DOCUMENT RESUME

ED 060 025 TM 001 107

AUTHOR Hambleton, Ronald K.; Gorth, William P.
TITLE Criterion-Referenced Testing: Issues and Applications.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY Charles F. Kettering Foundation, Dayton, Ohio.
REPORT NO TR-No-13
PUB DATE Sep 71
NOTE 27p.; Version of this paper presented at the Annual Meeting of the Northeastern Educational Research Association, Liberty, New York, 1970

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Computer Assisted Instruction; *Criterion Referenced Tests; Curriculum Evaluation; Individualized Instruction; Item Analysis; Measurement Techniques; *Norm Referenced Tests; Performance Specifications; *Program Evaluation; Psychometrics; *Student Evaluation; *Test Construction; Test Interpretation; Test Reliability; Test Results; Test Validity
IDENTIFIERS CAM; *Comprehensive Achievement Monitoring

ABSTRACT

This paper highlights some special characteristics of criterion-referenced tests while comparing them with norm-referenced tests. Psychometric considerations involved in constructing a criterion-referenced test, including item analysis, reliability, and validity, are discussed, along with application of criterion-referenced testing to individual assessment and program evaluation. (Author/CK)

Technical Reports

No. 13

CRITERION-REFERENCED TESTING: ISSUES AND APPLICATIONS

Ronald K. Hambleton and William P. Gorth
University of Massachusetts

September, 1971

CENTER FOR EDUCATIONAL RESEARCH

School of Education
University of Massachusetts
Amherst

CRITERION-REFERENCED TESTING: ISSUES AND APPLICATIONS[1,2]

Ronald K. Hambleton and William P. Gorth

University of Massachusetts

---

[1] This study was supported, in part, by a research grant from the Charles F. Kettering Foundation to the Principal Investigator, Dwight W. Allen, Dean, School of Education, The University of Massachusetts, Amherst.

[2] A version of this paper was presented at the annual meeting of the Northeastern Educational Research Association, Liberty, New York, 1970.

Criterion-Referenced Testing: Issues and Applications

Ronald K. Hambleton and William P. Gorth
University of Massachusetts

Over the years, standard procedures for constructing, administering, and analyzing tests and interpreting scores have become well-known to educators. But recently there have been numerous suggestions for and demonstrations of instructional models in the schools where the usual procedures for constructing tests and interpreting test scores are not so useful and in some cases are completely inappropriate. Examples of these instructional models include: A Model of School Learning (Carroll, 1963, 1970), Individualized Instruction (Glaser, 1968), and Project PLAN (Flanagan, 1967, 1969). With these models, tests are being used for the purpose of establishing an individual's achievement on specified content, i.e. instructional objectives, and of providing information for making a variety of instructional decisions. Since traditional norm-referenced tests are clearly inappropriate, we have seen the development of a new kind of testing, criterion-referenced testing. Criterion-referenced tests are specifically designed to meet the measurement needs of the new instructional models. The criteria for the measurements are standards defined when the instructional objectives are specified. For this reason, the tests are called criterion-referenced.

The term, criterion-referenced test, was introduced by Glaser (1963) to make the distinction between tests designed to compare individuals and tests designed to measure individual achievement relative to some specified domain of tasks. Of the various definitions proposed for criterion-referenced tests (Kriewell, 1969; Livingston, 1970; Ivens, 1970) we prefer the definition proposed by Glaser and Nitko (1971). That is,

> A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards.

According to Glaser and Nitko (1971):

> Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain, and such measurements are referenced directly to this domain for each individual measured.

Defining well-specified content domains, developing procedures for generating appropriate samples of test items, and setting performance standards represent significant problems for measurement specialists but they will not be discussed in this paper. Papers by Millman (1970), Glaser and Nitko (1971), Hively, Patterson and Page (1966), and Bormuth (1970) have addressed some of these issues.

Unfortunately, because of their newness and some rather unique problems to be described later, there is a lack of information on matters such as test construction procedures and psychometric properties of criterion-referenced tests. Seldom do even the most recent educational measurement textbooks include more than one or two pages on the topic. According to Cronbach (1970), "The testing movement has given too much attention to comparative interpretations (to individual differences) and too little to absolute, criterion-referenced measurement." However, the need for such information is easily seen when one considers the fact that more and more schools each year are adopting the new instructional models.

This paper will integrate existing information on criterion-referenced testing with some original research results. It is organized around three topics: (1) a comparison of norm-referenced and criterion-referenced testing, (2) item analysis, reliability, and validity of criterion-referenced tests, and (3) a description of two applications of criterion-referenced testing.

A Comparison of Norm-Referenced and Criterion-Referenced Testing

Norm-Referenced Tests

Almost all of the available aptitude and achievement tests can be classified as norm-referenced because they are designed to measure individual differences. The meaning which can be attached to any particular score depends upon a comparison of that score to some relevant norm distribution. A norm-referenced test is constructed specifically to maximize the variability of test scores, since such a test is more likely to produce fewer errors in ordering the individuals on the measured ability. Since norm-referenced tests are often used for selection purposes, it follows that minimizing the number of order errors is extremely important.

It is a well-known fact that norm-referenced tests are constructed using the traditional item analysis procedures (Gulliksen, 1950; Lord and Novick, 1968). It is partly because of this fact that the test scores cannot be interpreted relative to some well-defined content domain since items are normally selected to produce tests with desired statistical properties rather than to be representative of some content domain. Both easy and difficult test items do not usually appear in norm-referenced tests because they contribute very little to test score variance. Also items which do not measure the same ability as the majority of other items in the test are usually removed. Empirical evidence to support these conclusions is provided by Cox (1965). His work revealed that the selection of items from a total item pool by classical item analysis procedures resulted in tests which contained proportions of items measuring instructional objectives different from those in the total item pool.

Criterion-Referenced Tests

The emphasis on mastery learning in the new instructional models has lead to an interest by measurement specialists in criterion-referenced testing. Criterion-referenced tests can be used to serve two purposes. First, they

can be used to provide very specific information on the performance levels of individuals on the instructional objectives. This information can be used, for example, to determine whether an individual has "mastered" particular objectives (Block, 1971).

Second, criterion-referenced tests can be used to evaluate the effectiveness of instruction. Norm-referenced tests given at the end of a course are useless for making evaluative decisions on the effectiveness of instruction because they are not tailored to the instructional objectives. However, criterion-referenced tests combined possibly with the notion of item-examinee sampling are useful to the curriculum evaluator because of the specificity of the results to the instructional objectives (Lord, 1962; Cronbach, 1963; Shoemaker 1970a, 1970b; Hambleton, Rovinelli, and Gorth, 1971; and Gorth, Schriber, and O'Reilly, 1971).

What are the appropriate procedures for constructing a criterion-referenced test? It should be clear that since a score on a criterion-referenced test is compared to some performance standard rather than to the performance of other individuals that for the test to be a good measuring instrument it will be necessary to change the item selection and test construction procedures. However, it is only recently that any attention has been given to the problem (Hively, Patterson and Page, 1968; Bormuth, 1970; Lindeman, Gorth and Allen, 1969).

Since comparisons among individuals are of little or no interest when using a criterion-referenced test, it follows that a test constructor is not usually concerned with developing a test to maximize the variance of test scores. Therefore, a test developer cannot use classical item analysis procedures to choose items because they were specifically designed to result in a test with maximum variance of test scores. For example, criterion-referenced tests are often used either before students are taught specific instructional

objectives or immediately after students are taught specific instructional objectives. In the former situation, most students will answer few or none of the test items, i.e., low total scores, and in the latter situation, they will answer most or all of the items, i.e., high total scores. Both situations produce very little variation in total test scores within the group of students. Consequently, item discrimination indices, the biserial and point biserial correlation coefficients, will be very close to zero for most items which is considered an indication of a poor test item in classical test theory. However, item statistics based on correlational methods can be of some use in detecting poor items given that different standards are used to interpret the indices. More will be said about this and other psychometric issues in later sections.

Some measurement specialists have discussed criterion-referenced tests as ones which would be scalable in a Guttman sense (Popham and Husek, 1969; Guttman, 1950). In this case, knowing an individual's test-score would be sufficient information to reproduce his response pattern. We would know precisely which items he answered correctly and incorrectly. While this kind of test would be excellent for diagnostic purposes, these tests are difficult to construct (Cox and Graham, 1966).

More typically, the items on a criterion-referenced test can be thought of as a sample from some well-defined content domain. Knowing a student's test score does not allow us to accurately say which items were answered correctly, but we can make a pretty good estimate of the proportion of items in the domain that he could answer (Popham and Husek, 1969).

It would seem that what is needed now is some test theory developed specifically for criterion-referenced tests. Some progress has been made in this direction by Cronbach and Gleser (1965), Kriewell (1969), Glaser and Nitko (1971) and Hambleton and Novick (1971).

A Summary

While admitting that a test cannot be classified as either a norm-referenced or criterion-referenced test by simply looking at it, the two kinds of tests are designed for quite different reasons and constructed using different procedures. The norm-referenced test is constructed using traditional item analysis procedures for the purpose of making comparisons among individuals. In contrast, a criterion-referenced test is designed to facilitate decision-making relating to individual performance and effectiveness of instruction. Procedures for constructing the tests are only now being developed.

It is interesting to note, however, that criterion-referenced tests can be used to make comparisons among individuals and norm-referenced tests can be used to measure the extent to which individuals master instructional objectives. But, since the purpose of criterion-referenced tests and norm-referenced tests is basically different, one would in most cases be a weak substitute for the other.

Item Analysis, Reliability, and Validity

Item Analysis

Since the traditional approach to item analysis is of limited usefulness in developing criterion-referenced tests other procedures needed to be developed. Three approaches to item analysis of criterion-referenced tests will be discussed in this section: (1) modification of traditional item analysis procedures, (2) selecting items to measure change, and (3) item characteristic curves.

Modification of traditional item analysis procedures. In criterion-referenced test development the item difficulty index is useful for selecting "good" items. However, the item difficulty is used somewhat differently than when one is constructing a norm-referenced test. In that case, items with moderate difficulty are preferred because they increase the discriminating

power of the test. If such a strategy were employed in constructing a criterion-referenced test there is every likelihood that many of the best items would not be selected. How should the item difficulty index be used? If the content-domain is carefully specified, test items written to measure accomplishment of the objectives should also be carefully specified and closely associated with the objectives. Therefore all of the items associated with the same objective should be answered correctly by about the same proportion of examinees in a group, i.e., they should have approximately the same value for the item difficulty index. If an item has a value of the index quite different from all of the other items, it probably is measuring a performance which is identifiably different from the objective. If the indices of the items associated with an objective differ, several alternatives may be followed. Either the items which are least like the objective should be modified; (the item difficulty index would be obtained on the modified items and compared with the unaltered items for congruency) or the objective written more specifically to refer only to similar items with similar indices. Thus, the item difficulty index may be used in a new way to refine the items associated with an objective.

Similarly the item discrimination indices, mentioned earlier, can be useful in item analysis for criterion-referenced test construction, although they were developed specifically for norm-referenced tests. Negative discrimination indices serve as "warning flags" that items included on a criterion-referenced test may need modification. (There is also the possibility that a negative discrimination index is an indication of ineffective teaching and/or ineffective instructional materials.) The negative value indicates that students who have generally done best on the total test answered the item incorrectly more frequently than the students who did poorly on the test. A positive discrimination index is still meaningful; however, it is more likely to indicate some shortcoming of the instructional program. This follows since most of the new instructional programs using criterion-referenced tests are

designed to minimize post-test differences in achievement. (This is done by individualizing instruction to the extent that variables such as pace, sequence, and the instructional mode are optimally chosen for each individual.) Zero discriminating items may be quite acceptable for criterion-referenced tests.

Selecting items to measure change. To demonstrate the effectiveness of instruction, evaluators attempt to construct criterion-referenced tests which give very different total scores before and after instruction. A number of researchers have been concerned with item analysis and selection procedures for constructing these kinds of tests (Cox and Vargas, 1966). An interesting question concerns whether or not it matters what techniques are used to select items. That is, given a large pool of test items, how similar would the selection of items be if different item statistics were used. There is some evidence from a study by Englehart (1965) to suggest that with norm-referenced tests there is a high degree of agreement among items selected with various discrimination indices. For criterion-referenced tests is the situation similar?

Cox and Vargas (1965) investigated the effect of employing different item selection techniques to identify items for norm- and criterion-referenced tests and the extent to which two methods of item analysis yielded the same relative evaluation of items. Discrimination indices were computed for items on tests which has been administered as pre-tests and post-tests in an individualized instruction program. The first index was the common D statistic (Englehart, 1965) computed for items on the post-test data only. The second index was the difference in item difficulty between the pre-test and post-test data. (They also investigated a third index but it is of no interest here.) The results indicated that some items which are highly desirable for criterion-referenced tests would be discarded on the basis of their D statistic because they fail to discriminate between individuals. According to Cox (1970), "The pre-and post-test

method of item analysis produced results sufficiently different from traditional methods to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures."

Using the same methodology but different test items and groups of examinees, the Cox and Vargas (1966) study was replicated and extended to provide the results reported below. The test items came from two mathematics areas, algebra and trigonometry. The algebra test items were administered to 110, 11th grade students at Hopkins High School in Minneapolis, Minnesota. The trigonometry test items were administered to 102, 11th grade students at Kailua High School in Kailua, Hawaii. The items were administered to the students three times: (1) a pre-test, (2) an immediate post-test, and (3) a delayed post-test about one month after instruction.

The three item statistics considered in the study were r_g, p'_g, and p''_g where:

r_g = the biserial correlation for item g on the post-test,

p'_g = the difference between the proportion of individuals who correctly answered item g on the post-test and the pre-test, and

p''_g = the difference between the proportion of individuals who correctly answered item g on the delayed post-test and the pre-test.

From Table 1 it is apparent that there is little relationship between r_g and p'_g or r_g and p''_g for either set of test items. The correlation between p'_g and p''_g is higher than the other two but the statistics are based on the same pre-test data.

Tables 2 and 3 report the similarity of items selected using the three item statistics for test made up of 25%, 50%, and 75% of the initial item pool. It is clear from the results that the choice of statistics has a significant effect on the final selection of test items.

Table 1

Spearman's Rank Correlations Among
Three Sets of Item Parameters

	Algebra Test		Trigonometry Test	
Indices	Number of Items	Correlation	Number of Items	Correlation
r_g and p'_g	57	.38**	75	-.26*
r_g and p''_g	57	.28*	75	-.31**
p'_g and p''_g	57	.78**	75	.68**

*p<.05
**p<.01

Table 2

Percentage of Overlap Between Items Selected According to Each Pair of Item Analysis Indices (Algebra Test - 57 Items)

Proportion of the original item pool selected in the test	Baseline: Minimum possible overlap	r_g and p'_g	r_g and p''_g	p'_g and p''_g
1/4 (14 items)	0%	35.7%	35.7%	71.4%
1/2 (28 items)	0%	58.6%	62.0%	79.3%
3/4 (43 items)	66.7%	86.0%	81.4%	86.0%

Table 3

Percentage of Overlap Between Items Selected According to Each Pair of Item Analysis Indices (Trigonometry Test - 75 items)

Proportion of the original item pool selected in the test	Baseline: Minimum possible overlap	r_g and p'_g	r_g and p''_g	p'_g and p''_g
1/4 (19 items)	0%	21.0%	5.2%	57.9%
1/2 (37 items)	0%	39.5%	39.5%	76.3%
3/4 (57 items)	66.7%	71.9%	71.9%	90.0%

In summary, the differences are not surprising, but the magnitude of the differences is. This emphasizes the importance of choosing the appropriate item statistics to select items for criterion-referenced tests. Although Cox and Vargas (1966) endorse the change in item difficulty index as a criterion for item selection, they do point out, "the need for developmental work on item analysis procedures when only one test administration is possible."

Item characteristic curve. One of the more interesting suggestions for item analysis of criterion-referenced tests was made Wardrop (1970). He suggested that the item characteristic curve might be a useful alternative to some of the traditional item analysis procedures.

The notion of an item characteristic curve comes from the work of Lord (1952, 1968), Birnbaum (1968) and others in the area of latent trait theory. For the case of a unidimensional test, a latent trait model specifies a function which relates the probability of success on an item to the underlying latent trait or ability which the test measures. The choice of different mathematical forms for the item characteristic curve has led to the development of different latent trait models (Lord and Novick, 1968). The latent trait or ability for each individual could be conceptualized as his position on an ability scale ranging "from no proficiency at all to perfect performance" (Glaser, 1963). The measurement problem is to locate the individual in the correct location on the ability continuum.

As suggested earlier, various functions have been proposed for the item characteristic curve. For example, Birnbaum chose a two-parameter logistic curve,

$$P_g(x) = [1 + e^{-Da_g(x-b_g)}]^{-1}$$

as the form of the item characteristic curve in his model where $P_g(x)$, is the probability that an examinee with ability x answers item g correctly. The parameter b_g is usually referred to as the index of item difficulty, whereas a_g is referred to as an index of item discrimination. (The constant D is a

scaling factor.) What limited empirical work has been done on various latent trait models reveals fairly good fits to real data (Ross, 1966; Wright, 1968; Lord, 1968; and Hambleton and Traub, 1970). Information on assumptions underlying the latent trait models are discussed by Lord and Novick (1968).

Why is this such an attractive approach? First, in theory at least, the item parameters (difficulty and discrimination) remain invariant from group to group which is certainly not generally true of traditional item parameters. For example, the conventional item difficulty, defined as the proportion of examinees in a group who correctly answer the item, varies as a function of the ability of the group. The invariance of the item difficulty parameter would permit the construction of tests with specific characteristics without prior knowledge of the ability of the examinees. Also, it is theoretically possible to measure growth using the latent trait ability scale because it is an interval scale.

An important problem to solve before this particular approach to item analysis and ability estimation becomes practical is the development of an efficient procedure for estimating item parameters and abilities. Some progress on the problem has been made by Lord (1968) and by Bock (1971). Another problem for research concerns the empirical verification of the invariance property of the item parameters.

Reliability

In many situations where criterion-referenced tests are used, there is little or no test score variance. And, since it is well-known that the size of a reliability coefficient depends, among other things, on the variance of test scores, it is apparent that the common approaches to estimating reliability (such as internal consistency and parallel-form) will be of limited usefulness.

As Carver (1970) points out, the reliability of any test depends upon replicability, but replicability is not dependent upon test score variability.

If a group of examinees all obtained similar scores on parallel forms of some test, near perfect replicability exists even though test reliability, estimated using traditional methods, would be close to zero. This rather extreme example points out the shortcoming of traditional reliability indices and serves to indicate the need for the development of alternate approaches.

Cox and Graham (1966) report the use of the coefficient of reproducibility as an alternative to the classical approach to reliability estimation for one special type of criterion-referenced test. They calculate the coefficient for a sequentially scaled achievement test designed for use in an instructional model where performance objectives can be identified as being sequential in nature. Tests are said to be scalable if for a particular ordering of items, individuals are able to answer all questions up to a point and none beyond. The coefficient of reproducibility is a measure of the extent to which group performance satisfies this condition. As Cox (1970) says, "the pitfalls of using reproducibility as a reliability estimate for achievement tests have not been explored."

Validity

As in the case of reliability, the validity of criterion-referenced test scores will probably need to be determined by non-correlational techniques. This follows because of the lack of test score variance but also because in some cases, validity is best determined by alternative means such as assessing decision-making accuracy (Hambleton and Novick, 1971).

Above all else, a criterion-referenced test must have content validity. According to Popham and Husek (1969), content validity is determined by, "a carefully made judgment, based on the test's apparent relevance to the behaviors legitimately inferable from those delimited by the criterion." If techniques such as those advocated by Hively, Patterson and Page (1968) or Bormuth (1970) for defining content domains and item generation rules are followed, content

validity follows. If other procedures are used, the task of determining content validity becomes much more difficult.

For determining predictive and construct validity of criterion-referenced tests, both a non-correlational approach to validation and a suitable criterion must be found. Cox (1970) has suggested the use of experimental procedures to establish validity of a criterion-referenced test. For example, given that teaching is effective, one might determine the construct validity of a criterion-referenced test by observing the difference in performance between students who have been exposed to instruction and those who have not. The bigger the difference the more valid the test could be said to be.

Some Uses for Criterion-Referenced Testing

In this final part of the paper we will consider the application of criterion-referenced tests in the areas of individual assessment and program evaluation.

Individual Assessment

A new instructional model is the one used in the Jamesville-DeWitt (JD) High School in Syracuse, New York State (O'Reilly and Hambleton, 1971) in the 9th grade science course. It is organized into modules which consist of a series of instructional activities arranged into a hierarchy of objectives leading to mastery of a single concept or group of related concepts. The day to day instructional activities which, when taken together make up a module, are organized into a hierarchy of smaller submodules called learning activity packages (LAPs). Within each instructional module are four types of decisions. To provide information for decision-making the following criterion-referenced tests are administered: a module pretest, a module posttest, and several LAP pretests and LAP posttests.

Briefly let us consider each decision separately. As a student begins to work on a module, a module pretest is administered. Since items in the module

pretest are closely tied to the objectives of all of the LAPs in the module the student's correct responses to items measuring the objectives in a LAP would be used to decide to omit the corresponding LAP from the student's prescribed activities for the module. Such a procedure will insure that students will be working only on learning experiences directed toward goals which have not been mastered previously. The module posttest which is either the same test or a parallel form of the module pretest can be used for prescribing remedial work for a student, for grading, and for evaluating the effectiveness of instruction in the LAPs.

Analogous to the module pretests, the LAP pretests are used to prescribe a set of objectives within the LAP that the student must demonstrate competency in before moving on to the next LAP in his prescription. LAP posttests are used to determine the extent to which students have satisfactorily completed the objectives of the LAP.

The four decisions just described might conveniently be classified as either placement or mastery. Decisions relating to the diagnosis of learning difficulty can also be made from the criterion-referenced tests if the incorrect responses to the items have been carefully constructed, i.e., incorrect choices are included in an item because they are indicative of particular learning difficulties. Apparently this systematic construction of distractors for the purpose of diagnosing learning difficulty has not been carefully explored but offers much potential. In addition to being an excellent way of extracting more information from a criterion-referenced test, it offers a systematic way for constructing item alternatives.

One problem that remains to be solved for programs similar in format to the JD Model is the development of guidelines for establishing cut-off points (i.e., how many items must an individual pass to demonstrate mastery). At least one researcher (Nitko, 1971) has suggested different cut-off points for different individuals.

Another problem found in some of the new instructional models is the extensive amount of time which is taken up by testing. Although testing provides data for decision-making and one wants to maximize the number of correct decisions, it is apparent that the cost in terms of time is too much to allow tests to be of the length necessary to insure low probabilities of error for all types of decisions. Assuming that tests can be weighted according to their importance it should be possible to derive optimum test lengths for the case when the total testing time is fixed.

While increasing test length is an obvious way of reducing errors in decision-making, alternate means include tailored testing (Lord, 1969; Ferguson, 1969), differential weighting of response alternatives, and confidence testing (Wang and Stanley, 1970; Hambleton et al, 1970). All three approaches can be used with criterion-referenced test items, are based on intuitively appealing ideas, and offer more information per item on each examinee. However, there is little empirical data to support any of the approaches.

Comprehensive Achievement Monitoring (CAM)

Gorth, Schriber, and O'Reilly (1971) describe a model for the evaluation of student achievement in classrooms and for curriculum evaluation called Comprehensive Achievement Monitoring (CAM). All of the decision-making is made on the basis of criterion-referenced test results. The CAM design includes the following components:

1. The definition of a curriculum with behavioral objectives;
2. The writing of test items to measure student performance on each objective which are criterion-referenced test items;
3. The organization of a set of randomly parallel tests, where each test is made up of items measuring all or a sample of all of the objectives in the curriculum and therefore represents item sampling;

4. The design of longitudinal, usually every three or four weeks, schedule of test occasions throughout the course;

5. The analysis of the test data and the reporting of results by computer, usually within a couple of days;

6. The interpretation of the results by evaluators, teachers and students as a means for making better decisions about their instruction and curriculum; and

7. The modification of curriculum, instructional activities and the CAM design based upon the results.

The CAM methodology has been designed to work well with any grade level or curricular area. In fact, it has already been used successfully in more than 20 schools, with more than 15,000 participating students, and at grade levels from 3rd to 12th and in every academic subject area (Allen and Gorth, 1971). (Hambleton, Gorth and O'Reilly [1971] provide a detailed report on one of the many applications.)

Particularly important to the success of the evaluation is the use of the computer. It alleviates the frequently encountered bottlenecks of most evaluations, i.e., the analysis of data and the reporting of results. The computer allows maximum freedom in the design of evaluation which CAM has used by incorporating longitudinal testing with item sampling.

The information which is provided in the CAM system includes: (1) for individual students, (a) the total score on the current test and all previous tests, and (b) information on the correctness of their response to each item corresponding to course objectives on the current test; and (2) for any subgroup of students and any set of questions after each test administration, (a) the achievement level on each objective, and (b) achievement profiles which display graphically the level of achievement on all objectives on the previous test occasions.

The computer allows students' achievement to be plotted on any given objective (or group of objectives) for the entire course. This plot, called an achievement profile, gives a graphic presentation of the changes in group achievement throughout the course. Achievement profiles are a unique type

of information available from the CAM model.

Figure 1 presents hypothetical achievement profiles for four objectives from a course. In this example, objective 1 was taught between the first and second test administrations, objective 3 between the third and fourth testing and objective 4 between the fourth and fifth. For the reason given below objective 2 was not taught. On the pre-test in the example, all objectives except number two show achievement at the chance level or about 20% on the five option multiple-choice items. Using the achievement profiles after the second test administration the following decisions might be made: (a) objective 1 was not learned and should probably be retaught in a somewhat different way; (h) since the performance level on objective 2 was high on both the first and second test administrations one could safely skip instruction on it. After the sixth testing on the basis of the CAM data the following decision could be made: (a) the performance level on objective 3 is slipping and since it is an important objective it should be reviewed. It is also noted that the performance level on objective 1 has not changed. One might postulate that the topic is just too difficult for this particular group of students.

In Summary, CAM represents an application of criterion-referenced testing to program evaluation carried out using longitudinal testing and the notion of item-examinee sampling.

Summary

In this paper we have attempted to highlight some of the special characteristics of criterion-referenced tests and compare them with norm-referenced tests. Psychometric considerations involved in constructing a criterion-referenced test including item analysis, reliability, and validity were mentioned. Also the application of criterion-referenced testing to individual assessment and program evaluation was described.

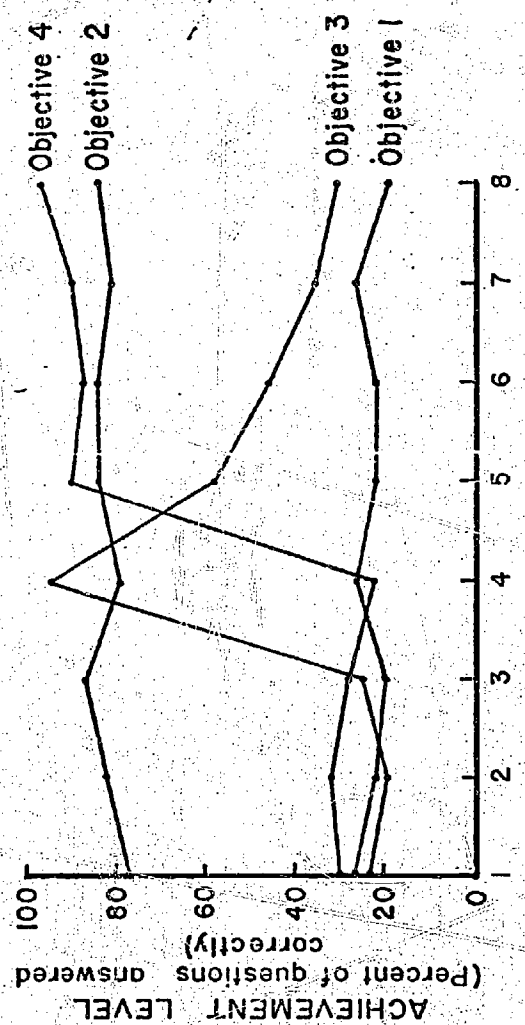

Figure 1. Achievement profiles of a group of students on four objectives across eight test administrations.

Throughout the paper an attempt has been made to indicate some problems and shortcomings of the current testing methodology. Hopefully the discussion of these problems will stimulate others to develop the methodology and models appropriate for criterion-referenced testing since these problems must rank among the most pressing in educational measurement.

References

Allen, D.W. and Gorth, W.P. Fourth Annual Report to the Charles F. Kettering Foundation. Amherst, Mass.: School of Education, University of Massachusetts, 1971.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R., Statistical theories of mental test scores, Reading, Mass.: Addison-Wesley, 1968.

Block, J.H. (Ed.) Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometika, 1971, in press.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Carroll, J.B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Carroll, J.B. Problems of measurement related to the concept of learning for mastery. Education Horizons, 1970, 48, 71-80.

Carver, R.P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Washington: American Institute for Research, 1970.

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 measurement. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970. (ERIC, Ed 038 679)

Cox, R.C. and Graham, G.T. The development of a sequentially scaled achievement test. Journal of Educational Measurement, 1966, 3, 147-150.

Cox, R.C. and Vargas, J.C. 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, 1966.

Cronbach, L.J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Cronbach, L.J. Validation of educational measures. Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1970.

Cronbach, L.J. and Gleser, Goldine C. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1965.

Englehart, M.D. A comparison of several item discrimination indices. Journal of Educational Measurement, 1965, 2, 69-76.

Ferguson, R.L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburg, 1969.

Flanagan, J.C. Functional education for the seventies. Phi Delta Kappan, 1967, 49, 27-32.

Flanagan, J.C. Program for learning in accordance with needs. Psychology in the Schools, 1969, 6, 133-136.

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

Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968, pp. 3-36.

Glaser, R. and Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement. Washington: American Council on Education, 1971, pp. 625-670.

Gorth, W.P., Schriber, P. and O'Reilly, R.P. Comprehensive Achievement Monitoring. Amherst, Mass.: School of Education, University of Massachusetts, 1971.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L.A. The basis for scalogram analysis. In S.A. Stouffer, et al. Measurement and prediction. Princeton, N.J.: Princeton University Press, 1950. pp. 60-90.

Hambleton, R.K., Gorth, W.P., and O'Reilly, R.P. A formative evaluation model for classroom instruction. Technical report #16, School of Education, University of Massachusetts, Amherst, 1971.

Hambleton, R.K. and Novick, M.R. Towards a theory of criterion-referenced tests. American College Testing Technical Report, Iowa City, 1971, in press.

Hambleton, R.K., Roberts, D.M., and Traub, R.E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 1970, 7, 75-82.

Hambleton, R.K., Rovinelli, R., and Gorth, W.P. Efficiency of various item-examinee sampling designs for estimating test parameters. In Proceedings, 79th Annual Convention, American Psychological Association, 1971, in press.

Hambleton, R.K., and Traub, R.E. Analysis of empirical data using two logistic test models. Technical Report #3, School of Education, University of Massachusetts, Amherst, 1970. (ERIC, Ed 042 816)

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

Ivens, S.H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Kriewell, T.E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Lindeman, R.H., Gorth, W.P., and Allen, D.W. The evaluation of item performance in an item sampling case. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1969.

Livingston, S.A. The reliability of criterion-referenced measures. Center for Social Organization of Schools Technical Report #73, The John Hopkins University, 1970.

Lord, F.M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F.M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F.M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M. Some test theory for tailored testing. In W. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Millman, J. Reporting student progress: a case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Nitko, A.J. A model for criterion-referenced tests based on use. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

O'Reilly, R.B. and Hambleton, R.K. A CMI model for an individualized learning program in ninth grade science. Technical Report #14, School of Education, University of Massachusetts, Amherst, 1971.

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

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Shoemaker, D.M. Allocation of items and examinees in estimating a norm distribution by item sampling. Journal of Educational Measurement, 1970, 7, 123-128. (a)

Shoemaker, D.M. Item - examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970, 7, 255-262. (b)

Wang, Marilyn D. and Stanley, J.C. Differential weighting: a review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

Wardrop, J.L. The use of item characteristic curves for criterion-referenced measurement. Urbana: University of Illinois, 1970. (mimeo)

Wright, B. Sample-free test calibration and personal measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.