DOCUMENT RESUME

ED 093 923 TM 003 741

AUTHOR Graham, Carol L.
TITLE An Examination of the Feasibility of Using Criterion-Referenced Measurement in Large-Scale, Survey Testing Situations.
PUB DATE [Apr 74]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Comparative Testing; *Criterion Referenced Tests; Educational Assessment; Item Sampling; *Testing Problems; Test Validity
IDENTIFIERS Mastery Testing; *Survey Achievement Testing; Test Item Format; Test Length

ABSTRACT

The adequacy of a test developed for statewide assessment of basic mathematics skills was investigated. The test, comprised of multiple-choice items reflecting a series of behavioral objectives, was compared with a more extensive criterion measure generated from the same objectives by the application of a strict item sampling model. In many instances, the two instruments provided different classifications of students regarding mastery of an objective. Many of the discrepancies were attributed to the small number of items per objective and to the multiple-choice format of the original test. Consequently, the use of criterion-referenced tests in situations that severely limit test length and item format options was questioned. In addition, the problems associated with the practice of assuming content validity for criterion-referenced tests were discussed. (Author)

AN EXAMINATION OF THE FEASIBILITY OF USING CRITERION-REFERENCED MEASUREMENT IN LARGE-SCALE, SURVEY TESTING SITUATIONS

Darol L. Graham
Florida State University

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1974

An Examination of the Feasibility of Using Criterion-Referenced Measurement in Large-Scale, Survey Testing Situations

Darol Graham

Florida State University

The topic of criterion-referenced measurement has received considerable attention during the past decade. Much of the initial controversy that was generated over the relative merits of criterion-referenced and norm-referenced measurement appears to have subsided. Today, most psychometricians seemingly agree that criterion-referenced and norm-referenced measurement have different purposes, and that each is appropriate under the circumstances for which it was intended. Norm-referenced measures are generally more appropriate in selection situations while criterion-referenced instruments facilitate classification decisions regarding an examinee's position relative to a specified objettive. The determining factor in the selection of a measurement technique is the type of information required by the decision maker.

The value of direct or absolute measures of student achievement relative to an instructional objective has been demonstrated repeatedly for at least two types of educational decision making. Instructional developers need highly specific information about the attainment of educational objectives in order to validate learning materials. Likewise, instructional managers need detailed information about the status of each of their students for monitoring the achievement of prescribed learning objectives. Each of these decision makers has become dependent upon criterion-referenced measurement for acquiring the necessary performance information.

Recently, the application of criterion-referenced measurement has been extended to survey achievement testing situations such as state assessment programs. The educational accountability movement has created a need for specific information concerning the achievement of common educational objectives in order to establish minimum educational standards. In discussing the accountability issue, Hartnett (1971), described the movement of education toward operational statements of educational objectives as a basis for more precise measurement of educational effectiveness. A typical example of the movement is the "accountability act" and "state assessment" programs adopted in Florida.

In establishing objective-based state assessment programs, the objective-based measurement techniques that have proven so useful for making instructional development and management decisions provided an obvious tool. The logic of such an extension in the use of criterion-referenced instruments cannot be argued; however, the decision to employ such instruments was made without evidence of the suitability of criterion-referenced measurement for large-scale testing situations. Utilization of the technique for survey testing may present additional problems to the theoretical and methodological problems faced by all users of criterion-referenced instruments. In particular, the magnitude of data collected in survey testing practically dictates the nature of usable instruments. For cost efficiency the responses must be readily obtainable and machine scoreable: thus, a multiple choice or similar format for such instruments appears mandatory. Kriewall (1969), suggested that the measurement error introduced by tests of reasonable length with such a format, severely limits the reliability of decisions concerning the proficiency of individuals. The present paper addresses some of the problems associated with the adaptation of criterion-referenced measurement techniques to situations which require the collection of voluminous data such as survey achievement testing.

Method

The Florida State-Wide Eighth-Grade Test includes a section designed to assess basic mathematics skills considered essential for everyday living. The test was designed to measure nine skills which had been defined by a set of behavioral objectives. For each objective three multiple choice items were written to assess the skill identified by the objective.

For the present investigation, ten-item domain-referenced tests (Millman, 1973) were constructed to serve as criterion measures of a selected subset of the nine objectives. Construction of the items followed an item form approach (Hively, et.al., 1969; Osborn, 1968). Common wording was adopted for each item in a given criterion measure, but unique numbers were randomly generated for each item by a stratified sampling plan. In an effort to keep the items as similar as possible to the items found in the Eighth-Grade Test, numbers used in the criterion measure were restricted to a range consistent with the numbers in the Eighth-Grade Test. The results reported herein were obtained by administration of the two different measures of the following objectives:

1. Cost Comparison: Given the prices of two articles, the student will determine the difference in cost.

2. Travel Time: Given the distance between two points and a rate of travel, the student will determine the required travel time.

3. Time Differences: Given two times of day, the student will determine the differences in time.

1. Cost Comparison:

Format: Two articles are priced at \$ $\underline{\ P_1\ }$ and $\underline{\ P_2\ }$. What is the difference in cost of the two articles?

Parameters:

(1) Cost Difference (d):

$$d = P_1 - P_2$$

where, $\$0.01 \leq d \leq \$299.50$ (@ \$0.01 intervals)

(2) Cost of First Article (P_1):

$$P_1 = \$0.01\ a$$

where, $50 \leq a \leq 30000$

(3) Cost of Second Article (P_2):

$$P_2 = \$0.01\ b$$

where, $50 \leq b \leq 30000$

and, $b \neq a$

2. Travel Time:

Format: A car travels $\underline{\ d\ }$ miles at an average speed of $\underline{\ r\ }$ per hour. How many hours does the trip take?

Parameters:

(1) Travel Time (t):

$$t = d/r$$

where, 1/2 hr. $\leq t \leq$ 30 hrs. (@ 1/2 hr. intervals)

(2) Distance Traveled (d):

$$d = 10\ a$$

where, $2 \leq a \leq 30$

(3) Speed or Rate of Travel (r):

$$r = 10b$$

where, $1 \leq 6 \leq 8$

3. Time Difference:

Format: If the time is $\underline{\ t_1\ }$, how long will it be until $\underline{\ t_2\ }$?

Parameters:

(1) Time Difference (T):

$$T = t_2 - t_1$$

where, 1/4 hr. $\leq T \leq$ 12 hrs.

(2) Initial Time (t_1):
$$t_1 = 12{:}00 \text{ p.m.} + a/4 \text{ hrs.}$$
where, $0 \leq a \leq 95$

(3) Final Time (t_2):
$$t_2 = t_1 + b/4 \text{ hrs.}$$
where, $0 \leq b \leq 47$

The domain-referenced tests were developed to provide criteria for determining the concurrent validity of the objective-based subscales in the Florida Eighth-Grade Test. It was realized that any indication of the the validity of the subscales would be limited by the degree to which the criterion measure provided valid information concerning mastery of the objectives. Although the validity of the criterion measures could not be guaranteed, it was assumed that the specification and use of explicit item generation rules would at least facilitate the rendering of judgments about their apparent content validity. To the extent that the item generation rules reflect the original intent of the objectives, validity of the criterion measures would be expected to exist.

It was assumed that for a given objective there exists two populations, masters and non-masters. Based upon this assumption, a reliable test would produce two distinct distributions of scores, one for each population. Combining the observed scores of all examinees, i.e., both masters and non-masters, would be expected to produce a bimodal distribution with the mastery group receiving scores equal to the maximum possible score less the number of careless errors and the non-mastery group receiving scores of zero plus the number of lucky guesses. Thus, the degree of overlap of the two distributions could be taken as an indication of the amount of measurement error in the scores.

The criterion measures were administered to 151 eighth-grade students who had taken the Florida State-Wide Eighth-Grade Test. Compentency Classifications (i.e., mastery and non-mastery status) provided by the Eighth-Grade Test were compared with classifications obtained with the criterion measures of the same skills. Comparison of the two instruments was intended to provide an indication of the feasability of using survey tests for making criterion-referenced interpretations.

Results

Table 1 presents (1) the proportion of students declared masters of each objective according to their performance on the three-item sub scales of the Florida Eighth-Grade Test, (2) the proportion of students declared masters of each objective according to their performance on the ten-item criterion measures, (3) the proportion of cases in which examinees were given the same classification by both measures, and (4) the product moment correlations between the scores produced by the two measures. The Florida Eighth-Grade Program had specified a minimum standard for mastery classification of two out of three items correct. Primarily for consistency, a two-thirds standard, i e., seven out of ten items correct, was also adopted for the criterion measure. Other factors influencing selection of the cut-off score for the criterion measures are discussed in the next section.

Figures 1-3 present the distributions of scores obtained on the ten-item criterion measures for objectives 1-3 respectively. In addition, Figures 2 and 3 display the effect upon score distributions of broadening the objectives through modification of the item generation rules.

In Figure 2, the solid line indicates the score distribution produced when the domain of travel time items was restricted to the problem set having fractional solutions of one—half hour (e.g. $1\frac{1}{2}$ hr., $2\frac{1}{2}$ hr., etc.).

The score distribution represented by the broken line was produced by stratified sampling of items from the domain having both integer solutions and half-hour fractional solutions.

Table I

Indications of Agreement Between the Florida Eighth-Grade Test (E-GT) and Domain-Referenced Criterion Measures (CM) Concerning Examinee Proficiency of Certain Basic Mathematics Skills

Objective	Mastery Proportion on E-GT	Mastery Proportion on CM	Prop. of Agreement in Class.	Corr. Between E-GT & CM
1. Cost Difference	.91	.76	.84	.54
2. Travel Time	.85	.52	.65	.51
3. Time Difference	.74	.50	.68	.57

Figure 3 presents the score distribution for Objective 3 with a sample of items randomly generated from a domain containing various combinations of the following stratifications: (1) a.m. only, p.m. only, a.m. to p.m.; and p.m. to a.m., (2) time differences of whole hours, half hours, and quarter hours, and (3) initial times starting on the whole hour, half hour, and quarter hour.

Discussion

The instruments compared in the present investigation showed considerable discrepancy in the classification of examinees as masters or non-masters of the skills specified by the objectives. Both instruments had been judged to possess content validity by virtue of their apparent consistency with the pre-stated objectives. Undoubtedly, both of the tests were measuring the corresponding skills to some extent. Problems arose, however, because demonstration of the ability to perform a given objective often required

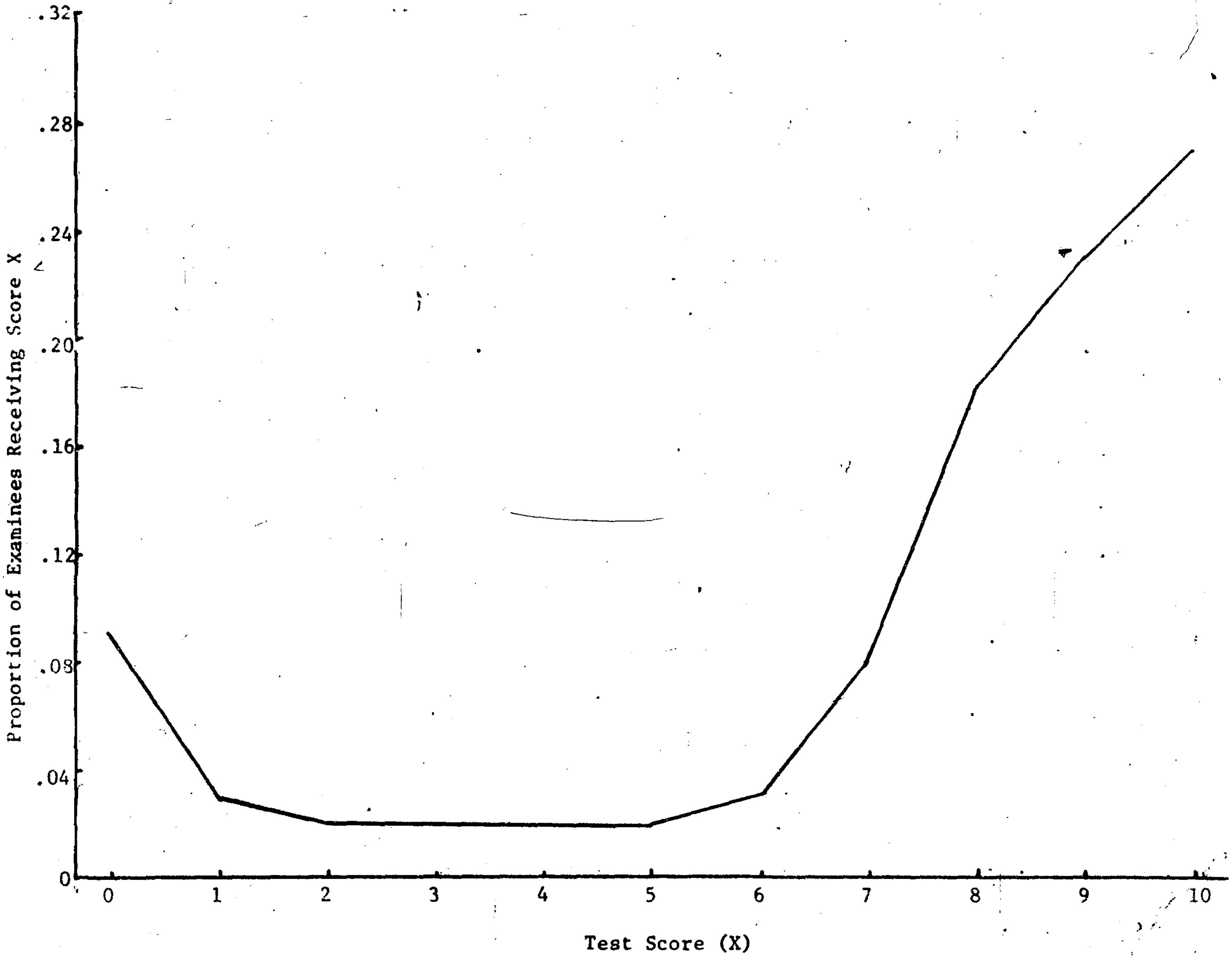


Figure 1. Proportion of Examinees Receiving a Score of X on Criterion Measure 1: Cost Difference.

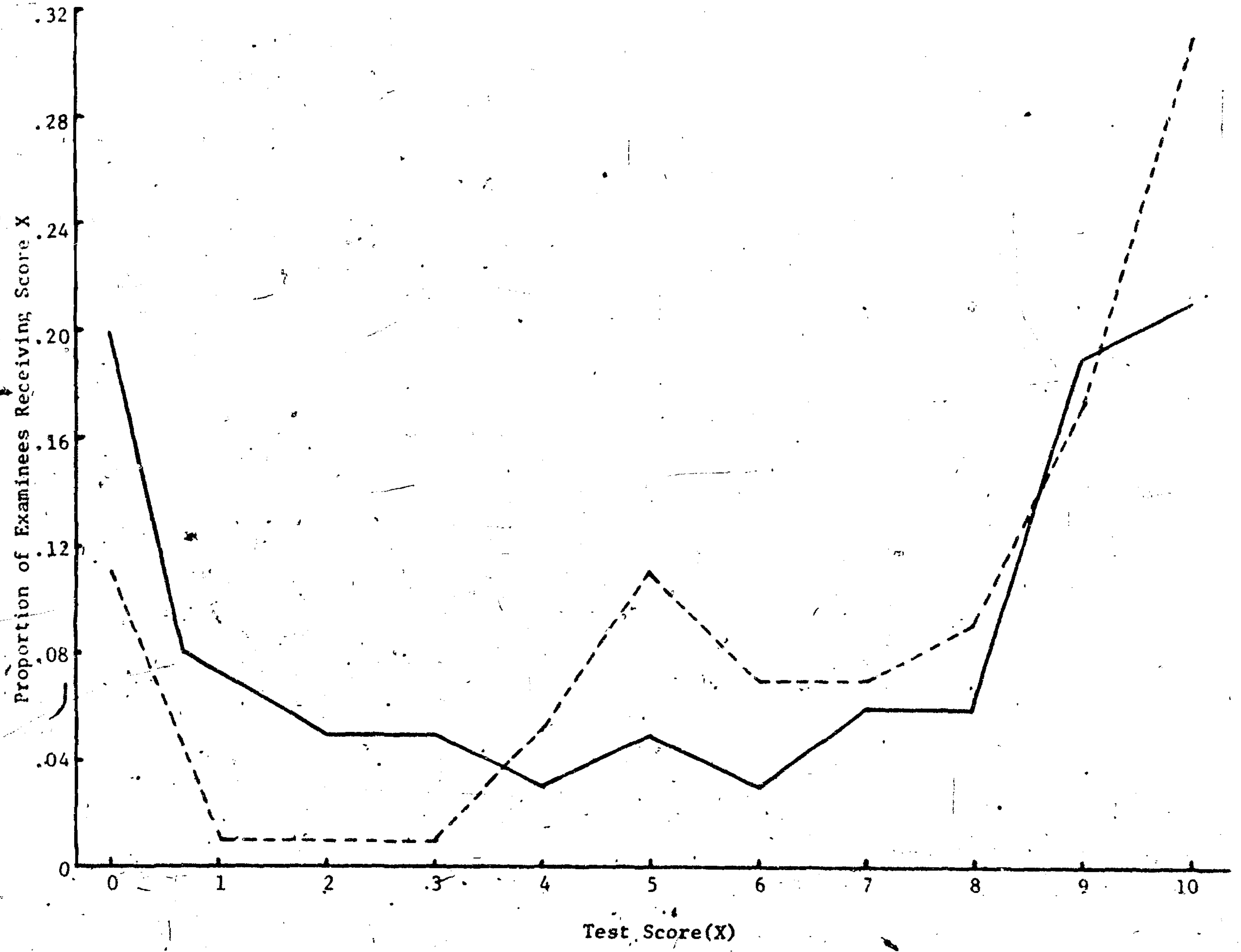
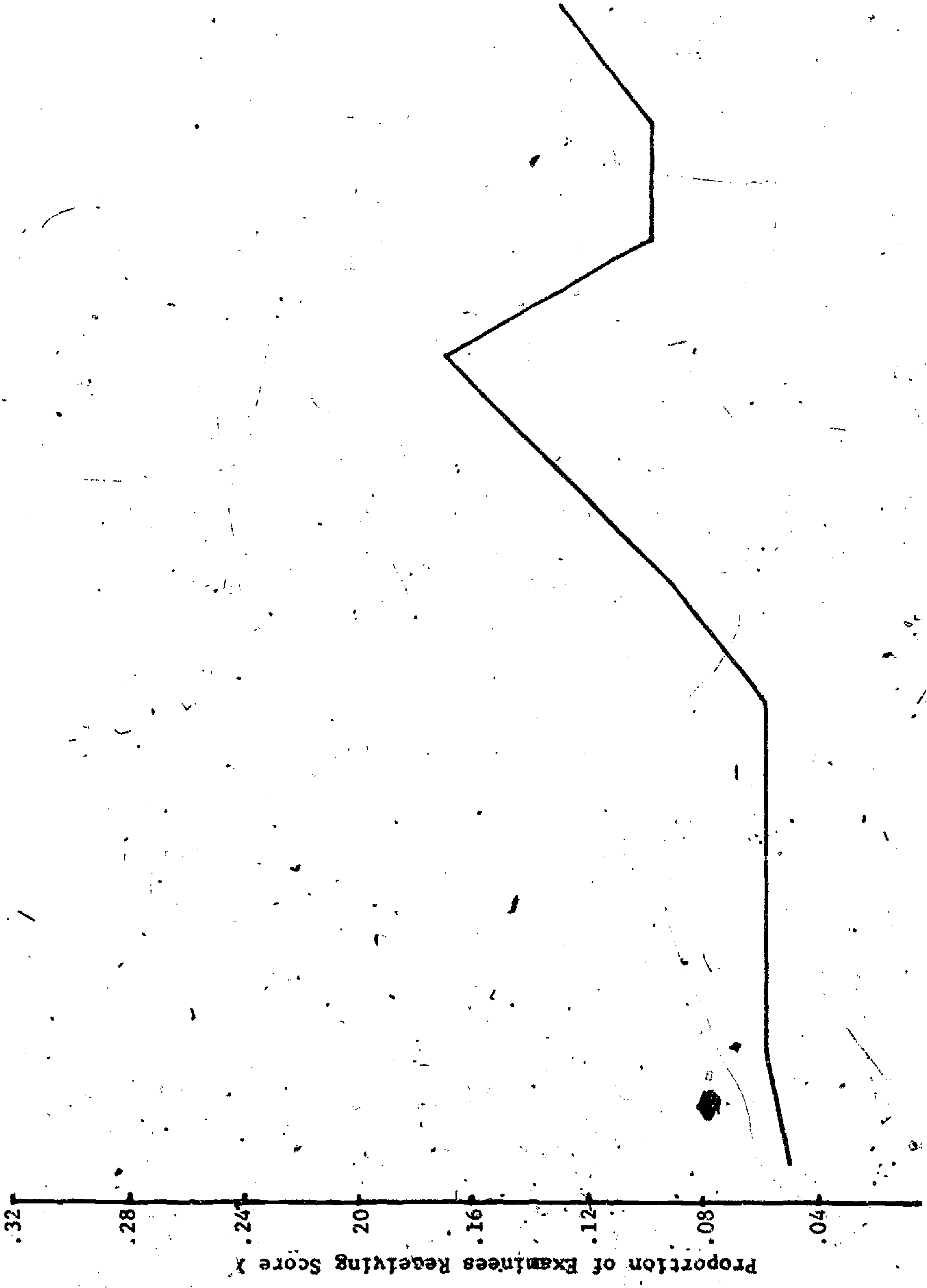


Figure 2. Proportion of Examinees Receiving a Score of X on Criterion Measure 2: Travel Time. (——— = Domain with integer solutions; ----- = Domain with integer and fractional solutions.)


.32
.28
.24
20
.16
.12
.08
.04
Proportion of Examinees Receiving Score X

subordinate or concommitant skills in addition to the primary skill specified by the objective. For example, the calculation of travel time, as specified by Objective 2, required the ability to perform certain basic mathematical computations in order to solve the verbal rate problems. As a result, minor changes of the item generation rules to include problems with whole-hour solutions as well as half-hour solutions produced quite noticeable changes in the score distributions. Thus, the broken line in Figure 2 seemingly identifies two types of masters of the skill of calculating travel time. One group of masters could solve travel time problems with either whole-hour or half-hour solutions while a lesser number of examinees could solve travel time problems but only for problems with integer solutions. It seems likely that the inclusion of problems involving other fractions that are less common than one-half would tend to confound the results even further.

Although representing different objectives, Figures 1 and 3 further demonstrate the effect of changing the item generation rules to broaden the domain of items included. The bimodal characteristics of the score distribution presented in Figure 1 suggest that parameters for Objective 1 define a rather narrow and homogeneous domain of items. In contrast, the measure of Objective 3, which included numbers representing three stratifications specified by the item generation rules, produced a more rectangular score distribution. Apparently, a number of examinees were able calculate time differences but either had not mastered the concept of a.m. and p.m. or had difficulty solving the problems that required the use of certain fractional portions of an hour.

It should be remembered that the verbal content of the items in each problem set used in the present investigation was held constant. Changes in the vocabulary and format of the items would be expected to exert additional influence upon the results. As the item domains increase in breadth and the items become more and more heterogeneors, this type of confounding influence tends to increase, and it becomes more and more difficult to make absolute statements about what examinees can and cannot do.

It is recommended that developers of criterion-referenced instruments devote considerable effort to activities that lead to increased precision of the objectives. It is often possible to employ procedures used in task analysis in the identification of capabilities that might be expected to influence performance of the skill identified by an objective. In particular, the test writer should look for pre-requisite capabilities that appear to be at a difficulty level that is relatively similar to that of the primary scale. For example, one might have predicted that the ability to manipulate mixed fractions would affect the performance of middle school students on travel time problems with fractional solutions. At the same time, one would not expect the inclusion of fractions in a set of wave mechanics problems to influence the performance of college physics majors. In instances where the potential influence of unspecified objectives is less obvious, it may be necessary to tryout the problems empirically in order to determine the extent of confounding for a given group of examinees.

The confounding of test results arising from the measurement of two or more skills simultaneously would be expected to increase as the item generation rules introduce more and more heterogeneity into the problem set. Since confounding increases the number of scores falling in the middle of the possible range, the degree of overlap between the mastery and non-mastery score distributions would also increase. Likewise, the number of scores at or near any selected mastery cut-off score would increase, thus increasing the likelihood of mis-classifying an individual with such an observed score. In this situation, classification results would be influenced to a much greater degree by the selection of a cut-off score. For example, if on a ten-item test no observed scores are found in the range from three to seven, the selection of any score within that range as the cut-off score will lead to the same classification of examinees. In the present investigation a cutting score of seven was arbitrarily adopted. Figures 1-3 suggest, however, that such a selection was fairly appropriate for minimizing the number of false positive classifications. If the consequences of a false negative classification were more important, a lower cutting score might be more suitable. In any event, for a homogeneous set of items such as the ones used to measure Objective 1, such changes in the cutting score adopted would have very little effect upon the results.

Reliability, in the sense of replicability of competency classifications relative to a given objective, would seem to be high for a bimodally distributed set of scores. In effect, each item from a homogeneous domain serves as a replication of the measurement of an individual's proficiency relative to a given objective. Naturally, such homogeneous measures are highly consistent internally, and as long as both masters and non-masters of a given objective are included in the test sample, KR-20 estimates of reliability will be high.

Much of the discrepancy in the classification of examinees that resulted from comparing performance on the two different measures can probably be attributed to the measurement error accompanying the subscales of the Florida Eighth-Grade Test. Primary factors contributing to this measurement error would be the use of three-item tests for each objective and the use of a multiple-choice format. Although the exact effect of the multiple-choice format upon the measurement of behavioral objectives cannot be determined, it seems likely that a test with such a format would require more items in order to yield a reliable measurement than would a test with a free response format.

Even with free response items, a number of factors appear to have an influence upon the number of items required to provide a reliable measure of a specified objective. First, as the objective becomes broader and the test becomes more heterogeneous, the length of the test must be increased to maintain measurement precision. Figure 1 suggests that even for highly homogeneous tests, four or five items may be necessary to minimize classification errors. Second, the number of items required to measure a given objective would also be influenced by the importance of the resulting decisions. For highly important decisions, where the consequences of misclassification are serious, the number of items would need to be increased. Finally, with the free response format, particularly in the measurement of mathematics objectives, test length may be related to the relative seriousness of type I and type II errors. For free response mathematics tests, the likelihood of careless errors would be far greater than the likelihood of lucky guesses. Thus, if false negatives are more serious than false positives, test length may need to be increased.

The adoption of criterion-referenced instruments for large-scale testing situations greatly increases the need for adequate theories and methodologies relating to criterion-referenced measurement. In classroom management situations, test quality is seldom critical. Other information sources provide a constant check on the criterion-referenced data. Since instructional management is a continuously ongoing process and most classroom decisions are of a temporary nature, decisions based upon invalid or inaccurate data can be readily modified at any time. On the other hand, survey testing often represents a single data collection effort and constitutes the sole information source for the decision maker. If the results of such testing are likely to have far-reaching effects upon the examinees or upon their schools or teachers, the integrity of the data is critical.

References

Hartnett, R. T. *Accountability in higher education: A consideration of some of the problems of assessing college impacts*. Princeton, N. J.: College Entrance Examination Board, 1971.

Hively, W., Patterson, H. L., & Page, S. H. A "universe-defined" system of arithemetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

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, Wisconsin.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95-104.