Comments on Criterion-Referenced Testing

Samuel A. Livingston
Educational Testing Service

The six papers in this issue summarize 10 years of theory development, empirical research, and practical experience in criterion-referenced testing. Much of the theory development has focused on questions and issues raised by Popham and Husek (1969), who pointed out that much of traditional psychometric theory did not work well when applied to criterion-referenced tests. The six papers, taken together, represent an attempt to answer four basic questions:

1. How should the reliability of a criterion-referenced test be measured?
2. How should it be decided how many items are needed in a criterion-referenced test?
3. How should criterion-referenced tests be used to make decisions about the people taking the tests?
4. What kind of evidence should be provided for the validity of a criterion-referenced test?

Attempts to answer these questions have been complicated by the lack of a universally accepted, unambiguous definition of the term "criterion-referenced test." Glaser's (1963) article, in which the term first appeared, defined criterion-referenced measures as those that "depend on an absolute standard of quality" (p. 519). However, Glaser went on to say that "the standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum" (p. 519) and that "we need to behaviorally specify minimum levels of performance. . ." (p. 520). These two ideas—absolute standards and behavioral test content specifications—received varying degrees of emphasis from the different individuals who attempted to develop criterion-referenced tests and to theorize about criterion-referenced testing. As a result, there are now several different answers to some of the questions that Popham and Husek (1969) raised.

State Models

Probably the most basic theoretical issue in criterion-referenced testing is the question of whether the skills measured by the test are effectively dichotomous or continuous. How this question is an-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 575–581

swered will, to a great extent, determine the approach to questions of reliability, test length, decision-making, and validity. Both the dichotomous approach and the continuous approach are useful, but in different testing situations. Which approach should be used depends on the skills to be measured and the population of persons to be tested. Some tests, particularly those used for diagnostic purposes, measure skills that are so highly specific that partial mastery is rare. For these skills, most of the persons to be tested at any given time will fit into one or the other of two categories: those who can use the skill correctly in most applications and those who can use it correctly in very few applications. For tests of these skills, state models are appropriate.[1]

In the past several years there has been considerable progress in the development of psychometric techniques based on state models. Much of this progress has been in the reduction of their dependence on assumptions not likely to be met in the real world. Macready and Dayton's paper in this issue describes the progress that has been made in this area, much of it by these two authors. For example, early papers on state models tended to include the assumption that all items on the test were equally difficult. Macready and Dayton have presented several models that do not depend on this assumption.

Standard Setting

Although state models are probably appropriate for some types of tests, they are clearly inappropriate for others. Once the assumption of an underlying dichotomy is abandoned, the problem of setting standards must be faced: deciding how much of the skill measured by the test is sufficient to place the examinee in the high-scoring group rather than the low-scoring group. Shepard has pointed out that "standards do not exist in nature, waiting to be estimated by statistical techniques." The important point of this statement is that the choice of a cutoff score on a test is not simply a statistical problem but also a psychological problem. Standards exist in people's minds but generally not in a form that can serve as the basis for an objective decision rule. For this purpose, the standard must be expressed in terms of some objective observable information such as a test score. Methods of choosing a cutoff score are actually methods of expressing the personal standards of one or more people in terms of the test score scale.

Some have used this fact to assert that the use of test scores to make decisions is basically an arbitrary process, no more objective than any other means of making decisions about people. Those who hold this view ignore the fact that there are varying degrees of subjectivity. Applying judgment by using a decision rule that is the same for all persons is more objective than applying judgment on a case-by-case basis. Standards cannot be objectively *determined*, but they can be objectively *applied*.

What, then, are the characteristics of a good standard-setting method? First, the judgments it is based on must be made by persons who are *qualified* to make them. Second, the judgments must be made in a way that is *meaningful* to the persons who are making them. Third, the process must take into account the *purpose* for which the test is being used. And fourth, the process must take into account the consequences of *both types of decision errors*.

There are three basic approaches to the task of translating implicit standards into a cutoff score on a test:

[1]Skills of this type may be more common in mathematics and related fields than in most other academic subject areas. One possible example of such a skill might be "finding the probability density of a function of two random variables with known densities."

1. The normative approach asks for judgments about groups of examinees, e.g., "What proportion of last year's students had adequate reading skills?"
2. The non-normative empirical approach asks for judgments about individual examinees. The "borderline group" method and the "contrasting groups" method are examples of this approach.
3. The conjectural approach asks for judgments about the performance expected of hypothetical "minimally competent" examinees. This approach includes the methods suggested by Nedelsky (1954), Angoff (1971, pp. 514–515), and Ebel (1972, pp. 492–495), which do not require actual test scores or responses of real examinees.

Shepard, taking an extreme position on the use of standards in program evaluation, states that "because standards impose an artificial dichotomy, they obscure performance information about individuals along the full performance continuum. Therefore, standards should not be used to interpret test data regarding the worth of educational programs." The first assertion is true, but the second does not follow. Any time data are summarized, information is lost. A statement such as "30% of the city's students scored below the level that a committee of teachers had selected as representing the minimum arithmetic ability acceptable for a high school graduate," is meaningful. It does not tell the whole story, but it tells an important part of the story. True, it focuses attention on achievement gains near the cutoff. In doing so, it may cause more educational resources to be directed at those students who have an acute need for them and the ability to benefit from them.

Decision Theory

One of the most promising recent developments in criterion-referenced testing is the increasing awareness of the relevance of decision theory. Van der Linden's paper in this issue is a thorough and precise mathematical presentation of decision theory as applied to educational testing. This very thoroughness and mathematical precision may tend to discourage test users without a strong background in mathematics. Fortunately, test users who want to use decision theory to set cutoff scores do not have to be able to follow a mathematical presentation of this type. Statistical decision theory, at its simpler levels, is really common sense expressed in mathematical language, and using contrasting-groups data to set a cutoff score is one of the simpler applications of decision theory. The data provide an estimate of the relationship between an examinee's test score and the probability that the examinee would be classified as a member of the higher group (e.g., mastery, success, adequate performance). There are two types of possible decision errors: passing a member of the nonmastery group and failing a member of the mastery group. Decision theory and common sense both tell us to minimize the total harm from all the decision errors. If passing a nonmaster is twice as bad as failing a master, two errors of the second kind (failing a master) can be tolerated for every error of the first kind (passing a nonmaster). Therefore, if a person's test score indicates that he or she is at least twice as likely to be a master as to be a nonmaster, that person should be passed; otherwise, not. In this case the cutoff would be the test score that corresponds to 2:1 odds, i.e., a .67 probability of mastery.

In using decision theory to choose a cutoff score, estimates are needed of the probability of mastery, given the person's test score. If these probabilities are to be estimated directly from the data, the sample of persons at each test score level must be representative of the persons in the entire population who have test scores at that level. This requirement will be met if the persons are selected at random from the population. It will also be met if they are selected on the basis of their test scores. It will *not* be met if they are selected on the basis of their mastery status (e.g., 100 masters and 100 nonmasters).

If there are test score data and mastery classifications from persons who have been selected on the basis of their mastery status, decision theory can still be applied if one additional number can be estimated: the proportion of the population who are masters. Bayes' theorem can then be applied:

$$P(M|X) = \frac{P(X|M)\,P(M)}{P(X|M)\,P(M) + P(X|N)\,P(N)} \quad [1]$$

where

P means "the probability that,"
M means "the person is a master,"
N means "the person is a nonmaster,"
X means "the person gets test score X," and
$|$ means "given that."

The process of collecting and using contrasting-groups data to choose a cutoff score on a test has been summarized as follows (Livingston, 1978):

1. Determine the measure of performance for which the standard is to be set. In general terms, we can tell this measure the *test score*. . . .
2. Determine the type of performance that will serve as the basis for judging a person's proficiency as adequate or inadequate. In general terms we would call this performance the *criterion performance*. . . .
3. Identify a population of persons qualified to judge examples of the criterion performance as adequate or inadequate. Select a sample of these persons to serve as *judges*.
4. Identify the population of persons taking the test for which a standard is to be set and obtain their test scores. Select a sample of these *examinees*, making sure the range of their test scores is broad enough to include both the lowest and the highest scores that might conceivably be selected as the standard.
5. Obtain *judgments* of the examinees' criterion performances by the judges.[2]
6. Analyze the data provided by these judgments to estimate the *probability* that an examinee's criterion performance will be judged adequate, as a function of the examinee's test score.

These six steps make up the empirical study. Two remaining steps complete the standard-setting procedure:

7. Determine the *relative seriousness* of the two types of possible errors: passing an examinee whose criterion performance is inadequate and failing an examinee whose criterion performance is adequate.
8. Set the *standard* at the test score level that results in an equal risk of the two types of possible errors, weighted by their seriousness in the particular decision-making situation for which a standard is to be set. (pp.269–270)

Incidentally, if there are very few nonmasters in the population, or if the cost of failing a master is high, the best decision rule may be to pass all persons regardless of their test scores. Koffler (1980) found such a result for one of the eight tests he investigated.

[2]In this step the judges should make their judgments *without* any information about the examinees' test scores.

Notice that it is *not* necessary to invoke the concept of "true score" when using decision theory with contrasting-groups data. The final decision rule and the probability estimates used to determine it will be in terms of observed scores.

Reliability

Probably the greatest challenge that criterion-referenced testing has posed to psychometric theory is in the area of reliability. As Popham and Husek (1969) pointed out, the general concept of reliability—the extent to which a person's test score is consistent over different occasions of testing, forms of a test, and so forth—is clearly relevant to criterion-referenced tests, but the classical definition of reliability as a correlation or a proportion of variance is not. The paper by Traub and Rowley in this issue presents an excellent review of the various approaches taken by different authors who attempted to resolve this dilemma. Traub and Rowley point out that the term "reliability" has been used to refer to several different characteristics of a set of test scores. They emphasize the important distinction between the reliability of *measurements* and the reliability of *decisions* based on those measurements. They also point out the distinction between two kinds of agreement: (1) between two parallel observed scores and (2) between an observed score and the corresponding true score. (They refer to the first type of agreement as "consistency" and the second type as "accuracy.") In traditional psychometric theory there is a simple mathematical relationship between the two types of agreement, but when agreement is considered as being more than a correlation coefficient, the relationship is not so simple.

One interesting fact that is apparent from Traub and Rowley's review is that the search for ways to describe the reliability of criterion-referenced tests led to the application of some earlier mathematical work on true-score estimation. This work, mostly by Lord (1965, 1969), had been available for several years, but its relevance to the questions raised by Popham and Husek (1969) had not been recognized.

Test Length

Closely related to the problem of describing the reliability of a test is the problem of assuring its reliability with respect to sampling of items, i.e., the problem of determining test length. Wilcox's paper in this issue deals directly with the question of how long a test should be. Wilcox presents formulas that allow the test user to specify four aspects of the situation:

1. Whether the main concern is with true scores on the test itself or with inferences to an infinite item domain;
2. The true score (or domain score) cutoff that defines mastery of the skills tested;
3. A desired minimum probability of correctly classifying a person as a master or nonmaster; and
4. How far the person's true score or domain score must be from the cutoff in order for the specified minimum probability of a correct classification to be required.

Wilcox's formulas translate these specified quantities into the minimum number of items required.

Validity

In the past few years the subject of test validity has received some serious attention in the form of articles questioning some of its basic concepts (e.g., Guion, 1977, 1978a, 1978b; Messick, 1975,

1979a, 1979b). Nevertheless, most testing professionals would probably agree with Cureton's (1951) definition of test validity as "how well a test does the job it is employed to do" (p. 621). Although they would agree on this general principle, they often disagree on its application to specific testing situations. The important, practical, and often controversial question of test validity is *what kinds of evidence* are necessary and sufficient to show that a test is doing its job adequately. The answer to this question undoubtedly varies from one testing situation to another. It may vary so much that any attempt to provide a general answer is doomed to failure. In some cases involving employment testing, the question has been answered by the federal courts, but not always satisfactorily from the point of view of those seeking guidance for future validation efforts.

Linn's article in this issue reviews the latest thinking on test validity as it applies to criterion-referenced tests and provides some practical guidance to those who must decide what kinds of validity evidence to provide. The article includes two examples of attempts to provide evidence of test validity, and the two attempts do indeed appear to be exemplary.

Probably the most important and controversial issue among experts writing about test validity is whether content validity evidence alone can ever be sufficient to justify the use of a test. Linn follows Messick (1975, 1979, 1980) in taking the negative position on this issue. The affirmative is represented by Ebel (1961, 1977) and by the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978). Ebel argues that in many cases the test serves to define the characteristic it is intended to measure. Therefore, if the test items form an adequate sample of the content to be tested, no additional evidence of the test's validity is necessary.

Future Developments

In the past 10 years a great deal of creative work in the theory and practice of criterion-referenced testing has been accomplished. What can be expected in the next 10 years, and what should we hope for? Probably the greatest need and the greatest opportunity for further theoretical work are in the area of test validity. What does it mean to say that a test is valid for a particular purpose? What kinds of evidence are needed to show that it is valid? Do criterion-referenced tests require special types of validity evidence? Notice that these questions involve both semantics and statistics. It seems likely that the answers will depend on several characteristics of the test and the testing situation. The answers will have to be expressed in practical terms to be of use to test makers and test users. Quite possibly, they will consist of a series of "if-then" statements: "If a test with characteristics x_1, x_2, x_3 is used in a situation with characteristics y_1, y_2, then the user should provide validity evidence of types z_1, z_2, z_3."

Another type of needed work is the translation of existing theoretical results into practical instructions that can be easily understood by professional test makers and test users who are not mathematicians. In some cases, this effort will require the creation of computer programs to perform the necessary calculations. Some work of this type has already been done in the areas of test reliability and decision theory, but much more is needed.

Probably the greatest need for empirical research is in the area of standard setting. Do the different standard-setting methods produce systematically different results? Are judges' conjectures about the performance of hypothetical minimally competent examinees consistent with the observed performance of examinees identified as minimally competent or with the outcome of contrasting-groups studies? Are judges' subjective estimates of the proportion of masters in a population consistent with their mastery judgments of individuals sampled from that population? Only a few such studies have appeared in educational measurement journals (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Koffler, 1980; Skakun & Kling, 1980).

None of these questions can be answered definitively with one or two research studies; each of them will require several studies with diverse tests and populations of examinees and judges. Clearly, considerable research remains to be done.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 1976, *36*, 45-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Brennan, R. L., & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 1980, *4*, 219-240.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1951.

Ebel, R. L. Must all tests be valid? *American Psychologist*, 1961, *16*, 640-647.

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Ebel, R. L. Comments on some problems of employment testing. *Personnel Psychology*, 1977, *30*, 55-63.

Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures. *Federal Register*, 1978, *43*, 38290-38309.

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

Guion, R. M. Content validity—the source of my discontent. *Applied Psychological Measurement*, 1977, *1*, 1-10.

Guion, R. M. "Content validity" in moderation. *Personnel Psychology*, 1978, *31*, 205-213. (a)

Guion, R. M. Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology*, 1978, *63*, 499-506. (b)

Koffler, S. L. A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement*, 1980, *17*, 167-178.

Livingston, S. A. Setting standards of speaking proficiency. In J. L. D. Clark (Ed.), *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: Educational Testing Service, 1978.

Lord, F. M. A strong true-score theory, with applications. *Psychometrika*, 1965, *30*, 239-270.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika*, 1969, *34*, 259-299.

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

Messick, S. Constructs and their vicissitudes in educational and psychological measurement (Research Report RR 79-11). Princeton, NJ: Educational Testing Service, 1979.

Messick, S. Test validity and the ethics of assessment. *American Psychologist*, 1980, *35*, 1012-1027.

Nedelsky, L. Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 1954, *14*, 3-19.

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

Skakun, E. N., & Kling, S. Comparability of methods for setting standards. *Journal of Educational Measurement*, 1980, *17* 229-235.

Author's Address

Samuel A. Livingston, Educational Testing Service, Princeton, NJ 08541.