficulty level and its discrimination power (Lindvall, 1967; Cox & Vargas, 1966). Such practice may result in the elimination of items which are important in the assessment of achievement in favor of less important items. Cox (1965) demonstrated that selection of test items solely on the basis of discrimination capabilities may result in distortion of the intended content of the test.

The net effect of a standardized test, therefore, is that the school district has some comparative data which is more or less meaningful depending on a set of variables such as those listed above. The degree to which the test is appropriate and meaningful can be determined only if one is aware of the values of the several variables for the school district and test in question.

What kinds of decisions can be made on the basis of the information provided by a standardized test? If the district is attempting to determine how well their educational goals are being accomplished and/or how

instruction might be improved to better attain the goals, the answer is not an encouraging one. Gross decisions may be made as to the standing of the district relative to the rest of the country on the particular test in question. A very high score would indicate that the district is apparently doing well. An average score would probably indicate that there are few really serious problems, while a very low score would express a probable shortcoming in instructional outcomes.

The foregoing discussion of standardized tests illustrates some of their limitations in terms of providing information which can be used to facilitate instructional improvement and gains in learner achievement. This problem, we suggest, can be alleviated by the use of a test which is specifically designed to yield information concerning student achievement of specific instructional goals.

OBJECTIVE-BASED TESTING

One system which seems to offer the diagnostic function missing from the standardized test will be referred to here as "objective-based testing." As the term is used here, objective-based testing refers to a system wherein test results are indicative of student achievement of educational goals. One large difference between objective-based tests and standardized tests is that the latter are generally norm-referenced. That is, the scores of standardized tests usually are referenced to norms which tell how well the student performs with respect to other students. They do not demonstrate to what degree the student has mastered the goals and, thus, do not supply information useful for planning future instruction.

Objective-based tests, on the other hand, are criterion-referenced. That is, the scores are related to previously specified learning criteria; in this case the criteria are learning objectives. The distinction between norm and criterion reference is extremely important in determining if a given item will appear on a test. An item which might be eliminated from a standardized (norm-referenced) test because it does not discriminate for a given group could very well be included in an objective-based test for that same group because it measures one of the intended educational objectives (Glaser, 1963).

Another major distinction between standardized and objective-based tests lies in their respective blueprints for construction. As was mentioned earlier, the standardized test blueprint contains the skills and content normally taught across the country. A blueprint for an objective-based test, however, consists of the educational objectives of the user. Since the user dictates the content of the objective-based test, its relevance is much greater than that of the standardized test.

By properly blueprinting an objective-based test, and by properly constructing a test to fit the blueprint, a wealth of information may be gained by the user. Consider the advantages offered by an objective-based achievement test. The test score can be reported in several ways. For example, the total score might give a rough idea of the overall achievement of the students in a broad content area. In addition, subscale scores can also be reported for each objective. Such a partitioning of the total score into component scores provides the user with diagnostic information not normally available from standardized tests.

Further, by preparing the test for a specific user, certain objectives may be more rigorously measured than others if desired.

The construction of objective-based tests on a large scale will obviously require considerable resources and organization. A bank of objectives and related test items must be prepared in such a manner that a test may be generated upon the request of a user. Such a system is in preparation at the Center for the Study of Evaluation (CSE), which is associated with the UCLA Graduate School of Education.

A prototype system for a reading curriculum, beginning with pre-kindergarden objectives, is under development at CSE. A classification scheme has been constructed to help design the test blueprint in cooperation with the user. A coding system based on the classification scheme will be used to retrieve items for test construction.

OBJECTIVE-ITEM CONGRUENCE

As noted above, objective-based testing relies first on the specification of the objectives to be taught and then on the availability of test items to measure the attainment of the objectives. The adequacy with which the test items measure the objectives is of prime importance and cannot be over emphasized. While the score attained on a norm-referenced test acquires meaning through comparison with scores other students have achieved on the same test, the score on an objective-based test is taken as meaningful even in the absence of such comparisons (Glaser, 1963). Meaningfulness is not a mystical consequence of criterion reference, however. A relationship must be obtained between objectives

and test items which will equate achievement on the test to achievement of the objective. This relationship between objectives and items is referred to here as objective-item congruence, hereafter simply referred to as "congruence". The attainment and measurement of congruence is probably the most important consideration in the construction and use of objective-based tests.

Congruence denotes a correspondence between an objective and the items which are produced to measure the objective. Consequently, the items which measure the same objective should share a similar correspondence. Moreover--and this is an essential point--these items share a correspondence which they do not share with items designed to measure other objectives.

Such a conceptualization of congruence is admittedly vague. The explication of the concept is an important process which will be accomplished at CSE through a series of theoretical predictions and empirical verifications. This continued investigation may be divided into two major areas - the measurement of congruence per se, and the definition of variables which determine or affect the attainment of congruence. Obviously, the dichotomy is not a perfect one, as the two parts are by no means independent. Nevertheless, the two-pronged attack seems to be a useful way of looking at the problem. Time will permit consideration of only the first area of concern -- the actual measurement of congruence. However, a word about the effect of variables on congruence is necessary.

A behavioral objective states an educational goal in terms of measurable behavior which the student must exhibit in order to demonstrate achievement of the goal. The essential aspect of the objective is the student's behavior, and an item which is produced to measure that objective must therefore elicit the behavior if it is to be congruent with that objective (Cox, 1970). Any variable which operates to produce a student response contingent on factors other than the stated behavior tends to reduce congruence. Such a variable may be related to the item or to the objective. Studies will be conducted at CSE to identify the important variables and to establish their influence on the attainment of congruence.

Measurement of Congruence

There seems to be considerable agreement that an appropriate measure of congruence does not exist. Recently, Cox (1970, p 6) agreed that the "use of experimental techniques to establish the validity of criterion-reference measures should be investigated." Popham and Husek (1969) concluded that traditional views of validity and reliability would probably have to be changed to be useful in describing the value of a criterion-referenced test.

In order to develop methods for ascertaining the extent of congruence between objectives and test items, the definitions of congruence, along with related assumptions, must be used. First, it was assumed that for an item to be congruent with an objective it must measure the behavior specified in the objective. Second, congruence was conceived

to be a correspondence beyween the item and the objective. It was reasoned that the correspondence is somehow shared by items which measure the same objective.

Most of the research to date has been concentrated in the area of the first assumption--measurement of the stipulated behavior. Two methods are widely recommended to ascertain if an item measures the appropriate behavior and both are intuitively satisfying. First, a straightforward judgment of subject matter experts is frequently used. In fact, such judgment is often the only measure of congruence obtained by researchers constructing criterion-referenced tests (e.g., Cox, 1965). Attainment of judgmental evidence of congruence is essential to the success of an objective-based evaluation system. Measuring instruments must have face validity or they will not be accepted by the user. Thus, determining the ability of persons knowledgeable in the content area to predict congruence of items is an important consideration.

The second widely recommended, and sometimes used, method for determining whether items measure the stipulated behavior is the use of group comparisons. That is, a group of students who have successfully completed a program should perform much better than a group which has not (Glaser, 1963; Cox, 1965; Cox, 1970; Popham & Husek, 1969; Rahmlow, et al., 1970). Problems which are inherent in this approach include the means by which "successful completion" of the program is determined, and the effectiveness of the program or instruction. In spite of such problems, this kind of instructional evidence is important to maintain the credibility of claims made in regard to the congruence of items.

Very little work has been done in the area of the second assumption, i.e., the existence of a correspondence between items. Using analysis of variance, Hively, et al., (1968) attempted to determine homogeneity of items generated by item forms. Their results were inconclusive, based on a priori expectations of relative contributions of the different sources of variance to the total test variance. Virtually all of the remaining articles reviewed restricted the analysis of the items to the two methods described above.

One reason, apparently, that investigators have tended to avoid the fundamental question of underlying item relationships, has been a confusion between inherent properties of items and instructional effects on item statistics. That is, the correspondence between an item and an objective, as well as the correspondence among items, is assumed in this paper to be an inherent property of congruent items. Investigators sometimes lose sight of this when considering an instructional setting. Since the goal of criterion-referenced teaching is mastery by most or all of the students, the reasoning goes, any analytical method of item analysis which depends on test variance is not applicable (Popham & Husek, 1969).

We take the view, however, that the aforementioned correspondence is a property of the items regardless of the instructional methods or results. It is certainly true that in the exceptional case where all examinees pass all the items on a posttest, no relationship between items can be determined by present methodology. However, as Warrington

(1970) points out, individual differences in students make even an approximation to this situation unlikely in a classroom setting. Therefore, the studies undertaken by CSE will include the investigation of analytical methods of determining the correspondence between items. If such methods reveal the expected correspondences under normal conditions, it will be asserted that the same correspondences pertain under extreme conditions such as complete mastery.

Monotonicity Analysis

According to the definition of congruence presented in this paper, an analytical technique for the measurement of congruence should somehow cluster those items which are congruent to a common objective in such a manner as to be distinct from all other item clusters. Unfortunately, the common techniques of cluster analysis and factor analysis do not work for the type of data to be obtained from many objective-based tests. First, the data will be binary. Second, when learning objectives are very similar, the amount of common variance that links items which are congruent to the same objective is apt to be small compared to the common variance associated with general skills which subsume the objectives.

One procedure which appears to meet to meet the demands for the measurement of congruence is a relatively new development known as monotonicity analysis (Bentler, 1968). This is a monotonic factor

analytic technique, which is appropriate for binary data, and which has been used successfully to define factors where the shared variance of the variables was relatively small. The interested reader should refer to Bentler for further details of this technique.

Unfortunately, no information is available as to the adequacy of monotonicity analysis in the determination of congruence. We have just collected data from 200 subjects on a 120-item test, but the results of the monotonicity analysis have not been received. The results of this and other testings should be available sometime during the summer of 1971.

SUMMARY AND CONCLUSION

The use of standardized tests as a means of gathering information to aid educational decision making has been shown to be of limited value. Factors which limit the useful application of standardized test data include the necessity of obtaining high test score variance, the misrepresentativeness of norms, the lack of commonality between local educational goals and those measured by the test, and questionable content validity. Objective-based testing was profferred as an alternative which may alleviate many of the aforementioned problems.

Probably the most important consideration in the evaluation of an objective-based test is the degree of congruence between objectives and test items. An analytic technique known as monotonicity analysis

was suggested as a possible method for measuring congruence. Although no information is currently available regarding the efficacy of monotonicity analysis for this purpose, data are now being collected and analyzed. The results of the analyses should be available during the summer of 1971.

REFERENCES

Bentler, P. M. Monotonicity analysis: an alternative to factor and test analysis. Paper presented at Symposium on Ordinal Scales for Development, Monterey, California, 1968.

Cox, R. C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Cox, R. C. Evaluative aspects of criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

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

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

Hively, W. II., Patterson, H. L., & Page, S. H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 275-290.

Hopkins, K. D., & Wilkerson, C. J. Differential content validity: The California spelling test, an illustrative example. Educational and Psychological Measurement, 1965, 25, 413-419.

Lindvall, C. M. Measuring pupil achievement and aptitude. New York: Harcourt, Brace & World, 1967.

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

Rahmlow, H. F., Matthews, J. J., & Jung, S. M. An empirical investigation of items analysis in criterion-referenced tests. Paper presented at a joint session of the American Educational Research Association and the National council on Measurement in Education, Minneapolis, Minnesota, March, 1970.

Thorndike, R. L., & Hagen, E. Measurement and evaluation in psychology and education. 2nd ed., New York: Wiley & Sons, 1961.

Warrington, W. G. Criterion related measures: Some general considerations. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.