DOCUMENT RESUME

ED 126 155 TM 005 412

AUTHOR Haladyna, Tom

TITLE The Paradox of Criterion-Referenced Measurement.

PUB DATE [Apr 76]

NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS Academic Standards; Achievement Tests; *Comparative Analysis; *Criterion Referenced Tests; Decision Making; Item Analysis; Item Sampling; *Norm Referenced Tests; Standard Error of Measurement; Test Construction; Test Interpretation; Test Reliability; Test Validity

IDENTIFIERS *Domain Referenced Tests; Test Theory; Variance (Statistical

ABSTRACT

The existence of criterion-referenced (CR) measurement is questioned in this paper. Despite beliefs that differences exist between two alternative forms of measurement, CR and Norm Referenced (NR), an analysis of philosophical and psychological descriptions of measurement, as well as a growing number of empirical studies, reveal that the common distinctions drawn between CR and NR measurement focus on what occurs prior to and following measurement, namely the writing of items and the interpreting of test scores. In this respect, the use of the term "criterion-referenced measurement" is paradoxical. (Author/DEP)

The Paradox of Criterion-Referenced Measurement

Tom Haladyna

Teaching Research Division

Oregon State System of Higher Education

A paper presented at the annual meeting of the National Council
on Measurement in Education, San Francisico, 1976

Traditionally, there has been one measurement construct, most commonly referred to as "norm-referenced". The need for more effective measures of classroom achievement coupled with the interest in using instructional objectives in teaching and testing has motivated the establishment of a new construct known as "criterion-referenced measurement". Norm-referenced (NR) measures are believed to be compatible with the study of individual differences, while criterion-referenced measures are more suitable for ascertaining any examinee's level of performance with respect to a well-defined achievement domain. The classical theory of measurement has been found to be unsuitable for application to CR measures (Popham and Husek, 1969), and a CR theory of measurement has been sought. Paradoxically, an analysis of the distinctions commonly drawn between CR and NR measurement, coupled with accumulating test data, suggests that there is only one measurement construct with two functions, NR and CR. This paper is devoted to an examination of this paradox.

Definitions

An inevitable problem in any analysis of CR and NR measurement is how these terms are defined and used in everyday test practices. The elusiveness of many definitions of CR measurement as well as the wide range of their applications had been reviewed by Hambleton and Novick (1973). The CR test, according to Popham and Husek (1969, p. 2) is "...one which is used to ascertain an individual's status with respect to some criterion." Explicit in their conceptualization of a CR measure is that these tests are objective-based and that a criterion level is employed to assign examinees to passing or failing categories. CR tests are believed to measure achievement of a well-specified domain (the domain consists of a well-defined set of tasks). There is no concern about how any examinee scores with respect to other examinees, as attention in testing is focused on how that examinee stands with respect to the entire domain of items. The variance of CR test scores is said to be restricted to the degree that traditional item and test statistics are useless.

In contrast, the NR test is primarily designed to measure the relative standing of examinees. It is intentionally constructed to maximize differences among examinees, consonant with the effort to measure individual differences. Traditional methods for estimating item discrimination and test reliability depend upon sufficient variance of the scores of examinees. When scores are restricted, these estimates are attenuated.

A recent outgrowth of the CR test movement has been the domain-referenced test (DRT). Any DRT is simply a random sample of items from a well-defined domain of items. This deceptively simple definition, however, does not capture the essence of a DRT. Millman (1974) states that the DRT is conceptually quite distinctive from CR and NR tests, both of which are considered differential assessment devices as contrasted with the DRT which is a true CR measure. The DRT is primarily distinguished from the more traditional CR test in terms of how items are created and how tests are constructed. That is, item-generating algorithms are used to write test items, and items are randomly sampled to test forms. Following these procedures will result in measures which have no reference to the sample of examinees, but yield clear-cut measures of achievement within that domain for each examinee.

In the balance of this analysis, both CR and DR tests will be treated as different cases of CR measurement. Distinctions drawn between NR tests and these two cases have been made in the areas of (a) what is measured, (b) how domains are measure, (c) test standards, (d) item selection, (e) reliability, measurement error, and decisionmaking, and (f) validity.

What Is Measured

The core of the difference between CR and NR measures is said to be what is intrinsically measured. In the context of classroom achievement testing, the objective of a NR measure is to ascertain the rank and relative differences for a group of examinees with respect to an achievement domain; the objective of a CR test is to estimate the level of functioning of any examinee so that level can be compared to the level of acceptable performance. Carver (1974) describes this difference as "the measurement of individual differences versus the measurement of the amount learned" (p. 512). This distinction has been made in a number of other discussions of CR tests (e.g, Popham and Husek, 1969; Anderson, 1972; Hambleton and Novick, 1973). In fact, this distinction pervades almost the entire body of literature on the subject!

Measurement has traditionally meant the obtaining of a numerical description of an object on a trait through the use of some rules. In achievement testing, this has come to mean the sum of correct responses on a test for a student. The object is the student, the trait is achievement and the rules involve taking the sum of correct responses. One kind of test (NR) gives relative information, and the other kind (CR) gives absolute information. Measurement has been viewed in other contexts as having two functions: (a) knowing the quantity thereof, and (b) making fine and subtle discriminations (Kaplan, 1963), and similar observations have been made by Cronbach (1970), Ebel (1975), and Messick (1975). Thus on the surface, the core of the difference between CR and NR measures has been widely discussed as one of what is inherently measured. In actuality, it appears that any achievement measure can yield one of two interpretations depending upon the function we wish to employ for the purposes at hand. This distinction is illustrated by noting the difference between a percent and a percentile (Glaser and Klaus, 1962). Using the above illustration, it should be apparent that any test can yield percentile or percentage interpretations, and that each provides unique information despite the single measurement that occurs.

How Achievement Domains Are Measured

The way items are created is another way to distinguish the CR, NR, and DR tests from each other. In a purely CR approach, items are constructed to directly represent instructional objectives. The information obtained from a CR test constructed in this manner permits inferences to be drawn about to what degree an objective or set of objectives have been achieved. This is contrasted with a traditional approach where (a) the achievement domain is abstractly defined, (b) items are written to represent the construct and (c) test results are used to confirm our predictions about the construct. These procedures may be recognizable as those recommended for the establishment of the construct validity of tests. Nonetheless, the procedures describe what goes on ideally in the creation of achievement tests using the classical theory. And it should be clear that test-making is often reduced to one of introspecti.e and subjective item writing to represent some vaguely-conceived domain. In the DR approach, items are created through the employment of an item-writing algorithm such as the ones suggested by Bormuth (1970) or Hively

(1974). Data has not been collected and reported as yet that attests to the distinctiveness of any of these item-generating approaches as leading to unique measures of the same achievement domain in question.

It has been noted that the DR test is formed by random sampling of items from an item pool to test forms. Despite the recency of DR testing, it needs to be noted that the practice of random sampling is neither unfamiliar nor antithetical to traditional test theory. The term "domain sampling" was used by Nunnally (1967) to describe the classical theory of measurement. In the truest sense, the classical theory has involved the random sampling of items from a well-defined set (Lord and Novick, 1968, p. 29) and the need for random sampling is explicit in the theory of generalizability, as well (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

The essence of the difference between DR and NR testing lies in the interpretation of the achievement domain. In the DR approach, the domain is operationally defined, while in traditional testing, the objective is one of inferring an abstract construct. While the use of item-generating techniques and random sampling may not distinguish a DR test from others, the interpretation of test results can be clearly DR or construct-referenced. In a DR approach, any test score represents an unbiased estimate of performance of all items in that domain. Thus, the interpretation is based upon our operational definition of that domain. In a traditional approach, one infers a construct which is more or less intangible. The acceptance of DR testing hinges on our willingness to accept an operational definition as a best indicator of a domain. For example, a domain of word problems in mathematics may be created through the use of an item writing algorithm.

Is it sufficient to know how many of all problems any student can correctly solve, or do we wish to know about something more intangible and seemingly remote, such as arithmetic reasoning? The issue presented here is more in the realm of values and meaning, and interestingly enough, this issue involves something that occurs after the process of measurement.

One study was recently completed which bears importantly on the issues of how domains are measures. Roid and Haladyna (Note 1) contrasted two types of item-writing procedures, one purely DR, the other CR. Bormuth's item-writing rules were used to construct a subtest over a 32 page, learner-verified programmed text, while another subtest was prepared using instructional objectives for the same material. The tests were administered prior to and following instruction to a group of students. One item writer consistently produced items of greater difficulty than the other item writer, and both item writers produced roughly the same number of faulty items regardless of the item-writing approach. In fact, both DR and CR item-writing approaches unexpectedly produced the same large number of faulty items that one could expect from using the traditional, subjective approach to item writing. A subsequent experiment is currently in progress where a NR approach is being compared to the CR and DR approaches. In light of the present stage of development of item generation theories and the rather negative supporting evidence, it appears premature to conclude that CR and DR tests offer distinctive and superior measures of achievement.

Standards

In DR and CR tests, a standard is typically used to assign students to a passing or failing category. Establishing a point on any test scale is an action which is independent of the measurement process, and something that can occur for any test. Doing so does not make a test CR or DR. It is merely using the test score to determine the worth of a student's learning effort or the value of the instructional program. In fact, it should be considered a CR use of a test score.

There is a more subtle and important problem with standards and the questionable existence of CR tests. Returning to the study by Roid and Haladyna (Note 1) where two types of CR test item-generating techniques were contrasted, clear-cut differences in the item difficulties of items written by the two item writers led to the creation of two scales, one hard and one easy, which were both reliable measures of the same achievement domain. Administering these forms to a group of students following instruction would create some serious problems in assigning students to pass or fail categories. Students receiving the hard test would more often be falsely categorized as passing. The use of either item-writing algorithm failed to reduce this difference between the item writers. The implications for this state of affairs is perplexing in light of the fact that CR measures are reputed to produce unbiased measures of student achievement. If either type of CR test is to be distinctive from typical NR test, the quality of items should be uniformly high and the difficulties be substantially reduced between various item writers. This has not occurred, and a standards problem continues to exist.

Test Variance

There has been considerable debate over the role and extent of test score variance in CR measures (Woodson, 1974a, 1974b; Haladyna 1974; Millman and Popham, 1974). Popham and Husek (1969) maintained that the meaning of scores flows from the item-objective congruence and not the notion of individual differences of which variability is a related concept. The variation of CR test scores is said to be restricted when learning and instruction is effective, and this seriously impedes the use of tradition item and test statistics.

Woodson (1974a, 1974b) has supported the idea that variability of test scores is a function of the sample of examinees. This has been clearly demonstrated in at least one study (Haladyna, 1974) and is logically realized when one considers the situation where a group of students have not learned content from a domain and another group has learned the content quite well. Resulting achievement test scores should be bimodal with the concentration of scores at the top and bottom of the achievement scale. The variability of these test scores is quite large. When high group averages 100 percent and the low group averages 0 percent, it can be deduced that the variance of test scores for the two group (when equally matched for sample size) is as high as possible. Therefore, any CR test could have substantial variance.

The usefulness of variability is questioned in CR tests by Millman and Popham (1974) in a rebuttal to Woodson's initial comments. The argument follows from the basic distinction made earlier, that CR tests are concerned with how a student has achieved rather than how different students are. The primary speculation about test variance appear to be what the role of variance in such tests should be. Perhaps the point of contention with variability involves the notion that a CR test is appropriate for determining how much a student has learned while a NR test is more appropriate for measuring individual differences. The measurement of individual differences, as earlier noted, is a function of all interval or ratio measurement scales. It is clear that CR and DR tests have substantial variability when one samples high and low achievers. The very same is true for a NR test. It may be more appropriate to say that any test is open to interpretations about individual differences if that test is sensitive to the trait being measured. If the test isn't sensitive, it probably contains too much measurement error to be useful for anything.

Given that variance can be substantial in any classroom achievement test when the sample obtained spans the full range of achievement, how do traditional item and test statistics work with these CR and DR tests? It is with item and test statistics that CR and NR tests should distinguish themselves.

Item Selection

There is a difference between the role of item analysis in CR and DR testing. Traditionally, CR item analysis has been one of ascertaining item quality in light of instructional sensitivity. The most commonly used CR item discrimination index is one derived by taking the difference between item difficulties of the item administered to pre- and post-instruction examiness (Cox and Vargas, Note 2). A number of empirical studies have been done comparing various indexes with the Cox-Vargas coefficient e.g., Rahmlow, Matthews, and Jung (Note 3); Popham, 1972; Tsu (Note 4), Haladyna, 1974; Helmstadter (Note 5); and Crehan, 1975). The scope and limitations of most of these studies were recently discussed in a study by Haladyna and Roid (Note 6).

In DR testing, empirical item analysis should not be used due to the possibility that the interpretation of DR test scores is destroyed when items generated for the DR test pool are tampered. Millman (1974 , p. 339) states:

> The use of item statistics destroys the random selection process, a defining characteristic of DRT's. Unless items are selected randomly, the estimate of a person's level of functioning loses meaning and interpretability of the test score is reduced.

A further criticism for the use of instructional sensitivity was offered by Cronbach (1975) who stated that useful items would not be sensitive to instruction and thus falsely discarded. Such items might be transfer items which should display a high difficulty index both preceding and following instruction.

There are a host of compelling reasons for the use of item analysis, and these reasons need to be established before the evidence for the distinctiveness of various CR item selection procedures can be discussed. First, the item writing procedures which distinguish DR tests from all others have not been shown to produce uniformly high quality items. The study by Roid and Haladyna (Note 1) indicated that a DR procedure produced as many faulty items as a CR procedure. In fact, the number of faulty items produced was comparable to the number of faulty items one would expect from using the traditionally subjective approach. Second, how doing an empirical item analysis destroys random selection is unclear. It seems reasonable to employ an empirical screening method to weed out faulty items as they surely seem to exist in any item pool. Random sampling is done from a item pool which has been gleaned of faulty items. Third, how empirical item analysis destroys interpretability is also unclear. Defective items are one source of measurement error. To rid item pools of measurement error can only improve the precision of test scores as well as interpretability. Finally, alternative and non-empirical procedures for item analysis have been advocated by CR advocates (e.g., Hambleton, et al., Note 7). These procedures involve committees of content experts who judge items for their appropriateness. These non-empirical procedures are aligned with the logical analysis that precedes item pools and tests. There is no available evidence of the soundness of these non-empirical procedures on DR or CR items, and there is reason to believe that these procedures are nothing more than traditional approaches to establishing content validity.

Perhaps a more compelling reason for empirical-item analysis is a consideration of the inferences one draws from test data and its basic unit of measure, the item. This reason is also at the core of the supposed difference between CR and DR measures. Thorndike (in Jackson & Messick, 1967, p. 203) stated that "Each item is in a very real sense a little test all by itself. Each item must necessarily be judged on its own merits as far as validity is concerned." Traditionally, item validity has been the correlation between item and test performance for a group of examinees, the item discrimination index. If a CR test, as Popham and Husek (1969) describe it, is sensitive to treatments (i.e., instruction), we expect pre-instructed students to score low and post-instructed students to score high. Messick (1975, p. 959) has conveyed a similar impression when he states:

> ...the most sensitive and soundest evidence is likely to come from experimental studies of groups receiving different instructional treatments or of test administrations under different conditions of motivation and strategy.

The instructional sensitivity might apply not only to test scores, but item difficulties as well. Since the item is the analogue of the test. Empirical item analysis appear to be a necessary feature of any achievement test regardless of its purpose, CR or DR; and the Cox-Vargas index comes conceptually closest to measuring CR item discrimination. Is this index actually distinctive from others?

The discussion up to this point in this section has been directed at the necessity of item analysis in DR testing. In CR testing, it has been advocated for some time (Kosecoff and Klein, Note 8; Harris, Note 9). However, the traditional point biseral correlation which serves as an item discrimination index has been criticized due to the variance problem. Whenever a correlation is computed from a restricted range of scores, that estimate of relationship is attenuated. Haladyna (1974) used samples of pre- and post-instructed students to compute traditional item statistics, which compared very favorably to CR item statistics. Haladyna and Roid (Note 6) examined a host of CR and other item statistics with a CR test and discovered that all were uniformly highly related. These included a Baysian index (Helmstadter, Note 5), a Rasch instructional sensitivity index, the Cox-Vargas coefficient, and the full-scale traditional item discrimination coefficient (Haladyna, 1974). The intercorrelations among these statistics approached unity, 1.00. Thus, it would seem that all four statistics provide identical information.

The contention that CR and DR tests are unique is not supported when examining the rational for item analysis and the empirical evidence for item discrimination indexes. Perhaps the reason for this lack of support can be found by closely examining a distinction drawn by Millman (1974) when he labeled typical CR and NR tests as "differential assessment devices" (DAD). In contrast, the DRT is not a DAD. Whenever group or individual differences are considered, the concepts of variability and traditional item and test statistics are quite acceptable. And it follows the CR as well as the NR test are DAD's. However, it should be clear that these concepts apply to the DR test as well; and when traditional statistics are employed, the DRT resembles all others. Thus the CR and DR tests, as defined in the paper, produce test results that lend themselves to conventional item analysis. In the area of item selection, the distinctions drawn among CR, DR, and NR tests don't withstand empirical tests.

Measurement Error, Reliability, and Decisionmaking

With any test, a certain degree of measurement error will occur. Typically the construct of reliability assists us in understanding how much measurement error can be found in the test responses of a group of examinees. The problem in instruction is knowing how much error exists when deciding who has passed and who has failed. Invar'ably, a number of examinees scores will fall near the passing standard, and the risk of misclassifying these persons is high. There have been several suggestions to establish confidence bands around the passing standard and assign passing, failing, or conditional status to examinees based on this confidence band (Hambleton and Novick, 1973; Millman, 1974). The procedures minimize the misclassification errors tha? often occur.

What is required is a statistical procedure which permits the valid estimation of a standard error from which the confidence band can be constructed. The believed low variability of CR and DR test scores has led some persons to reject classical reliability estimates (e.g., Popham and Husek, 1969). However, reliability was shown to be reasonably estimated from combined samples of pre- and post-instruction examinees in one study by Haladyna (Note, 10). Although reliability estimates are dependent upon variance, the estimation of measurement error in traditional test theory is not a function of the variability of test scores. Therefore, the traditional standard error of measurement can be usefully estimated for any test including putative DR and CR tests.

Other procedures have been recommended as alternatives. One of these is a straightforward item sampling approach where the binomial distribution is employed (Millman, 1974b). Traditional and item sampling approaches were compared in one study (Haladyna, Note 10) with a slight superiority for the traditional approach. However, both approaches were found to be lacking in terms of feasibility with student populations. Baysian techniques and a traditional approach were compared in a Monte Carlo study by Hambleton, Hutten, and Swaminathan (unpublished). While one Baysian technique showed a superiority to others in the accuracy of decisionmaking, the differences were slight.

One example of the usefulness of traditonal reliability estimates for CR tests can be found in the statewide assessment of fourth grade mathematics in the state of Oregon (Haladyna, 1976). Following procedures similar to those recommended by Millman (1974), a content panel, represented by mathematics educators, was established to judge the congruence of items with 29 instructional objectives. All data was subjected to traditional item and test analysis. Scales for objectives were classified into five achievement domains and were quite reliable as judged by traditional KR-20 estimates. Variance was not restricted, and test results appeared quite similar to those one might obtain in any NR mathematics achievement test of fourth graders. Interestingly, this test is CR by virtue of the way it was constructed, and DR in the loosest sense, and yet, traditional statistics were usefully employed to gain an understanding about how much measurement error occurred with their fourth grade sample.

There are three salient observation which follow from this discussion of reliability, measurement error, and decisionmaking:

1. Traditional reliability estimates have been usefully employed in CR and DR tests to estimate standard errors. If one can judge by these limited number of studies, the traditional approach is slightly superior to the item sampling approach and slightly inferior to the Baysian approach. What is conclusive from these empirical findings is that all three approaches lead to measures of the same construct, test error. And all approaches (e.g., Baysian, Rasch, traditional, and item sampling) are based on the concept of true scores.

2. Any test may be used to make decisions and can, therefore, be CR by simply establishing a decision point and using it accordingly. Therefore, use of passing standards does not distinguish a CR or DR test from a NR test.

3. Regardless of the ways tests are categorized (i.e., CR, NR, or DR), it seems clear that reliability estimates and standard errors are very comparable regardless of the procedure used to obtain these estimates. If the CR, DR, and NR tests are truly distinct measurement constructs, NR

approaches to measurement error, reliability, and decisionmaking should be ineffective. But clearly they are as effective as any other approach including those specifically created for these CR and DR tests.

<u>Validity</u>

A CR test is constructed to reveal an examinee's relationship to a behavioral repertory (Glaser, 1963) or to measure an examinee's standing with respect to a criterion level (Popham and Husek, 1969). If DR, CR, and NR tests are indeed different, then how might they differ in terms of validity? Are conventional concepts of validity applicable to DR and CR tests?

One of the unique features of DR tests is the strict adherence to the item-writing algorithm and the random sampling of items to test forms. While these item-writing procedures have reached an operational level, they are by no means unfamiliar. As noted earlier, the need to clearly define domains, to construct items representing the domain,aand to randomly sample test forms have been hallmarks of classical test theory. The random sampling of items to test forms is actually one very desirable form of content validity called "sampling validity" by Helmstadter, (1964). What is unfortunate is the lack of attention given to the principles espoused in classical test theory. Seldom has achievement tests in the past been carefully constructed to represent domains and randomly assigned to test forms. Nunnally (1967) admits to the fact that traditional theory is not true to life. (Or perhaps test practitioners are not true to theory).

A DR approach to item analysis has been the use of content panels to judge the item-domain correspondence (Millman, 1974; Hambleton, <u>et al</u>., (Note 6).

The use of such content experts is actually a type of face validity, the weakest and least justifiable form of content validity according to Helmstadter (1964).

The role of variance in CR and DR tests had led to the conclusion that conventional correlation-based statistics are typically useless. As a consequence, the role of predictive validity was said to be quite limited. If one considers the potential of using pre- and post-instructed students in studies of predictive validity, there is no restriction of test scores and predictive validity may be usefully employed. For example, successful students should have a high probability of success in future units of instruction or on a task from which instruction was designed and unsuccessful students should not have as good a chance.

There is a more compelling reason to study predictive validity in the context of systematic instruction. Where the passing standard is set determines who will pass and who will fail. Instruction is planned to establish high degrees of achievement in students for the purposes of either continuing in a sequence of instructional units or giving evidence of competence so that examinees may perform a task or series of tasks. Setting high standards will ensure higher levels of achievement with the risk of obtaining greater failures and more frustration on the part of students. Where the ideal criterion level is and how to maximize the success of students in future endeavors are problems of predictive validity. However, the establishment of a criterion level is not a distinguishing trait of a CR or DR test. In the truest sense of the word, it is a CR use of a test.

The need for more concern for the construct validity of educational achievement tests was expressed by Messick (1975, p. 957):

> ...all measurement should be construct referenced. A measure estimates how much of something an individual displays or possesses. The basic question is, What is the nature of that something? It may be answered by referring to evidence insupport or particular attributes, processes or traits constructed to underlie and determine task performance.

While it is an eloquent plea for greater concern for the inferred construct behind each achievement measure, the opposing approach, epitomized in DR testing, is equally compelling. The essence of the disagreement is based in our willingness to accept interpretations of achievement in strict behavioral language. Or, as Cronbach and Meehl (1955) contend, when our operational definitions conflict, one is compelled to become concerned with the construct validity of our test interpretations. Borrowing an example from Millman (1974 , p. 321) a DRT can be constructed to ascertain if a student can solve profit and loss word problems. While a domain can be defined algorithmically, is this sufficient to define the domain of mathematics achievement of which we are interested. The differences here are in the realm of a philosophy of scientific inquiry and well beyond what is intended here. Regardless of one's stance on interpretation, it is clear that traditional concepts of validity work well in the context of systematic instruction where achievement tests are geared to the learner.

Conclusion

Despite the many efforts to construct a theory of CR measurement, there has been understandably little progress. Perhaps the *raison d'etre* is that there are really not two or three different measurement constructs, but only one. That one construct has two primary functions: (a) the first is knowing how much of that trait an examinee possess--CR; and (b) the second is knowing how different one examinee is from another--NR.

The contentions that CR and NR tests are distinguished by the way items are constructed (i.e., item-writing algorithms) has not empirically been supported. In fact, DR tests look and behave like any other test of the same domain when administered to a equivalent or same group of examiners.

The belief that variance is greatly reduced in CR ests when compared to NR tests is also quite unsupported. What does occur in effective instruction happens to any test which is geared to the content begin taught. Thus, the restriction in range of achievement test scores of the post-instruction students is a function of the instruction and not the test. And, it has been demonstrated that variance is actually maximized in situations where the tests are directly geared to the instruction that occurs.

The role of variance and the variance-reliant statistics has also been questioned by many CR advocates. With variance not restricted as originally believed, traditional reliability and item discrimination indexes can be usefully estimated. When they are computed, they are found to be quite comparable to statistics uniquely compatible to CR and DR tests, thus giving credibility to the argument that a host of reliability and item discrimination procedures lead to measures of the same constructs, measurement error, and item quality.

In effect, any achievement measure is simply that. It is neither NR, CR, nor NR. The advent of the CR test and later the DR test, may be an reaction to what traditonal test theory has evolved, a degenerate use. That is, classroom teachers are unable to cope with the intricacies of test theory and the demands to construct and analyze classroom achievement tests in the recommended way. This has lead to testing practices which are actually reproachable and have come to be labeled "NR". It is undeniably clear that classroom achievements tests have in the past and will in the future be misused. The movement toward CR and DR testing has created an interest in unifying instruction and testing. For the most part, this creates testing which is well suited to the needs of effective evaluation of instruction and student progress. It does not constitute a new form of measurement, as the arguments presented here and accumulating test data has and will continue to attest.

REFERENCE NOTES

1. Roid, G. H., & Haladyna, T. M. *A comparison of objective-based and Bormuth item writing techniques.* A paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

2. Cox, R. C., & Vargas, J. *A comparison of item selection techniques for norm-referenced and criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, 1966.

3. Rahmlow, H. F., Matthews, J. J., & Jung, S. M. *An empirical investigation of item analysis in criterion-referenced tests.* A paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

4. Hsu, T. *Empirical data on criterion-referenced tests.* A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

5. Helmstadter, G. C. *A comparison of Baysian and traditional indexes of test item effectiveness.* A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

6. Haladyna, T. M., & Roid, G. *The quality of domain-referenced test items.* A paper presented at the annual meeting of the American Educational Research Association Meeting, San Francisco, 1976.

7. Hambleton, R. K., Swaminathan, H., Alsing, J., & Coulson, D. *Criterion-referenced testing and measurement: A review of technical issues and developments.* A symposium presented at the annual meeting of the American Research Association, Washington, D.C., 1975.

8. Kosecoff, J. B., & Klein, S. P. *Instructional sensitivity statistics appropriate for objective-based test items.* A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

9. Harris, C. W. *Techniques for analyzing test response data.* A paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

10. Haladyna, T. M. *An investigation of full and subscale reliabilities of criterion-referenced tests.* A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

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

Carver, R. P. Two dimensions of tests--Psychometric and edumetric. *American Psychologist*, 1974, *29*, 512-518.

Cronbach, L. J. Review of "On the theory of achievement test items", *Psychometrika*, 1970, *35*, 509-511.

Cronbach, L. J. Dissent from Carver, *American Psychologist*, 1975, *30*, 602-603.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 281-302.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements.* New York: John Wiley, 1972.

Ebel, R. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*, 273-279.

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

Glaser, R., & Klaus, D. Proficiency measurement: Assessing human performance, pp. 419-474. In R. M. Gagne, (Ed). *Psychological principles in system development*, New York: Holt, Rinehart & Winston, 1962.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 93-99.

Haladyna, T. M. *Technical report of the pilot assessment of mathematics in grade 4, Oregon.* Monmouth, OR, Teaching Research, 1976.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. *A comparisón of several methods for assessing student mastery in objective-based instructional programs.* Unpublished.

Helmstadter, G. C. *Principles of psychological measurement.* New York: Appleton-Century-Crafts, 1964.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*, 5-10.

Kaplan, A. *The conduct of inquiry.* San Francisco: Chandler, 1964.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham, (Ed). *Evaluation in education: Current applications.* San Francisco: McCutchan, 1974.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 1975, *30,* 955-966.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement,* 1974, *11,* 137-138.

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Popham, W. J. Indices of adequacy for criterion-referenced tests. In W. J. Popham, (Ed). *Criterion-referenced measurement.* Englewood Clifts, NY: Educational Technology Publications, 1972.

Popham, W. J. Selecting objectives and generating test items for objective-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham, (Eds). *Problems in criterion-referenced measurement.* CSE monograph series in evaluation, No. 3, Los Angeles: CSE, 1974.

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

Thorndike, R. M. The analysis and selection of test items. In D. N. Jackson & S. Messick, (Ed). *Problems in human assessment.* New York: McGraw-Hill, 1967.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. *Journal of Educational Measurement,* 1974, *11,* 63-64, a.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests: A reply. *Journal of Educational Measurement,* 1974, *11,* 139-140, b.