Reliability of Test Scores and Decisions

Ross E. Traub
The Ontario Institute for Studies in Education

Glenn L. Rowley
La Trobe University

A criterion-referenced test can be viewed as testing either a continuous or a binary variable, and the scores on a test can be used as measurements of the variable or to make decisions (e.g., pass or fail). Recent work on the reliability of criterion-referenced tests has focused on the use of scores from tests of continuous variables for decision-making purposes. This work can be categorized according to type of loss function—threshold, linear, or quadratic. It is the loss function that is used either explicitly or implicitly to evaluate the goodness of the decisions that are made on the basis of the test scores. The literature in which a threshold loss function is employed can be further subdivided according to whether the goodness of decisions is assessed as the probability of making an erroneous decision or as a measure of the consistency of decisions over repeated testing occasions. This review points to the need for simple procedures by which to estimate the probability of decision errors.

It has been more than 70 years since Charles Spearman invented the concept of test reliability. In the time since then, practitioners of the discipline of educational and psychological measurement have acquired, perhaps by some Lamarckian process, a distinctive response tendency: Whenever a new approach to measurement is devised, the response is somehow to investigate its reliability. It is hardly surprising, therefore, that during the 1970s, when the notion of criterion-referenced measurement (CRM) captured and held the attention of the measurement profession unlike any other idea of the decade, the literature on the subject of CRM reliability expanded year by year in apparent conformity with an exponential function.

After reviewing this literature, however, the authors' assessment is that much of the literature on CRM reliability is confused and confusing in at least two interrelated ways: (1) the word reliability is used to denote several different concepts and (2) the technical requirements that should be met by the indices designed to describe the consistency of pass-fail decisions are not always clearly distinguished from the requirements met by traditional indices of reliability. This assessment is elaborated in the first of the two main sections of this paper, an elaboration that leads, ultimately, to the development of a framework for organizing the methods presently available for assessing CRM reliability. A review of these methods is presented in the second major section of this paper.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 517–545

Reliability Issues and Criterion-Referenced Tests

The term reliability carries with it a wealth of meanings and implications. In general, it describes not just one property of a person or object but a collection of attributes generally looked upon as desirable. Different attributes might be important in different circumstances. For example, the particular set of attributes that would cause a secretary to be described as reliable (speed, accuracy, punctuality) are not necessarily the same set of attributes as those for a jockey or a deep-sea diver. Although reliability is often thought to imply consistency or predictability, this is not always the case. The secretary who has a consistent error rate of 20% will probably be considered to be distinctively unreliable, even though predictable. Reliability does not necessarily imply a small variance, either. In cricket, an opening batsman who scores 50, 80, 45, 250, 165, and 93 runs in six successive matches will be regarded as exceptionally reliable, whereas the opening batsman who scores no runs at all in six matches will be seen as totally unreliable. A car that always starts at the first attempt is reliable; a car that never does is unreliable. In short, the term reliability has been used to describe not just one attribute, but a range of attributes, with the emphasis shifting with changing circumstances and requirements.

When scientists borrow a term from everyday language and use it as a label for a scientific concept, they must define it with care and precision. It takes on a much narrower meaning in science than it had in everyday life (consider, e.g., work, power, intensity); but if scientists are to communicate with one another, they must restrict themselves to these narrower definitions or else create confusion and misunderstanding. The advancement of science is dependent on the development of concepts that are useful, clearly defined, and well understood. The label carried by the concept is not as important as a clear and agreed-upon definition. The label, in fact, may be created for the concept (e.g., temperature) or may be borrowed from elsewhere and attached to it (e.g., energy). Good science has always involved the choosing of labels to apply to newly developed concepts, and the advancement of science has depended much more on the usefulness of the concepts than on the labels chosen to identify them.

In the science of measurement, certain terms have been borrowed from everyday language and applied to measurement concepts. Reliability is one of those words. Validity, for better or for worse, is another. The concept of reliability of measurement is among the better understood and more useful concepts to have emerged in the science of measurement. Defining reliability as the ratio of true-score variance to observed-score variance allows the prospect of shared meaning among measurement specialists. Although they might disagree as to how reliability should be assessed, the agreed definition allows them to see such disputes for what they are—differing views as to what should be regarded as true-score variance and what should be attributed to error. The work of Cronbach, Gleser, Nanda, and Rajaratnam (1972) has helped further to clarify such discussions; the substitution of universe score for true score allows the conception of different universes of generalization, for each of which the partitioning of total variance into universe-score variance and error variance is different.

The concept of reliability so developed refers to a property of a set of measurements. It has generally come to be regarded as describing an important property of those measurements—perhaps best described as the consistency with which the measurement process can distinguish among individuals with respect to the property or attribute being measured. Reliability is thus a property of measurements including both norm-referenced measurements as well as criterion-referenced measurements. If the persons or objects being measured do not differ, or differ very little, on the attribute being measured, reliability will be zero or very low. Similarly, if they do differ from one another but the measurement process is not precise enough to detect the differences, reliability will be low or zero.

Until recently, these features of the concept of reliability were understood but rarely thought important enough to warrant comment. There are few attributes upon which people do not differ substantially, so the problem of low or zero reliability arising from a low or zero variance among persons generally did not arise. The science of measurement was directed towards identifying and quantifying individual differences, and the concept of reliability was a useful one.

Then, in 1969, Popham and Husek published their influential article "Implications of Criterion-Referenced Measurement." In this article, considerable attention was paid to the question of variability, which "is at the core of the difference between norm-referenced and criterion-referenced tests" (p. 3). With norm-referenced tests, score variability is important; with criterion-referenced tests, it is irrelevant. Therefore, it was argued, traditional notions of reliability, dependent as they are on score variability, are unsuitable for use with criterion-referenced tests, as are the traditional ways of assessing validity and of performing item analyses. (See Linn, 1979, for a review of arguments on the importance of score variability in CRM.)

In their paper, Popham and Husek distinguished between criterion-referenced tests that yield essentially dichotomous scores, "that is, the individual has either mastered the criterion or he hasn't," and those for which "a range of acceptable performance exists" (p. 7). In the former case, the test can do nothing but sort the candidates into two categories; the latter case is what has traditionally been thought of as measurement. The characteristics that would be demanded of these two tests are, naturally, quite different.

The assumptions made about the nature of measurement are spelled out all too rarely in writings on CRM. A welcome exception was provided by Graham and Bergquist (1975), who presented two different criterion-referenced measurement models and explored the consequences of each. The models were termed binary and continuous, corresponding to the distinction made by Popham and Husek (1969) and referred to in the previous paragraph. (Meskauskas, 1976, refers to these as the state model and the continuum model.)

Under the binary model, the assumption is made that "certain capabilities enable an individual to perform an entire class of behaviors [universe of items], and if the capability is not acquired, the individual cannot perform any of the class of behaviors [items in the universe]" (p. 4). For example, a test of the ability to add pairs of single-digit numbers might well fit this assumption. As Graham and Bergquist (1975) pointed out, two implications of this model for an infinitely large population comprised of persons, some of whom possess and some of whom do not possess the capabilities in question, are these: (1) the universe of items is homogeneous in that all the items are of equal difficulty, and scores on all possible pairs of items are perfectly intercorrelated (except for the attenuating effects of measurement error); and (2) an examinee's true score on the universe, and hence on every sample of items from the universe, is either 0% or 100%. (If the multiple-choice item format is used, then examinees who do not possess the required capabilities can be expected to score at the level of chance.) Observed scores on a test composed of items from a universe conforming to the binary model can, of course, be expected to deviate upwards from 0 and downwards from n, the number of items in the test, to the extent that errors of measurement occur. Note that the true scores of this binary model are *not* the true scores of classical test theory; for if they were, the associated observed scores would of necessity be 0 or n. Instead, this is an example of Platonic true scores (Lord & Novick, 1968, sec. 2.9).

Under the continuous model of measurement, it is assumed that differences among test scores can be meaningful across the full range of the test score scale. According to Graham and Bergquist (1975), measurement on a continuous scale will arise whenever (1) each item in the universe taps "certain learned capabilities . . . from which only a single behavior can be demonstrated" (p. 6), (2) the

universe is composed of many such items, and (3) the capabilities required to answer one item correctly are not perfectly correlated with the capabilities required to answer another item correctly. (The reason for less than perfect inter-item correlation must be that the capabilities needed to answer different items are, in some way or other, truly different, and not just because of measurement error.) Ebel's (1962) vocabulary test, which would be described as domain referenced (in the sense that this expression is used by Hively, 1974; Hively, Patterson, & Page, 1968; Millman, 1974; and Shoemaker, 1975) conforms to this continuous model.

Two additional points about the continuous model should be noted. First, the items of a universe could conform to this model and yet have the same difficulty indices and inter-item correlation coefficients. It seems more likely, however, that most universes of interest will be heterogeneous in one or both of these respects. Second, unlike the binary model, the continuous model provides no basis for expecting test scores to be distributed in any particular way. How scores are distributed will depend on the distribution of true capabilities in the population being tested and on the statistical properties of measurement error.

In the case of the binary model, where the distinction is clearly between mastery and nonmastery, there should be little difficulty in establishing an appropriate passing score for the test. The score distribution will be bimodal, and if the test is long enough, there should be a negligible proportion of examinees in the middle of the score range. It is when there is a test that is better described by the continuous model, and it is used to classify examinees as masters and nonmasters, that difficulties arise. Not only is the choice of a passing score made arbitrarily (see Glass, 1978), but this score is likely to be in the region of the score range that is quite heavily populated, with the result that minor differences in passing score can have serious effects on the resulting classification. Although the authors recognize that continuous measures often have to be used to make binary decisions (e.g., promotion, hiring, selection), it is preferred that this be made explicit and that the argument not be couched in terms of masters and nonmasters. If a measure is best described by the continuous model, its primary role is to ascertain degrees of skill or competence, not to separate two distinct groups.

Examining score distributions should provide a clear guide to the types of criterion-referenced tests that are in common use. The expectation is that only a tiny minority could be described by the binary model, the vast majority being clearly continuous measures. Graham and Bergquist (1975) described some carefully constructed tests of unitary skills that yielded bimodal score distributions, but these appear to be exceptional.

The point is that there are at least two distinct types of criterion-referenced tests and that the desirable properties for one type are different from the desirable properties for the other. A single concept of reliability cannot be expected to suffice for the two of them. In the case of the continuous model, where a range of acceptable performance exists (and/or a range of unacceptable performance), and the purpose of the test is seen as being to ascertain where a person's performance lies along that continuum, there may be legitimate inquiry about the precision with which that person's position is located. There may also be interest in determining the consistency with which the test can distinguish among a group of people with respect to their location along that continuum. Reliability, as traditionally defined, describes how effectively the measurement process can do this. It is a useful property by which to describe measurements and one that can be investigated by traditional means. It has not, however, been greatly favored by writers on CRM.

Why not? One would think that whether the focus of the measurement is criterion referenced or norm referenced, and whether individual scores are compared with one another or with externally derived standards, a measurement process powerful enough to distinguish small differences would be prized. That this is not always seen to be so reflects the emphasis placed by those who work in the area

on classification, which is often a result of the measurement process, rather than on the measurement process itself. Hambleton and Novick (1973), for example, argued that with criterion-referenced tests "the pertinent question is whether or not the individual examinee has attained some prescribed degree of competence on an instructional performance task. . . . Questions of precise achievement levels and comparisons among individuals on these levels seem to be largely irrelevant" (p. 160). This idea has led to the view that reliability for criterion-referenced tests should mean consistency of decision-making. Hambleton and Novick went on to declare (1973, p. 166) that "in order to evaluate the test, it would be necessary to know something about the consistency of decision-making across parallel forms of the criterion-referenced test or across repeated measurements (i.e., reliability)." Implicit in this is an interesting new definition of reliability, one which has little connection with previous (technical) usage of the term. This definition was made explicit in a later paper (Swaminathan, Hambleton, & Algina, 1974) in which the authors wrote, "Specifically we define reliability of a criterion-referenced test as the measure of agreement between the decisions made in repeated test administrations" (p. 264).

There can be no argument that reliability, as so defined, is an important attribute of a criterion-referenced test, particularly where the test is used in conjunction with the new instructional models that involve individualized instruction, self-pacing, and regular checking of progress (see Hambleton, 1974). But it should be recognized that this meaning of reliability refers to a property of the *decisions* that are made when the test is used in a particular manner, rather than a property of the measurements provided by the test. In spite of all the work done on the problem of setting cutting scores (see, e.g., Brennan & Lockwood, 1980; Fhanér, 1974; Huynh, 1976c, 1977; Millman, 1973; Wilcox, 1976), the methods usually employed by educators to do this are arbitrary. (This is evident from the fact that the mathematical work in choosing a cutoff score along the observed score scale starts with the assumption that a standard has already been defined, either on the true scale or on the scale of a criterion measure; how this standard gets defined is never dealt with satisfactorily.) For a single criterion-referenced test, different sets of decisions should be expected depending on where the cutoff score is located. Some of these decisions would presumably be better than others; and some would presumably be more consistent than others, particularly if the measurement process has greater precision at some points along the scale than at others.

Users of tests, whether norm referenced or criterion referenced, have long recognized that it makes little sense to claim that a test is valid, but that questions about the validity of particular uses or interpretations of test scores can and should be asked. Cronbach (1971) put the argument as follows:

> The phrase "validation of a test" is a source of much misunderstanding. One validates, not a test, but an "interpretation of data arising from a specified procedure." A single instrument is used in many different ways—Smith's reading test may be used to screen applicants for professional training, to plan remedial instruction in reading, to measure the effectiveness of an instructional program, etc. Since each application is based on a different interpretation, the evidence that justifies one application may have litttle relevance to the next. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test "is valid." (p. 447)

Similar considerations have to be made with regard to questions of reliability. In particular, it is important to distinguish between the measurements made using a test and the decisions made on the basis of those measurements. A test that yields relatively good measurements (i.e., measurements that

on the average have a relatively small standard error) would be expected to yield decisions that are, in some sense, also relatively good (e.g., consistent); but the choice of cutoff score may be such that for a particular application this same test yields relatively poor decisions. On the other hand, a test that provides relatively poor measurements would be expected to result in relatively poor decisions; but again, the choice of cutoff score may be such that the test actually supports decisions that are relatively good.

The view of the authors is that in relation to criterion-referenced tests the term reliability has been used to describe a range of attributes, some belonging to the measurements yielded by the test, and others describing the decisions made when the test is used in a particular manner. Both are important, although referring to both as reliability encourages confusion and is therefore questionable. In particular, it seems unwise to use the phrase "test reliability" when it is the use of a test to make particular kinds of decisions that is being referenced. The use of more explicit terminology such as "reliability of decisions in domain-referenced testing" (Huynh, 1976b) and alternative names such as "index of dependability for mastery tests" (Brennan & Kane, 1977a) is thus encouraging. Hambleton, Swaminathan, Algina, and Coulson (1978) have reinforced this important distinction of differentiating the reliability of domain score estimates from the reliability of mastery classification decisions.

Because of considerations such as these, the two-dimensional framework defined by the following questions is offered as an aid to the organization of discourse about the reliability of criterion-referenced tests:

1. Is what is being assessed a property of the measurements yielded by the test, or of the decisions made on the basis of those measurements?
2. What assumptions are made about the nature of the measurements yielded by the test? In particular, is what is being measured a variable that, given sufficient accuracy of measurement, should be dichotomous, or one that should be continuous?

It is the contention of the authors that the issues of interest in a systematic treatment of test reliability will be different, depending on how these questions are answered. Four situations, which can be depicted as in Figure 1, are distinguished. The variable itself may be seen as continuous or binary, and the intention may be to measure the variable or to take some decision on the basis of the measurements made of the variable. The four situations may be described as follows.

Situation 1. Measurements of a continuous variable are used to make decisions about pupils. It will be assumed that the decisions are binary (e.g., pass-fail, promote-retain), although some procedures are available for dealing with classification into three or more categories (see, e.g., Huynh, 1978; Swaminathan, Hambleton, & Algina, 1975). In Situation 1, the intended use of test scores is to make decisions. Questions can be raised about the consistency of those decisions or about the losses associated with incorrect decisions.

Situation 2. Measurements of a binary variable are used to make decisions, which must be binary as well. The intended use of the test scores is again to make decisions. As in Situation 1, concern may be with the consistency of the decisions or with an examination of losses. This situation differs from Situation 1, though, in that correct classifications can be distinguished from misclassifications, whereas in Situation 1 serious misclassifications (e.g., a failed pupil who had achieved at a level well above the cutoff score) can also be distinguished from minor misclassifications (e.g., a failed pupil who had actually surpassed the cutoff score by the barest of possible margins).

Figure 1
A Framework for Organizing Issues Pertaining to the Reliability of Criterion-Referenced Tests

Intended use of test score	Type of variable: Continuous	Type of variable: Binary
Decision-making	1	2
Measurement	3	4

Situation 3. The variable being measured is continuous, and the focus of attention is on the measurements themselves, rather than the making of binary decisions. Reliability can be defined in terms of consistency over repeated measurements or over equivalent measures. A domain-referenced test, for which the purpose of measurement is to estimate an examinee's standing on some well-defined universe of tasks or items, fits this situation. Conventional wisdom would recommend the use of traditional reliability coefficients in this situation, but they do have at least one serious shortcoming. Taking a (randomly) parallel forms correlation as an example, it is possible that the two forms could rank pupils in almost the same way, although yielding score distributions that are different in central tendency and/or variability. There would then be a high correlation, even though the two forms yielded scores that differed substantially. For a domain-referenced test, where the aim is to estimate the proportion of items known in a well-defined universe of tasks, and not just order the examinees with respect to one another, this hardly seems adequate.

Situation 4. The measurement properties of a binary measure may be documented in a number of ways. The most desirable property of such a measure is accuracy of classification, and a corollary of this is consistency of classification. Certainly in practice, and perhaps also in concept, Situation 4 will be difficult, if not impossible, to distinguish from Situation 2.

A Review of the Literature

For the most part, the conceptualizations and associated procedures that can be found in the literature on CRM reliability fall into Situation 1 of Figure 1. Consequently, this situation is treated first and at greatest length in the subsections that follow.[1] The other situations defined in Figure 1 are treated only briefly near the end of the paper.

The outline followed in the ensuing description of conceptualizations and procedures for Situation 1 is this: A conceptual framework for decision-making is constructed and three kinds of loss due to decision errors are defined; each kind of loss is treated separately, the presentation of related developments being concentrated on concepts and issues. Readers who wish to delve deeply into technical details are advised to consult the original sources cited.

[1]For a different but very useful treatment of the concepts and coefficients the authors have assigned to Situation 1, see Kane and Brennan (1980).

A Conceptual Framework for Decision-Making

Situation 1 of Figure 1 is focused on the measurement of continuous variables, the resulting measurements being used to make binary decisions. This decision-making situation can be described as follows. Suppose that the application of a CRT to a person yields an observed score, say x. In the CRM literature it is usual to assume that the test is composed of dichotomously scored items, hence that x is the number of items answered correctly or the proportion of items answered correctly, or for multiple-choice items a score corrected for the effects of chance success. For the present, however, no restriction on the derivation and meaning of x need be imposed. If x exceeds or equals a predetermined cutoff score, say c, the person is said to have passed; otherwise, he or she is said to have failed. The decision, then, involves comparing x with c to decide whether or not the person has passed. (Note that the word pass is used simply to describe the fact that the person's test score is above or equal to the cutoff score.)

The possibility and source of decision error in this situation becomes clear if it is imagined that x is the realization of a true or latent measurement, say τ, plus a random perturbation, say ε. Depending on the circumstances, τ can be taken to be either a true score in the classical sense or a universe score in the sense of generalizability theory. If, corresponding to cutoff score c on the observed score scale, there is cutoff score γ on the true score scale, then the ideal or true decision would be to pass a person if his or her true score τ equals or exceeds γ; otherwise, the person should be failed. A comparison of the decision based on the observed score x with that based on the true score τ reveals whether or not a decision error has occurred.

Decision-making, as it has been outlined, is summarized in Figure 2. Incorrect decisions fall into cells (1,0) and (0,1), the former being false negative decisions and the latter being false positive decisions. The other two cells contain the correct decisions.

The problem is that of somehow describing the goodness of the dichotomous decision situation. One approach is to work with the information that is either contained in, or can be derived from, the situation depicted in Figure 2. Note that there is a precedent in the literature on CRM for viewing the correspondence between decisions based on the observed score and true score variables as decision validity (see, e.g., Harris, 1974; Millman, 1979). But just as the reliability of a test in classical theory can be defined as the squared correlation between observed and true scores (Lord & Novick, 1968, p. 61), so too, the goodness of decision-making as reflected by the correspondence between observed-score and true-score-based decisions can be seen as decision reliability.

Figure 2
The Dichotomous Decision Situation

		True Score Variable	
		$\underline{\tau} < \gamma$	$\underline{\tau} \geq \gamma$
Observed Score Variable	$\underline{x} \geq c$	(0,1)	(1,1)
	$\underline{x} < c$	(0,0)	(1,0)

A second approach to describing the goodness of the dichotomous decision situation is to study the correspondence between decisions based on the observed scores for two or more parallel tests. This approach, too, has its analogue in traditional reliability theory, the correlation between observed scores on parallel tests.

Each of these approaches to describing the goodness of decision-making has stimulated the development of a corpus of research and writing; hence, they have been used implicitly in organizing this review. In addition, however, explicit use has been made of an overriding organizational concept, that of loss. A function, say L, has been defined that specifies the losses that can accrue whenever decisions are made. The losses for the four cells of Figure 2 can be specified as in Figure 3 (adapted from van der Linden & Mellenbergh, 1977).

Three ways of assessing the losses due to decision errors have been employed in the CRM literature.

1. *Threshold loss*, in which case all losses associated with a particular decision error are assigned an equal value. To be more specific, under threshold loss the values of the loss function given in Figure 2 might be defined as follows: $\ell_{01}(\tau)=a$, $\ell_{10}(\tau)=b$, $\ell_{11}(\tau)=\ell_{00}(\tau)=0$, where a and b are real numbers.
2. *Linear loss*, in which case the magnitude of the loss associated with an error of a particular kind is a function of the distance between γ and τ. In this case, there might be $\ell_{01}(\tau)=b_1(\tau-\gamma)+a_1$, $\ell_{10}(\tau)=b_0(\tau-\gamma)+a_0$, and $\ell_{11}(\tau)=\ell_{00}(\tau)=0$, where b_1, b_0, a_1, and a_0 are real numbers.
3. *Squared-error loss*, in which case the loss associated with an error is a function of the square of the distance between γ and τ.

Given L, γ, and either the joint distribution of x and τ or, equivalently, the distribution of τ and the conditional distribution of x given τ, it is possible to define a risk function, say R, as the expected loss over the population of examinees. With R so defined (note that γ is presumed known) it is possible to choose c in such a way that risk, i.e., the value of R, is minimized by the decision-making procedure. Here, however, it is assumed that c is also given; the interest is in describing, in some sense, the goodness of the decision-making procedure. Methods for doing this, such as they exist, are considered for each type of loss.

Figure 3
Values of Loss Function L for Dichotomous Decisions

		True Score Variables	
		$\underline{\tau} < \gamma$	$\underline{\tau} \geq \gamma$
Observed Score Variable	$\underline{x} \geq c$	$\ell_{01}(\underline{\tau})$	$\ell_{11}(\underline{\tau})$
	$\underline{x} < c$	$\ell_{00}(\underline{\tau})$	$\ell_{10}(\underline{\tau})$

Threshold Loss: Probability of Decision Error

Suppose c, γ, and the joint distribution of x and τ are known for some population of interest. If the losses $\ell_{01}(\tau)=a$ and $\ell_{10}(\tau)=b$ are also known for false positive and false negative decision errors, respectively, then it is possible to compute the expected loss that would be incurred due to decision errors for a sample of any given size drawn at random from the population. Alternatively, the losses due to different kinds of decision errors can be ignored, in which case a natural and useful way to describe the goodness of the decision-making procedure is by the probability of committing each type of decision error or by the overall probability of committing a decision error, regardless of type. These quantities can be obtained by double-integration (where "integral" is interpreted in the Stieltjes sense, Kenny & Keeping, 1951, p. 24) of the bivariate distribution of x and τ over the appropriate portions of the ranges of these variables (see Figure 2).

An obvious problem with the foregoing proposals is that in practice the bivariate distribution of x and τ is not known. Moreover, the available methods for estimating this distribution are not very satisfactory, particularly if the sample size is small (Lord & Novick, 1968, p. 527). Nevertheless, in the literature on CRM reliability, Livingston (1978), Livingston and Wingersky (1979), and Wilcox (1977a) refer to work by Keats and Lord (1962) and Lord (1965, 1969) that addresses this problem. It is necessary here to restrict attention to tests composed of dichotomously scored items, in which case an observed score on the test is simply the number of items answered correctly. Also, the proportion-correct true score ξ must be used, not the number-correct true score τ. (Note that $\tau = n\xi$, where n is the number of items in the test. Note, also, that the symbol γ is everywhere used to represent the true cutoff score; the context in which it appears indicates whether it is in the scale of τ or ξ.)

Three crucial assumptions are made in the work of Keats and Lord (1962): (1) that the conditional distribution of observed scores x for given true score ξ is the binomial distribution; (2) that the regression of ξ on x is linear; and (3) that the distribution of ξ is continuous. It can be deduced from the first two assumptions that the observed scores are distributed as the negative hypergeometric distribution. The addition of the third assumption leads to the conclusion that the beta distribution (see, e.g., Kenny & Keeping, 1951, pp. 95–97) is a "reasonable" (Keats & Lord, 1962, p. 71) distribution for ξ. It is interesting to note that when this model applies, the squared correlation between x and ξ equals Kuder-Richardson coefficient 21 (KR-21) for the reliability of a test.

Wilcox (1977a) has employed this negative hypergeometric model (also known as the beta binomial model) to derive the equations needed to estimate α, the probability of committing a false-positive decision error, and β, the probability of committing a false-negative decision error. In addition, Wilcox has derived equations for this same purpose that stem from use of a variance stabilizing inverse sine transformation of proportion-correct scores.

Wilcox (1977a) evaluated the probability estimates that his equations yield in a monte carlo study. He concluded, "In general, but not always, increasing [test length] will decrease the value of α. Consequently, we lower the value of w_α [the difference between true α and its estimated value $\hat{\alpha}$] by increasing [test length] primarily because we obtain more accurate estimates of α when α is small. Increasing [sample size] with [test length] fixed also lowered w_α but at a somewhat slower rate" (p. 304). As regards the two different procedures that Wilcox devised for estimating the probability of decision error, the one based on the variance stabilizing transformation "usually performed as well as [the other procedure], and frequently it gave substantially better results" (p. 304). Regardless of which procedure was employed, however, the discrepancy between the estimated and true values of α were strikingly large relative to the true values of α. This points to the desirability of large samples of subjects and long tests when this procedure is used.

Further evidence bearing on this general conclusion was provided recently by Huynh (1980). He has developed formulas for obtaining asymptotic estimates of the standard errors of α and β. Huynh found that the differences between these standard errors and the actual standard errors computed in a simulation study seem to vary as a function of test reliability. Of the five sets of test characteristics for which Huynh simulated data, the two sets having a relatively large value of KR-21 were associated with relatively small differences in standard errors, regardless of sample size. When KR-21 was relatively small, however, the difference between asymptotic and actual standard errors was also relatively small only when the sample size was large.

The work of Lord (1965, 1969) employs the assumption that the conditional distribution of x given ζ is the compound binomial distribution (Lord & Novick, 1968, pp. 524–526). This assumption means that the probability that an examinee answers a question correctly is free to vary from one item to another. In Lord's work this freedom exacts a price—the need to estimate for each item the relationship between ζ and the probability of responding correctly. (Note here ζ_i is the mean over the items in the test of the probability that examinee i answers an item correctly.) This relationship cannot be estimated without incurring sampling errors, which, as Lord and Novick (1968, p. 525) have observed, are likely to be relatively large when cumulated over a number of items. Lord (1965) proposed a solution of a different sort—the development of a finite series expansion of the compound binomial and use of only the first few terms of this expansion to approximate the compound binomial distribution. It is necessary to specify an unknown function of ζ in the second term of this expansion. Lord has chosen to do so in such a way that the resulting squared correlation between x and ζ is equal to Kuder-Richardson coefficient 20 (KR-20).[2]

Lord's work with the compound binomial underlies two developments of interest here. One of these was by Wilcox (1977a). He employed the two-term approximation to the compound binomial distribution, while retaining the assumption that the true scores ζ are distributed as the beta distribution, to derive other formulae for estimating the error probabilities α and β. In the monte carlo study he conducted, Wilcox found that the compound binomial approximation effected very little reduction in the size of the difference between the estimated and true probabilities of a decision error over what that difference was when the binomial distribution was used. In the other development, Lord (1969) devised a method in which the form of the distribution of ζ is not specified. This is Method 20 (Lord,

---

[2]The CRM literature is not particularly clear as to when it is appropriate to assume that the distribution of x for given ζ is binomial and when it is appropriate to assume that this distribution is compound binomial. The authors' present understanding of this matter stems in part from a reading of Keeping (1962): Given a universe of items, and provided (1) that the proportion of items in the universe that a particular examinee can answer correctly is fixed during the time the examinee is being tested, (2) that the items an examinee answers are chosen at random, and (3) either that the number of items in the universe is infinitely large or that items are sampled with replacement, then the distribution of x for an examinee with a given true score ζ will be binomial. Note that this distribution will be binomial whether or not the items are equal in difficulty, where the difficulty of an item is the proportion of examinees in the population who can answer it correctly. The compound binomial distribution is the appropriate model when the probability that an examinee answers an item correctly varies from item to item (i.e., from trial to trial of the binomial process). Note that the probability of a correct response being referenced here is the probability for a particular examinee and a particular item; for a fixed item the probability in question will vary from examinee to examinee. Hence, the aforementioned measure of the difficulty of an item does not provide a satisfactory estimate of the probability that a given examinee will answer the item correctly. Moreover, it is impractical in educational testing to contemplate estimating this probability by repeatedly administering an item to an examinee. What is required instead is strong theory. Lord (1965, 1969) found a use for the compound binomial distribution in the context of one-dimensional latent trait theory. If the responses of a population of examinees to the items in a universe are well modeled by a one-dimensional latent trait theory, then knowledge of an examinee's score on the latent trait and knowledge of the item characteristic curve is sufficient to determine the probability with which the examinee answers the item correctly (Lord & Novick, 1968, chap. 16).

1969). Other features of Method 20 have been summarized by Livingston and Wingersky (1979) as follows:

> Method 20 is based on the assumptions that (1) the (conditional) distribution of observed scores, for persons with a given true score, is a compound binomial distribution (approximated by a four-term Taylor series expansion), and (2) the true score distribution is "smooth." Before estimating the true-score distribution, Method 20 divides the observed-score range into intervals. The model for the true-score distribution contains one parameter for each interval. Method 20 estimates these parameters and uses the resulting estimated true-score distribution to predict the number of persons in each observed-score interval. It then compares these predictions to the actual observed-score distribution and computes a chi-square goodness-of-fit. (p. 257)

As noted by Livingston and Wingersky, this procedure can be repeated using more or fewer observed-score intervals. A preferred solution is then chosen on the basis of the "smoothness of the estimated true score distributions and the goodness-of-fit of the predicted observed-score distributions" (p. 257).

Method 20 is not without its limitations:

> The test must be *unspeeded*, and there must be *no correction for guessing*. Ideally, every person should have a chance to answer every item, and the person's score must be simply the number of items answered correctly. The only limit on the length of the test is a practical one; the existing computer program will not accept a test of more than 100 items. Also, a very short test limits the complexity of the estimated true-score distribution. But even as few as six items will allow for a bimodal distribution with unequal modes. (Livingston, 1978, p. 3)

In addition, the use of Method 20 is not recommended for sample sizes smaller than 1,000 (Wingersky, Lees, Lennon, & Lord, 1969). This requirement limits use of the method to relatively large-scale test administrations. An example of the application of Method 20 to data arising from the administration of a 30-item test to over 3,000 examinees is reported by Livingston and Wingersky.

Two other ways of dealing with the probability of decision error should be noted. One is attributable to Divgi (1978). He proposed that the probability of misclassification be plotted as a function of true score ζ. This curve will have a cusp at cutoff score γ; the probabilities associated with the values of $\zeta < \gamma$ pertain to false positive errors, and the probabilities associated with values of $\zeta \geq \gamma$ pertain to false negative errors. As Divgi noted, "Such a curve provides a complete description of the error in decisions based on the test. However, a single number is much more convenient. The natural candidate, which does not require any arbitrary choice, is the rate at which [the error probability] decreases from its maximum value at [γ]" (1978, p. 4). Divgi went on to say that "a higher value of [this rate index] implies a smaller probability of incorrect classification of any true score, and therefore a smaller expected loss for any given distribution of true scores in the population" (p. 4).

To employ Divgi's approach, it is necessary to assume the form of the distribution of x for given ζ. Divgi has provided an illustration in which Lord's two-term approximation to the compound binomial distribution is used. Divgi has shown that his rate index increases directly as the square root of any increase in test length, and within limits, it increases directly as the variance of item difficulty indices increases. This rate index also varies with the cutoff score γ, being smallest, as would be expected, for cutoff scores in the midrange and largest for cutoff scores at the extremes of the true score scale.

A criticism of Divgi's rate index is that it has no upper bound. Divgi himself suggested a transformation of the index so that it lies in the interval [0,1]. Unfortunately, however, this transformed quantity has "no simple interpretation" (Divgi, 1978, p. 5).

Another way of dealing with decision error probabilities has been suggested by Wilcox (1979b). He proposed the calculation of upper and lower bounds on the probabilities of false positive and false negative decisions. This can be done making no assumptions whatsoever concerning the form of the distribution of true scores in the population of examinees. The method outlined by Wilcox requires only that it be possible to estimate the mean and variance of the distribution of true scores. It is necessary, of course, to assume the form of the conditional distribution of observed scores x given true score ζ. For this purpose, Wilcox considered both the binomial distribution and Lord's approximation to the compound binomial distribution. Given either of these distributions, Lord (1965) has provided the required estimates of the mean and the variance of the (unobserved) distribution of true scores. Wilcox developed the rationale and equations needed to calculate upper and lower bounds for the probabilities of decision error.

Threshold Loss: Decision Consistency[3]

Describing the goodness of the decision-making process by the probability of making an error has not been promoted by many writers on the topic of CRM reliability, despite its conceptual appeal. This circumstance is due in part, no doubt, to the unavailability of really good small sample methods for estimating the probabilities of false-positive and false-negative errors.

An approach that does stem reasonably directly from psychometric tradition is suggested by the concept of decision consistency. This phrase refers to the proportion of the examinees who are classified the same way with respect to cutoff score c on two tests that are administered independently, yet designed to measure the same characteristic. Figure 4 depicts this situation. Consistency, as just defined, is the sum of the proportions p_{00} and p_{11} in Figure 4.

The approach to evaluating a decision-making process that is suggested by the notion of decision consistency embraces several virtues: First, it can be implemented on relatively small samples of persons, e.g., samples the size of a school class, although it must be remembered that when the sample size is small, the estimated proportions have sampling errors that are correspondingly large. Second, this approach does not require the adoption of any particular method for deriving test scores from item responses. Third, there is flexibility in the fact that replicate measurements can be obtained in several different ways—by readministering the same test, by administering two test forms that are parallel in the classical sense (Lord & Novick, 1968, pp. 47-50) or by administering two test forms that are parallel in the sense that they were formed by random sampling from the same universe. Given the existence of a universe, it is even possible to imagine the situation in which each person is administered a different pair of tests, these having been formed by sampling at random from the universe. Conceptions of true score and error will differ, of course, depending on the method used to generate replicate measurements.

Aside from the fact that decision consistency fails to provide important information, namely the probability of making a wrong decision, it has another defect. The coefficient of decision consistency will usually be substantial even when the results on one test are independent of the results on the other test. If a standard statistical treatment of contingency tables (see, e.g., Keeping, 1962, §11.18) is

[3]See Subkoviak (1980) for a review focused solely on decision-consistency approaches.

Figure 4
Joint Decision Probabilities for Two Tests
$(p_{10}+p_{11}+p_{00}+p_{01} = p_{.0}+p_{.1} = p_{1.}+p_{0.} = 1)$

		Test 1 Results: $\underline{x}_1 < c$	Test 1 Results: $\underline{x}_1 \geq c$	
Test 2 Results	$\underline{x}_2 \geq c$	p_{10}	p_{11}	$p_{1.}$
	$\underline{x}_2 < c$	p_{00}	p_{01}	$p_{0.}$
		$p_{.0}$	$p_{.1}$	

followed and if the scores on Test 1 are statistically independent of scores on Test 2, then the proportion of times that x will be greater than or equal to c on both tests is equal to the product of the corresponding proportions of times x is greater than or equal to c on each test separately. A similar argument defines the expected proportion of times x will be less than c on both tests. This consideration led Swaminathan et al. (1974) to propose the use of coefficient kappa (κ) as a measure of decision consistency. The formal definition of κ for the bivariate decision situation depicted in Figure 4 is as follows:

$$\kappa = (p_0 - p_c)/(1 - p_c) \quad [1]$$

where

$$p_0 = p_{11} + p_{00} \quad [2]$$

and

$$p_c = p_{0.}p_{.0} + p_{1.}p_{.1} \quad [3]$$

In effect, κ defines the proportion of times two tests yield consistent decisions beyond the level of consistency that can be expected when scores on the two tests are independent.

The relationship between p_0 and κ is not a simple one. It depends on the shape of the distributions of scores on Tests 1 and 2, whether they are unimodal and symmetric or not; on the magnitude of the correlation between scores on the two tests; and on the location of the cutting score c. In view of this, Millman (1979) recommended that both indices be computed and reported, especially in view of the fact that the additional work involved is negligible.

The statistic κ has been the subject of much study (e.g., Brennan & Prediger, 1977; Cohen, 1960; 1968; Everitt, 1968; Fleiss, Cohen, & Everitt, 1969; Hubert, 1977; Kraemer, 1979). One of the interesting facts about κ is that if the marginal proportions for Tests 1 and 2 are the same (see Figure 4), κ

is equal in value to ϕ, the Pearson product-moment correlation coefficient between dichotomous scores on the tests (Cohen, 1960). (Dichotomous scores can be achieved by converting the scale of x to the binary scale w through the following transformation: Set $w = 1$ when $x \geq c$; otherwise, set $w = 0$.) Even when the marginal distributions of Tests 1 and 2 are not identical, Reid and Roberts (1978) found that κ and ϕ correspond very closely under simulated test conditions and recommended the use of ϕ. Their recommendation should not be accepted uncritically, however, because the interpretations of κ and ϕ are very different.

Brennan and Prediger (1977) and Livingston and Wingersky (1979) questioned the use of κ as a measure of decision consistency. At issue is the fact that if a sampling theory were desired for κ, that theory would have to be based on only those samples from the population of examinees that provide test results of a very restricted kind; the marginal proportions of the contingency table constructed for each sample would have to be exactly the same as the marginal proportions of the contingency table that were observed and used in the calculation of κ. This is because, under the assumption that the results for the two tests are independent, the expected proportions of the sample falling in the cells of the contingency table are estimated using the observed marginal proportions as estimates of the marginal proportions in the population (Keeping, 1962, p. 316). Brennan and Prediger (1977, pp. 6, 7) argued from this fact to the conclusion that p_c provides a "reasonable" measure of the extent of agreement that can be expected when the scores on two tests are statistically independent only when the marginal proportions of the contingency table are fixed in advance. Livingston and Wingersky (1979, p. 250) drew much the same conclusion. Of course, to fix the marginal proportions in advance would be anathema to those users of CRM who set cutoff score c before administering the test and then let the results determine the proportions of examinees who fall above and below c.

There seems to be no basis in the statistical treatment of contingency tables with which the authors are familiar, however, for concluding that the marginal proportions must be fixed prior to the administration of the tests. In the statistical treatment referred to, the marginal proportions are accepted as observed, without prior restriction; but the sampling distribution of κ is restricted to a limited set of examinee samples, namely, those that give contingency tables having the same marginal proportions as were observed. Whether or not a sampling theory for κ is needed depends on whether or not one wants to test the significance of, or construct a confidence interval for, κ. In any case, κ provides a measure of decision consistency in the same way that a Pearson product-moment correlation coefficient provides a measure of association, whether or not the available sampling theory for the correlation coefficient is used.

A more serious problem with κ has also been noted by Brennan and Prediger (1977, p. 7). If the marginal proportions for Tests 1 and 2 are not identical, then, as Cohen (1960) has shown, the maximum value of κ is less than 1, just as the maximum value of ϕ under the same circumstances is less than 1. For this situation, Brennan and Prediger (1977) noted the possibility of using the statistic $\kappa/\max(\kappa)$, where $\max(\kappa)$ is the maximum value of κ that is possible for a given contingency table in which the marginal proportions for Test 1 differ from those for Test 2.

Finally, with reference to κ, note another suggestion from Brennan and Prediger (1977, p. 10). For examinees being categorized into two mutually exclusive categories on the basis of test results, Brennan and Prediger have suggested the use of "½" as the expected marginal proportions for each category. It has been suggested that these proportions apply if the assignment of examinees to categories is made at random and, moreover, that these proportions remove the necessity, when formulating a sampling theory for contingency tables, of restricting that theory to samples that yield tables having the same marginal proportions as the observed table. Following this suggestion to its conclusion, if the assignments made by one test are independent of the assignments made by the other, then the

probability that both tests will assign an examinee to the same category (i.e., the probability either that $x \geq c$ on both tests or that $x < c$ on both tests) will be $(\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{2} \times \frac{1}{2}) = \frac{1}{2}$. A measure of decision consistency can then be defined (Brennan & Prediger, 1977, p. 10) as $\kappa_2 = (p_0 - \frac{1}{2})/(1 - \frac{1}{2})$. This measure of agreement was also suggested by Livingston and Wingersky (1979, p. 250), who show that it is formally identical to the G index of Holley and Guilford (1964). Prospective users should note that κ_2 will differ from κ, except, of course, when the observed marginal proportions of the contingency table are equal to ½, in which case the two coefficients are identical.

An obvious difficulty with the decision consistency approach is that results from two test administrations are required. This difficulty has sparked the development of methods whereby p_0 and κ can be estimated from the results of a single test administration.

Huynh (1976b) has described a method that can be viewed as stemming from the negative hypergeometric or beta-binomial model described by Keats and Lord (1962), although Huynh derived the method using a Bayesian approach. The three main assumptions stated by Keats and Lord, presented earlier, lead to another deduction: that "the bivariate distribution of randomly parallel tests is the bivariate hypergeometric distribution" (1962, p. 71). Using this deduction, Huynh's proposal can be summarized as follows:

1. Use the available distribution of observed scores to compute the required parameters of the two-parameter beta distribution of true scores;
2. Assume that these same parameters would characterize the beta distribution of true scores for a randomly parallel test were it available;
3. Apply the formulas provided by Keats and Lord (1962, pp. 62, 66) or by Huynh (1976b, pp. 254, 255) to compute the theoretical relative frequencies of the marginal and joint distributions of scores on the available test and its hypothetical, randomly parallel counterpart; and
4. Aggregate the appropriate ones of these relative frequencies to obtain the proportions required to compute p_0 and κ (see Figure 4).

Huynh (1979) shows how to compute the standard errors of these quantities.

A second method of estimating decision consistency from the results of a single test administration has been proposed by Subkoviak (1976). The fundamental assumption underlying this method is that for a specified person, the distribution of observed scores on each of two randomly parallel tests is binomial and therefore a function only of test length and the person's true (proportion correct) score for the universe of items from which the test items were sampled. If a person's observed scores on different randomly parallel tests are independent, then the probability that an observed score for this person will exceed cutoff c is the same for all tests of length n. Given both c and the estimated true score for person i, say $\hat{\xi}_i$, and assuming that the conditional distribution of x for given ξ_i is the binomial distribution, the estimated probability that this person's observed score equals or exceeds c is

$$p_i(x \geq c) = \sum_{x=c}^{n} \binom{n}{x} \hat{\zeta}_i^{\,x} (1 - \hat{\zeta}_i)^{n-x} \qquad [4]$$

The probability that this person's observed scores on two randomly parallel tests are either both less than c or both greater than or equal to c, under the assumption that the observed scores on the two tests are independent except, of course, for their mutual dependence on $\hat{\xi}_i$, is

$$p_{0i} = p_i(x \geq c)^2 + [1 - p_i(x \geq c)]^2 \qquad [5]$$

Averaging p_{0i} over i yields, for the group of persons being studied, an estimate of p_0, the index of decision consistency. As noted by Subkoviak (1976, p. 268), it would be feasible, possibly even desirable, to modify the foregoing model by substituting Lord's approximation to the compound binomial distribution for the binomial distribution itself.

In applying Subkoviak's approach, one problem that must be resolved is that of estimating the true score ξ_i. Subkoviak suggests several possibilities, of which three are noted here.

1. The proportion of test items answered correctly. This is the maximum likelihood estimate of ξ_i, and according to Subkoviak, it "should lead to reasonably accurate results if [the number of items in the test is greater than] 40, particularly if the mastery level of most students is well above or below [a true score equal to] .50" (1976, p. 269).
2. The linear regression estimate of true score, i.e., $\hat{\xi}_i = \hat{\rho}_{xx'}(x_i/n) + (1-\hat{\rho}_{xx'})\bar{x}/n$, where $\hat{\rho}_{xx'}$ is the estimated coefficient of reliability, x_i/n is the proportion-correct score of person i, and $\bar{x}/n$ is the mean proportion-correct score. As Subkoviak emphasizes, it is reasonable to employ this regression only when the distribution of observed scores is negative hypergeometric and the conditional distribution of x for given ξ can be presumed to be the binomial distribution, in which case the regression (reliability) coefficient is KR-21 (Keats & Lord, 1962).
3. The non-linear regression estimate due to Lord (1959). Unfortunately, this regression is "not uniquely determined by the observed score distribution" (Lord & Novick, 1968, p. 514), and the method requires large samples, but it does not rest on any distributional assumption other than that the conditional distribution of x for given ξ is binomial.

A third method for estimating decision consistency from the results for a single test has been proposed by Marshall and his collaborators (Marshall, 1976; Marshall & Haertel, 1975; Marshall & Serlin, 1979). The index that has been proposed is "the mean (over persons) proportion (over all possible test splits) of consistent mastery [i.e., pass-fail] decisions on a hypothetical double-length test" (Marshall & Serlin, 1979, p. 3). Several ways have been suggested for modeling the distribution of scores on the double-length test (Marshall & Serlin, 1979), including the negative hypergeometric or beta binomial model, Lord's compound binomial model, and the binomial model with linear regression estimates of true scores that was employed by Subkoviak.

Some attention has been paid in the recent literature on CRM reliability to studying the characteristics of different single-trial procedures for estimating decision consistency (Algina & Noe, 1978; Subkoviak, 1978). Interest extends to Subkoviak's procedure because its use with small (class-sized) samples can be easily justified (Hambleton et al., 1978, p. 22). Algina and Noe (1978) focused their study on Subkoviak's procedure. Simulating tests that might be described as homogenous—"the average within-examinee variance of the [item true scores] was small" (Algina & Noe, 1978, p. 105)—and keeping responses to different items by the same examinee independent, these investigators compared the proportion-correct estimate with the linear regression estimate of true scores. The results suggest that the true proportion of consistent classifications can be estimated most accurately using the linear regression estimate of true scores. In addition, Algina and Noe concluded that under the conditions studied, there is substantial bias in the estimated proportions of consistent classifications only when the cutoff score is near the mean score and the classical reliability coefficient (i.e., $\rho_{xx'}$) exceeds .48.

Subkoviak (1978) compared three of the single trial procedures using real data, the responses of 1,546 students (the population) to 50 items (the universe) drawn from the verbal sections of a Scholastic Aptitude Test. The cutoff score and length of test were varied; class-sized samples of students were

drawn. Not much difference was found among the estimated proportions of consistent classifications that were obtained using the different procedures. All procedures gave biased estimates under some conditions and all had small sampling variances as compared with the sampling variance of the estimates obtained by administering two tests to the examinees. (This latter result should be expected because the single test procedures model scores for a second test, and these modeled scores are likely to conform more closely to scores on the available real test than would the scores on a second real test.) Subkoviak recommended use of Huynh's procedure for the reason that "it is mathematically sound, requires only one testing, and produces reasonably accurate estimates, which appear to be slightly conservative for short tests" (1978, p. 115).

This conclusion, however, should be treated cautiously. Wilcox (1979a) has summarized the negative features of the beta-binomial model on which Huynh's procedure rests:

1. Estimates of the parameters of the beta distribution of true scores can sometimes yield unacceptable negative values of the probability density;
2. The parameter estimates can be very different from their true values even when sample size is relatively large;
3. The beta-binomial model permits the distribution of true scores either to have only one mode or to be U-shaped; and
4. The assumption that the conditional distribution of x for a given ζ is binomial "must be viewed as an oversimplification of the 'true' situation when an item sampling model applies. More specifically [as noted by Lord and Novick (1968, p. 524) and as acknowledged by Huynh (1976c)], the binomial conditional distribution is justified when the observed scores of different examinees are distributed independently of one another; however, this independence does not exist in the usual case where all examinees take the same random sample of n items" (Wilcox, 1979a, p. 246).

The discussion of one-trial estimates of decision consistency is concluded on the following note of warning: These methods all require strong assumptions. Whichever assumptions are made, the resulting estimates of decision consistency are only approximations to the estimates that would result were two tests used instead. The one-trial estimates are likely to be biased, and they almost certainly will have sampling variances that are unrealistically small compared with the sampling variances of the two-trial indices they approximate.

Linear Loss

As in the consideration of threshold loss, let knowledge of c, γ, and the joint distribution of x and ζ be assumed. Suppose, too, that the loss due to a decision error is specified as a linear function of the difference between the true score ζ and the true cutoff score γ; e.g.,

$$\left.\begin{aligned} \ell_{01}(\zeta) &= b_1(\zeta-\gamma) + a_1 \\ \ell_{10}(\zeta) &= b_0(\zeta-\gamma) + a_0 \\ \ell_{11}(\zeta) &= \ell_{00}(\zeta) = 0 \end{aligned}\right\} \qquad [6]$$

where a_0, a_1, b_0, and b_1 are real numbers. The expected value could then be computed of the loss that would be incurred under these conditions for a randomly constituted sample of any given size drawn

from the population. This expected value, also called the Bayes risk or the risk (van der Linden & Mellenbergh, 1978), would not be directly interpretable, being a function of sample size and the numbers a_0, a_1, b_0, and b_1, as well as the choice of c, γ, and the form of the functional relationship between x and ζ.

Attempts to define a standardized index of risk, or its complement, can be found in the CRM literature for linear loss functions (and, as will be seen shortly, for squared-error loss functions). These attempts are peculiar in a way that motivates this digression. Gains due to correct decisions—in the complement, losses due to incorrect decisions—are not set to zero as Equation 6 suggests they should be. If it is supposed, therefore, that incorrect decisions contribute a negative quantity to the complement of a standardized index of risk, then correct decisions contribute a positive quantity to this index, with the size of the contributions, both positive and negative, being determined in the case of linear loss by the extent to which an examinee's true score departs from the true cutoff score. An index obtained in this way reflects the relative extent to which the positive contributions of correct decisions outweigh the negative contributions of incorrect decisions. It is debatable, perhaps, whether or not indices of this sort, because they do not conform to expectations for what a standardized index of risk or its complement should be, should be treated as exemplars of work on Situation 1 of Figure 1. They should, however, because these indices are sensitive to variation in the location of the cutoff score c. In other words, these indices have been designed for use in situations where pass/fail decisions will be made, and they are sensitive to changes in the rule for making these decisions.

Two attempts to define standardized indices were encountered in the literature on CRM. Livingston and Wingersky (1979) have described an index of decision-making efficiency, symbolized here as I_{DME}:

$$I_{DME} = \frac{\text{Expectation}[(\zeta-\gamma).\text{Sign}(x-c)]}{\text{Expectation}[(\zeta-\gamma).\text{Sign}(\zeta-\gamma)]} \qquad [7]$$

In computing this index, all decisions are considered—correct as well as incorrect—and false-positive errors are seen as just as serious as false-negative errors. (Livingston and Wingersky offer a variant of Equation 7 for use when the two kinds of decision errors are weighted differently.) The requisite information for computing I_{DME} was obtained in an application described by Livingston and Wingersky through use of the previously noted Method 20 attributable to Lord (1969). Recall that Method 20 serves to estimate the joint distribution of x and ζ. It is clear from Equation 7 that losses in efficiency due to decision errors—these occur when the sign of $(x-c)$ differs from that of $(\zeta-\gamma)$—are a function of the (linear) difference between ζ and γ.

Another standardized index has been suggested by van der Linden and Mellenbergh (1978). This index, whose lineage can be traced to Huynh (1976a), as defined as follows:

$$\delta = 1 - (R-R_c)/(R_n-R_c) \qquad [8]$$

where

R is the expected risk for the decision-making situation as it has been observed,

R_n is the risk for the situation in which observed scores are a monotonically increasing function of true scores, and

R_c is the risk for the situation in which observed scores are independent of true scores.

As defined, δ increases as risk decreases. Moreover, if R lies in the interval from R_c to R_n inclu-

sive—this is not necessarily true (van der Linden & Mellenbergh, 1978, p. 212; see also Wilcox, 1978)—δ will fall in the interval from 0 to 1 inclusive: "a value of 0 signifies that the test is worthless, and a value of 1 signifies that the test is perfect for the decision situation" (van der Linden & Mellenburgh, 1978, p. 121).

Suppose, now, that the following linear loss function is used to evaluate R:

$$L(\zeta) = \begin{cases} b_0(\zeta-\gamma) + a_0 & \text{for } x<c \\ b_1(\zeta-\gamma) + a_1 & \text{for } x\geq c \end{cases} \quad [9]$$

[Note that the correct decisions make a nonzero contribution to $L(\zeta)$]. Suppose, too, that the linear regression of ζ on x is used to estimate the unknown value of ζ for given x. Then, the standardized index δ assumes the value $\rho_{xx'}$, where $\rho_{xx'}$ is the reliability coefficient. In this development, $\rho_{xx'}$ is the same reliability coefficient as the one used to obtain the linear regression estimate of ζ.

Squared-Error Loss

Three lines of development in the CRM literature employ a squared-error loss function to measure the seriousness of decision errors. Each of these developments is similar in that the losses due to correct decisions are *not* set to zero.

One line of development is due to van der Linden and Mellenbergh (1978). They define the loss function

$$L(\zeta,x) = [f(x)-\zeta]^2 \quad [10]$$

where $f(x)$ is an estimate of ζ based on the observed score x. It can be shown (van der Linden & Mellenbergh, 1978, pp. 122–123) that when $f(x)$ is the linear regression estimate of ζ given x, the risk is

$$R = \sigma_\zeta^2(1-\rho_{xx'}) \quad [11]$$

where σ_ζ^2 is the variance of true scores ζ. The standardized index δ is the reliability coefficient $\rho_{xx'}$, provided that when ζ and x are independent, the variance of ζ is greater than zero. (In classical test theory σ_ζ^2 is zero when ζ and x are independent, in which case δ is undefined. This was not noted by van der Linden & Mellenbergh, 1978.)

Wilcox (1978) has identified several problems with this line of development. The linear regression used to estimate ζ includes $\rho_{xx'}$, which is not known and must be estimated. The risk function becomes complicated when an estimate of $\rho_{xx'}$ is substituted for $\rho_{xx'}$ itself, and the function "cannot be evaluated theoretically" (p. 611). In this case δ, too, cannot be evaluated.

A second line of development involving squared-error loss was initiated by Livingston (1972). His by now familiar treatment of CRM reliability is an adaptation of the derivations leading to the coefficient of reliability in classical test theory. In essence, Livingston substituted cutoff score c for the means μ_x and μ_τ in the expressions for the observed and true score variances of classical test theory. This leads to the coefficient

$$k^2(x,\tau) = D^2(\tau)/D^2(x) \quad [12]$$

where

$$D^2(x) = \sigma_x^2 + (\mu_x-c)^2 \quad [13]$$

and

$$D^2(\tau) = \rho_{xx'}\sigma_x^2 + (\mu_x - c)^2 \qquad [14]$$

Livingston's index has many interesting properties. Normally, it is greater than $\rho_{xx'}$, the exceptions being when $\mu_x = c$, and $\rho_{xx'} = 1$. It does not shrink to zero or become undefined through lack of score variance as long as c is sufficiently removed from μ_x. And, of course, when $\mu_x = c$, it becomes identical with $\rho_{xx'}$.

Upon superficial examination, many of these results might seem paradoxical when compared with the results of classical test theory. They are not, and if they seem so, it is only because of false expectations engendered by familiarity with classical results. As Livingston has explained,

> The point is simply that the farther any person's obtained score is from the criterion [cutoff] score, the more confident we can be in saying that his true score is on the same side of the criterion score. Then if two groups of scores have equal variance and equal reliability in the norm-referenced sense, the group of scores whose mean is farther from the criterion score must have the greater criterion-referenced reliability. (1972, p. 18)

Since the publication of Livingston's (1972) seminal article, the formulation it contains has been extended by Lovett (1977, 1978), who has provided two alternative derivations of Livingston's coefficient. One of these is from the perspective of the analysis of variance of item scores, and the other is based on the assumption that the distribution of dichotomous item scores for a given true score (i.e., for a given examinee) is binomial. Although these derivations yield results that appear to be equivalent to Livingston's, it must be remembered that Livingston used only the assumptions of classical test theory; these are not necessarily valid for dichotomously scored items. (See, for example, Lord & Novick, 1968, chap. 23.) At the very least, for Lovett's (1977, 1978) and Livingston's (1972) derivations to yield similar interpretations, the items that Lovett considers must be strictly equal in difficulty for each examinee.

The ideas of Livingston (1972) have been elaborated in the third line of development from the perspective of squared-error loss by Brennan (1977, 1978, 1980), Brennan and Kane (1977a, 1977b), and Kane and Brennan (1980). This elaboration is founded on generalizability theory (Brennan & Kane, 1979; Cronbach et al., 1972) in the same way that Livingston's (1972) proposal is founded on classical test theory. The principal consequence of this substitution of one theory for the other is a conception of measurement error that Brennan and Kane, among others, think is highly appropriate for domain-referenced measurement, whenever this kind of measurement is achieved.

The nature of this conception of measurement error can be easily grasped if it is recalled that for a domain-referenced test the existence of an infinite pool or universe of items and an infinite population of examinees is imagined. Let ξ_i be the proportion of items in the universe that examinee i can answer correctly, π_j be the proportion of examinees in the population who can answer item j correctly, $(\xi\pi)_{ij}$ be the effect due to the interaction of examinee i with item j, and ε_{ij} be the residual error. Then the (dichotomous) observed score, say x_{ij}, of randomly selected examinee i on randomly selected item j can be modeled as follows:

$$x_{ij} = \mu + (\xi_i - \mu) + (\pi_j - \mu) + (\xi\pi)_{ij} + \varepsilon_{ij} \qquad [15]$$

where μ is the expected value over examinees and items of x_{ij}. Under the required assumptions of analysis of variance (see, e.g., Lord & Novick, 1968, pp. 162–166), the variance of x_{ij} over items and

examinees can be expressed as follows:

$$\sigma^2(x) = \sigma^2(\zeta) + \sigma^2(\pi) + \sigma^2(\zeta\pi, \varepsilon) \quad [16]$$

where

$\sigma^2(\xi)$ is the variance, over the population of examinees, of ξ_i;
$\sigma^2(\pi)$ is the variance, over the population of items, of π_j; and
$\sigma^2(\xi\pi,\varepsilon)$ is the variance, over the population of examinees and the universe of items, of $(\xi\pi)_{ij}$ and ε_{ij}.

(These interaction and residual effects are inextricably confounded if each examinee responds only once to each item.)

Consider now the usual definition of measurement error as the difference between observed score x_{ij} and true score ξ_i. From Equation 15 this error, say Δ_{ij}, is as follows:

$$\Delta_{ij} = (\pi_j - \mu) + (\zeta\pi)_{ij} + \varepsilon_{ij} \quad [17]$$

in which case the variance over items and examinees of Δ_{ij} is

$$\sigma^2(\Delta) = \sigma^2(\pi) + \sigma^2(\zeta\pi, \varepsilon) \quad [18]$$

This is the error variance recommended by Brennan and Kane (1977a, 1977b), by Cronbach et al. (1972), and much earlier by Lord (1955), for judging an examinee's observed score in relation to a predetermined cutoff score γ. There can be no quarrel with this recommendation when different randomly selected examinees or groups of examinees take different randomly selected sets of items from the universe.

Suppose, however, that all students take the same test. If the purpose of this test is to estimate the universe (or domain) scores of the examinees, then reference must be made to Lord and Novick (1968, pp. 187-191) for a means of describing, within the framework of generalizability theory, the quality of the measurements provided by the test for this purpose. Alternatively, if interest does not extend to the examinees' probable performances on any of the other distinct tests that could be formed, by whatever means, from the universe, then the appropriate error of measurement would appear to be what Cronbach et al. (1972) call δ. It is defined as follows:

$$\delta = (x_{ij} - \zeta_i) - (\pi_j - \mu) \quad [19]$$

and has variance

$$\sigma^2(\delta) = \sigma^2(\zeta\pi, \varepsilon) \quad [20]$$

This is the error variance that Lord and Novick (1968, p. 167) refer to as the "*generous estimate* of the specific error variance" for the test. (The "specific error variance" for a test is the variance of errors in the classical test model.) For a full discussion of these different conceptions of measurement error, see Lord and Novick (1968), Cronbach et al. (1972), Brennan and Kane (1977a, 1977b), and Brennan (1978, 1980).

Using the error of measurement Δ, Brennan and Kane (1977a) have defined an index of dependability for CRM, symbolized here as ID(γ). This index is analogous to Livingston's (1972) $k^2(x,\tau)$; it can be specified as follows:

$$ID(\gamma) = \frac{\sigma^2(\zeta) + (\mu-\gamma)^2}{\sigma^2(\zeta) + (\mu-\gamma)^2 + \frac{1}{n}\sigma^2(\Delta)} \quad [21]$$

(Note that ID(γ) is an index for proportion-correct scores.) A means of estimating ID(γ) is provided by Brennan and Kane (1977a, pp. 280–282).

Some of the characteristics of ID(γ) are worthy of note. It has the form of a reliability or generalizability coefficient except, of course, for the presence of the term $(\mu-\gamma)^2$ in the numerator and denominator. The presence of this term means that, like $k^2(x,\tau)$, ID(γ) can be relatively large even when the true score variance $\sigma^2(\xi)$ is 0 or nearly so. Also, ID(γ) is at its smallest when $\gamma = \mu$. In this case, there is the special index,

$$ID(\mu=\gamma) = \sigma^2(\zeta)/[\sigma^2(\zeta) + \frac{1}{n}\sigma^2(\Delta)] \quad [22]$$

which Brennan and Kane (1977b) refer to as a "general purpose index of dependability" (p. 617). In practice, of course, μ is not known. It is possible, however, for γ to be set equal to the sample mean proportion-correct score (i.e., $\bar{x} = \sum_i\sum_j x_{ij}/nN$, where N is the size of the examinee sample and n is the size of the item sample). Brennan (1977) has shown that in this case, ID(γ) is equal to coefficient KR-21, a result that would be expected on the basis of Lord's (1955) work. Brennan has also shown that $ID(\mu=\gamma) \geq$ KR-21. This supports the recommendation of Brennan and Kane (1977b) to compute $ID(\mu=\gamma)$ as a general-purpose index of dependability whenever possible. Of course, KR-21 is easily computed, so if it can be assumed that $\bar{x}$ is not very different from μ, then KR-21 would be preferred on practical grounds.

The Remaining Situations

Situation 2 of Figure 1 is defined by a binary variable that is measured for the purpose of decision-making. In the earlier discussion of binary variables, the point was made that repeated attempts to measure a binary variable, as when students respond to n dichotomously scored test questions and receive for their efforts a score equal to the number of questions answered correctly, invariably result in scores that range from 0 to n, even though the belief for this situation is that all scores should be either 0 or n. This might suggest that all of the proposals for CRM efficiency that were considered for Situation 1 can also be employed for Situation 2. Close inspection of this suggestion, however, shows it to be false. Under the binary model, the distribution of relative true scores ξ will possess nonzero density at only two points, $\xi = 0$ and $\xi = 1$. (Note again that these are Platonic true scores, *not* the true scores of classical test theory, Lord & Novick, 1968, chap. 2.)

All approaches for assessing decision efficiency in which it is assumed that the distribution of true scores is smooth and continuous—this is true of approaches based on the beta-binomial model and Lord's Method 20—cannot apply to Situation 2. Also inapplicable are those approaches in which it is assumed that the conditional distribution of observed scores for a given true score is either the binomial or the compound-binomial distribution; this assumption is invalidated by the appearance of observed scores different from 0 and n. Only one of the approaches discussed for Situation 1 is clearly applicable to Situation 2: the decision consistency approach of Swaminathan et al. (1974).

There are two other approaches. One of these is Harris' (1974) squared point-biserial coefficient of correlation between the observed test scores and the corresponding dichotomous decision scores. The second development is by Macready and Dayton (1977). In their approach, examinees are assumed to

fall into one of two groups—masters, all of whom have Platonic true scores of 1, and nonmasters, all of whom have Platonic true scores of 0. The examinees are also assumed to respond to a set of n dichotomously scorable items; but instead of summing the scores on these items to produce an observed score on the scale from 0 to n, Macready and Dayton attend to the vector of n dichotomous item scores that can be formed for each examinee. The probability of occurrence of a given score vector is specified as the sum of the probabilities of two mutually exclusive events: (1) that the examinee was a master who answered items incorrectly because of forgetting, and (2) that the examinee was a nonmaster who answered items correctly by guessing. (In this formulation, the probability that a master answers an item correctly and the probability that a nonmaster answers an item incorrectly are, respectively, the complements of the forgetting and guessing probabilities.)

Macready and Dayton considered two different ways of modeling the probabilities associated with these two events. In one model the probability that a master forgets the answer to an item is left free to vary from item to item. Similarly free to vary is the probability that a nonmaster guesses the correct answer to an item. In the second model, the probability that a master forgets the answer to an item is constrained to be the same for all items. A similar constraint is placed on the probabilities that nonmasters correctly guess answers to items. The first of these models requires the estimation of $2n+1$ parameters—two probabilities for each of n items plus the probability that an examinee in the group tested is a master. The second model requires the estimation of only three parameters—the probability that masters forget, the probability that nonmasters guess correctly, and the probability that an examinee is a master.

The approach of Macready and Dayton should be attractive to all who take seriously the proposition that test variables can be binary. The following reasons apply:

1. Macready and Dayton show how to fit the two models to data using an iterative Newton-Raphson procedure and have prepared computer programs to accomplish this.
2. Using a statistic that appears to be distributed as chi-square, the adequacy of the fit of one or the other models to data can be assessed in absolute terms, and the adequacy of the fit of the two models can be compared.
3. Once a model has been fit to data, an examinee can be assigned to the mastery or the nonmastery group, depending on whether the probability of his or her score vector under the assumption that he/she is a master is greater or smaller than the probability of his/her score vector under the assumption that he/she is a nonmaster. (This assignment can be refined by taking the costs of incorrect decisions into account, if they are known.)
4. The probability of classification errors can be computed as a measure of the goodness of the decision rule.
5. It is possible to determine the minimum number of items needed to bring the proportion of misclassified examinees down to, or below, a level deemed acceptable.

Although Macready and Dayton considered only two ways of modeling the forgetting and guessing probabilities that are associated with each item, alternative models are possible. Indeed, Macready and Dayton themselves suggested several alternatives, but these were rejected on the ground that they "are not seen as having as much general applicability to criterion-referenced testing as [the models that were considered]" (p. 104). This conclusion undoubtedly merits examination by others interested in Situation 2.

Situation 3 of Figure 1 is defined by a continuous variable; moreover, a test of this variable is used for measurement purposes. Almost all of the CRM reliability literature reviewed does not bear on

Situation 3. (For an exception, see Kane & Brennan, 1980, who discuss the use of the coefficient ID(μ=γ) when domain-referenced test scores are not compared to a cutoff score.) The circumstance should not be viewed as a problem, however, because the vast literature on reliability theory and on generalizability theory can be referenced whenever a test serves the purpose of measurement defined by Situation 3.

Situation 4 of Figure 1 encompasses binary variables intended to serve the purpose of measurement. Dichotomously scored items can, perhaps, be viewed as tests for this situation. If so, interested readers should see Wilcox (1977b, 1977c).

Conclusions and Recommendations

Several conclusions are supported by the results of this review. The first is that most authors who have written recently on the topic of CRM reliability have addressed the problem of describing the goodness of pass/fail decisions based on test scores. Of primary interest has been the use of tests to decide whether or not a student should progress to the next unit of work in the subject being tested. This conclusion suggests the need to develop language that clearly conveys the message that it is a quality of the decisions based on test scores that is being described, not test reliability as it is traditionally understood. The requisite language has not been coined in this paper, but a classification of approaches to assessing the quality of decisions has been offered, a classification that may spur others to develop the needed terminology.

A second conclusion is that when a threshold loss function is employed, the goodness of decision-making is most meaningfully described as one probability or two—the probability of making an erroneous decision or, alternatively, the probability of making a false-positive decision error and the probability of making a false-negative decision error. What is needed, if possible, are methods that provide satisfactory estimates of these probabilities for class-sized groups of students.

A third conclusion is related to the second: Most of the proposals encountered for evaluating the goodness of decisions involve relatively strong theory and require difficult computations. There are no procedures now available that the classroom teacher can be expected to use routinely. Even the proposal from Swaminathan et al. (1974), which is simple in conception, will be resisted by classroom teachers because it requires the administration of two tests. If it is important for classroom teachers to be concerned about the goodness of decisions based on tests—because the quality of decisions based on short classroom tests will often be abysmally low—and if teachers are to make their judgments in meaningful terms (e.g., as the probability of a decision error), then procedures must be developed that are simple to use and that provide reasonably accurate results. Therein lies no small challenge.

A final conclusion follows only indirectly from the results of the review. There is an increasingly common use of CRM for which the procedures described above are, for the most part, not helpful. Both in curriculum evaluation and in surveys of educational achievement, the focus of interest is not on the individual student and the decisions being made about that student. Instead, the focus is on group characteristics—the proportion of a group responding correctly to a particular item or demonstrating mastery (however defined) on a subset of items. A good example is provided by the Australian Studies in School Performance, in which the authors wrote:

> It is important to emphasize that the tests were not constructed as normative tests for the purposes of grading and comparing students, but as criterion-referenced tests for assessing the

> achievement of mastery by students in the basic skills of reading, writing and number work. The tests were designed to find out what students could do and the skills they had mastered.
>
> As a consequence of employing tests of this type, the distribution of scores on these tests were markedly skewed and the use of traditional test statistics was considered to be largely inappropriate . . . .
>
> In reporting the results on individual items in this study, the research workers have presented the proportions who answered each item correctly and have invited readers to examine the item and the extent to which students achieved success on the underlying task. However, in reporting the results on the tests and subtests employed in the study, they have presented the proportions who have achieved mastery on each test or sub-test and the associated tasks. (Keeves, Matthews, & Bourke, 1978, p. 11)

For uses like this one, the focus is on the adequacy of group description rather than of individual decisions. Just as it is well known that the standard error of a group mean is considerably less than the error in the individual scores that contribute to it, so is it clear that the adequacy of a test for group description is somewhat greater than the adequacy of the same test when used for the description of individuals. One of the most important future lines of development will be in this direction. Better ways of describing the adequacy of tests when they are used for the purposes described above are needed.

References

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement*, 1978, *15*, 101-110.

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests (ACT Technical Bulletin No. 27). Iowa City, IA: American College Testing Program, 1977.

Brennan, R. L. Extensions of generalizability theory to domain-referenced testing (ACT Technical Bulletin No. 30). Iowa City, IA: American College Testing Program, 1978.

Brennan, R. L. Applications of generalizability theory. In R. E. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement*, 1977, *14*, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika*, 1977, *42*, 609-625. (b)

Brennan, R. L., & Kane, M. T. Generalizability theory: A review. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement* (No. 4). San Francisco: Jossey-Bass, 1979.

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.

Brennan, R. L., & Prediger, D. J. Coefficient kappa: Some uses, misuses, and alternatives (ACT Technical Bulletin No. 29). Iowa City, IA: American College Testing Program, 1977.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37-46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *71*, 213-220.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Divgi, D. R. *A new index for the accuracy of a criterion-referenced test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March 1978.

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97-103.

Fhanér, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 172-175.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, *15*, 237-261.

Graham, D, & Bergquist, C. *An examination of criterion-referenced test characteristics in relation to assumptions about the nature of achievement variables*. Paper presented at the annual meeting of the American Educational Research Association, Washington, March 1975.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research*, 1974, *44*, 371-400.

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., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Harris, C. W. Some technical characteristics of mastery 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: University of California, Center for the Study of Evaluation, 1974.

Hively, W. Introduction to domain-referenced testing. In W. Hively (Ed.), *Domain-referenced testing*. Englewood-Cliffs, NJ: Educational Technology Publications, 1974.

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

Holley, J. W., & Guilford, J. P. A note on the *G* index of agreement. *Educational and Psychological Measurement*, 1964, *24*, 749-753.

Hubert, L. Kappa revisited. *Psychological Bulletin*, 1977, *84*, 289-297.

Huynh, H. *On mastery scores and efficiency of criterion-referenced tests when losses are partially known*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976. (a)

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976, *13*, 253-264. (b)

Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, *41*, 65-78. (c)

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. *Psychometrika*, 1977, *42*, 601-608.

Huynh, H. Reliability of multiple classifications. *Psychometrika*, 1978, *43*, 317-325.

Huynh, H. Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, 1979, *4*, 231-246.

Huynh, H. Statistical inference for false positive and false negative error rates in mastery testing. *Psychometrika*, 1980, *45*, 107-120.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 1980, *4*, 105-126.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika*, 1962, *27*, 59-72.

Keeping, E. S. *Introduction to statistical inference*. New York: van Nostrand, 1962.

Keeves, J. P., Matthews, J. K., & Bourke, S. F. *Educating for literacy and numeracy in Australian schools*. Melbourne: Australian Council for Educational Research, 1978.

Kenny, F., & Keeping, E. S. *Mathematics of statistics* (Part 2; 2nd ed.). New York: van Nostrand, 1951.

Kraemer, H. C. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, 1979, *44*, 461-472.

Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement*. Washington, DC: National Council on Measurement in Education, 1979.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Livingston, S. A. *Reliability of tests used to make pass/fail decisions: Answering the right questions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March 1978.

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 1979, *16*, 247-260.

Lord, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 1955, *20*, 1-22.

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.

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

Lovett, H. T. Criterion-referenced reliability estimated by ANOVA. *Educational and Psychological Measurement*, 1977, *37*, 21-29.

Lovett, H. T. The effect of violating the assumption of equal item means in estimating the Livingston coefficient. *Educational and Psychological Measurement*, 1978, *38*, 239-251.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 1977, *2*, 99-120.

Marshall, J. L. The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data. (Technical Report No. 350.) Madison, WI: Wisconsin Research and Development Center for Cognitive Learning, 1976.

Marshall, J. L., & Haertel, E. H. *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement*. Paper presented at the annual meeting of the American Educational Research Association, Washington, March 1975.

Marshall, J. L., & Serlin, R.C. *Characteristic of four mastery test reliability indices: Influence of distribution shape and cutting score*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, *46*, 133-158.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, CA: McCutchan Publishing Co., 1974.

Millman, J. Reliability and validity of criterion-referenced test scores. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement* (No. 4). San Francisco: Jossey-Bass, 1979.

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

Reid, J. B., & Roberts, D. M. *A monte carlo comparison of phi and kappa as measures of criterion-referenced reliability*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Shoemaker, D. M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-147.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265-276.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, *15*, 111-116.

Subkoviak, M. J. Decision-consistency approaches. In R. E. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263-267.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87-98.

van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, 1977, *1*, 593-599.

van der Linden, W. J., & Mellenbergh, G. J. Coefficients for tests from a decision theoretic point of view. *Applied Psychological Measurement*, 1978, *2*, 119-134.

Wilcox, R. R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, *1*, 359-364.

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics*, 1977, *2*, 289-307. (a)

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox (Eds.), *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977. (b)

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox (Eds.), *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977. (c)

Wilcox, R. R. A note on decision theoretic coefficients for tests. *Applied Psychological Measurement*, 1978, *2*, 609-613.

Wilcox, R. R. A lower bound to the probability of choosing the optimal passing score for a mastery test when there is an external criterion. *Psychometrika*, 1979, *44*, 245-249. (a)

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. *Journal of Educational Statistics*, 1979, *4*, 59-73. (b)

Wingersky, M. S., Lees, D. M., Lennon, V., & Lord, F. M. A computer program for estimating true-score distributions and graduating observed-score distributions. *Educational and Psychological Measurement*, 1969, *29*, 689-692.

Acknowledgments

We are grateful to R. L. Brennan, H. Huynh, S. A. Livingston, F. M. Lord, M. J. Subkoviak, and R. K. Hambleton for their constructively critical comments on a draft of this manuscript. All remaining errors of fact and interpretation are, of course, our responsibility.

Authors' Addresses

Ross E. Traub, Department of Measurement, Evaluation and Computer Applications, The Ontario Institute for Studies in Education, 252 Bloor St. West, Toronto, Ontario, Canada M5S 1V6; Glenn L. Rowley, La Trobe University, Education Department, Bundoora, Victoria, Australia 3083.