A CONSUMERS' GUIDE TO CRITERION-REFERENCED TEST RELIABILITY

RONALD A. BERK
Johns Hopkins University

Over the past decade more than a dozen different statistics have been devised for measuring the reliability of criterion-referenced tests. Recent summaries and critiques of those measures have been conducted by Hambleton, Swaminathan, Algina, and Coulson (1978), Linn (1979), Millman (1979), Shepard (1980), and Traub and Rowley (in press). Hambleton et al. (1978, pp. 15–23) defined three major categories of reliability: (a) reliability of mastery classification decisions—consistency of mastery-nonmastery classification decision making across repeated measures with one test form or parallel test forms; (b) reliability of criterion-referenced test scores—consistency of squared deviations of individual scores from the cutting score across parallel or randomly parallel test forms; (c) reliability of domain score estimates—consistency of individual scores across parallel or randomly parallel test forms. Subsequently, in-depth analyses of the indices that fall into the first category that are based on a threshold loss function and the indices that fall into the second category that are based on a squared-error loss function have been presented by Subkoviak (1980) and Brennan (1980), respectively. These sets of indices are appropriate for criterion-referenced tests where mastery-nonmastery decisions are made on the basis of a cutting score. The indices that fall into the last category that are used to estimate the stability of an individual's domain score or proportion correct have received the least attention.

Despite the number of indices and the available sources noted above, a reliability index is rarely reported for criterion-referenced and minimum competency tests. When an index is reported, it is usually the Kuder-Richardson 20 or 21. In order to facilitate increased and proper use of the reliability indices by teachers, district and state level test makers, and test publishers, the indices have been compiled and evaluated in the form of a "consumers' guide." A comparison of the characteristics of the indices and an evaluation of their precision and practicability will be presented in the guide.

METHOD

A dozen different approaches that yield 13 reliability indices for criterion-referenced tests were identified and grouped into three categories: threshold loss function, squared-error loss function, and domain score estimation. Selection of a reliability index for a specific application was conceived as a two-stage decision process: first, choosing the appropriate category of reliability and second, choosing a specific index within the given category.

The initial choice of a reliability category is based on certain preliminary considerations related to the assumptions, interpretations, and uses of the indices. These considerations include: a clarification of terminology, test forms assumption, setting the cutting score, and test score interpretation and decision making.

The author acknowledges the very helpful comments of Robert L. Brennan and Samuel A. Livingston on an earlier version of this paper. This does not imply their endorsement of the structure and content of the review.

323
The choice of a specific index is based on the characteristics, precision, and practicability of each index. Practicability pertains to the ease with which an index can be computed and interpreted. A critical review of the dozen approaches and corresponding indices according to these factors is presented.

PRELIMINARY CONSIDERATIONS

A Clarification of Terminology

The use of the term reliability coefficient to characterize the indices recommended for criterion-referenced tests is inappropriate. This term is, in fact, a misnomer since the threshold loss and squared-error loss function approaches are neither defined nor interpreted in the language of the standard psychometric reliability components of observed-score and true-score variance (cf. APA/AERA/NCME, 1974). It is certainly possible to relate many of the indices to classical reliability theory or to describe them within a broad framework of variance components. However, their properties suggest that the term agreement index provides a more accurate label. The threshold loss function indices are measures of agreement between categorical data sets based on mastery-nonmastery classifications; the squared-error loss function indices are measures of agreement between the scores on two test forms.

Unfortunately, the domain score estimation procedures can not be explained as easily in terms of an agreement function. Four of the procedures involve a consideration of different types of standard error, while the other consists of a reliability-like index. Therefore, the statistics in this last category will not be relabeled.

Test Forms Assumption

All of the statistics reviewed in this paper require an assumption about test forms. It is assumed that two or more forms of a criterion-referenced test are either classically parallel or randomly parallel. The assumption is related to the characteristics of the item domain from which the item samples are generated and to how the item sample that defines one test form and the samples that define alternate forms are selected.

Classically parallel test forms. When items are written from objectives-based specifications using traditional item construction rules, those items viewed collectively constitute a sample from a theoretical domain of “all possible items” that could be written for the objective. When two or more samples are developed for the same objective and domain, those samples should contain content and yield means, variances, and item intercorrelations identical to the first sample. Item samples or alternate test forms with these properties are said to be classically parallel or equivalent. This statistical equivalence imposes restrictions on the domain. Very often when a test maker does not actually produce parallel forms, the alternate form(s) and its statistical properties are simply “assumed.” All of the agreement indices except Brennan’s (1980) index and one domain score estimation statistic (Lord & Novick, 1968) are based on this assumption.

Randomly parallel test forms. In contrast, the recent technological advances in the specification of behavioral domains offer alternative strategies that can generate “all possible items” in a domain. The progress that has been made over the past decade is reflected in the methodologies of amplified objectives, IOX test specifications, mapping sentences (facet theory), item transformations, item forms, and algorithms (See...
Millman, 1980; Popham, 1980; Roid & Haladyna, 1980). Proponents of the latter three strategies, in particular, claim that generating an item domain is an objective, mechanical process (Berk, 1979, 1980a). While the precision of the different strategies varies markedly and much more research lies ahead, the strategies do provide some evidence to refute the skepticism expressed by Thorndike in 1967, "As soon as we try to conceptualize a test score as a sample from some universe, we are brought face to face with the very knotty problem of defining the universe from which we are sampling (p. 285) ... the universe is considerably restricted, is hard to define, and the sampling from it is hardly to be considered random" (p. 288).

When one of the item-generation strategies is employed in test construction, the items to be included on one test form or alternate test forms are selected from the item domain using a random or stratified random sampling plan. These test forms are said to be randomly parallel. None of the statistical properties mentioned previously apply to these forms. When the test maker can not actually build test forms by randomly sampling from the domain, the alternate form(s) and its characteristics can more reasonably be "assumed" (See Cardinet, Tourneur, & Allal, 1976). Only one agreement index (Brennan & Kane, 1977a) and four domain score estimation statistics (Berk, 1980b; Brennan, 1980; Lord, 1957) assume randomly parallel test forms.

Setting the Cutting Score

Computation of the threshold loss and squared-error loss function indices presumes a cutting score for mastery of an objective has been selected. All of the indices are sensitive to the position of the cutting score in the score distribution. This relationship makes it essential that the interpretation of an index includes a specification of the cutting score. Therefore, the setting of standards is a crucial step prior to the consideration of an agreement index.

Setting standards is one of the most important, controversial and misunderstood validity issues in criterion-referenced measurement. A decision maker must confront the accuracy of mastery-nonmastery classification decisions. Evidence of accuracy expressed usually as probabilities of correct and incorrect (false mastery and false nonmastery) classifications pertains to decision validity (Hambleton, 1980, pp. 97–99). Since an extensive discussion of this topic is not within the scope of this paper, interested readers are referred to Hambleton (1980) and Hambleton, Powell, and Eignor (1979) for reviews of methods for setting cutting scores and estimating probabilities.

The major point of this section is that without validity evidence or a sound justification for setting the cutting score it seems pointless even to compute an agreement index. Certainly one can compute an index based on a cutting score that has little or no justification. Since the final interpretation links the index value to the cutting score, however, a high agreement index associated with an "invalid" or "unjustified" standard, for example, might indicate that a test can consistently classify students into the wrong groups. Consistent decision making without accurate decision making has questionable value in criterion-referenced evaluation.

In addition, a priori decisions regarding the relative seriousness of the losses related to the probabilities of false mastery and false nonmastery errors must be reached before choosing a loss function. Whether such losses are perceived to be equally serious (threshold loss function) or to vary as a function of how far above or below from the cutting score the misclassified students' scores are located (squared-error loss function) will be the determining factors in that choice.
Test Score Interpretation and Decision Making

The scores on a criterion-referenced test can be used for decision making about individuals and programs at the classroom, school, district, state and national levels. The decisions made by the classroom teacher within the context of mastery learning theory and diagnostic-prescriptive teaching have best demonstrated the utility of criterion-referenced test scores. Typically, those decisions are derived from an objectives-based assessment of individual strengths (mastery) and weaknesses (nonmastery). They involve placing students at the appropriate points in the instructional sequence, monitoring students' successes and failures as they progress through the program, and assigning letter grades/pass-fail to certify competencies in a subject area or course. Recently the latter type of summative evaluation has been extended in the form of minimum competency certification for grade-to-grade promotion and high school graduation. Even outside the public school arena, similar score interpretations are being contemplated for the certification and/or licensing of professionals in medicine and the allied health fields, social work, psychology, law, and engineering.

In addition to these different types of individual decisions, test scores may be used to assess program effectiveness at the decision levels mentioned previously. These evaluation decisions focus usually on the effectiveness of curricular materials or instructional methods. The actual decisions may be to continue, to modify, or to terminate a given program. In order to supply information that will be meaningful for these decisions it is often necessary to summarize the test score results into various forms such as: (a) percentage of students demonstrating mastery of each objective, (b) percentage of students demonstrating mastery of sets of objectives, (c) estimated mean performance on the total test, and (d) estimated mean performance based on the matrix sampling of students and items from their respective populations or domains. The appropriateness of these summaries will vary as a function of the level of the decision maker, from classroom teacher to state superintendent, and the nature of the decision. The summary that is needed for a particular decision is composed by aggregating the individual test scores at

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score and Decision Characteristics for Three Categories of &quot;Reliability&quot;</td>
</tr>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Score interpretation</td>
</tr>
<tr>
<td>Type of decision or information required for decision</td>
</tr>
<tr>
<td>Losses associated with decision errors</td>
</tr>
<tr>
<td>Threshold Loss</td>
</tr>
<tr>
<td>Individual score is referenced to cutting score</td>
</tr>
<tr>
<td>Mastery or nonmastery classification</td>
</tr>
<tr>
<td>Losses are equally serious</td>
</tr>
</tbody>
</table>
the appropriate decision level. Examples of several report formats have been assembled by Fremer (1978) and Mills and Hambleton (1980).

The initial choice of a "reliability" category is contingent upon the intended interpretation of the test scores, type of decision to be made with those scores, and the consequences or losses associated with false mastery and false nonmastery decision errors. A brief list of score and decision characteristics related to the three categories of "reliability" is shown in Table 1. Since the score interpretations for the numerous levels of program evaluation decisions are derived from aggregates of individual scores and their corresponding uses and interpretations, only individual score and decision characteristics are specified in the table.

THRESHOLD LOSS AGREEMENT INDICES

The concept of classification decision consistency using a threshold loss function was first proposed by Hambleton and Novick (1973, p. 168). The use of the threshold loss function assumes (a) a dichotomous, qualitative classification of students as masters and nonmasters of an objective based on a threshold or cutting score and (b) the losses associated with all false mastery and false nonmastery classification errors are equally serious regardless of their size.

The characteristics and evaluation of the six approaches in this category are outlined in Table 2. The various approaches actually involve only two agreement indices: \( p_o \), proportion of individuals consistently classified as masters and nonmasters across (classically) parallel test forms, and \( \kappa \), proportion of individuals consistently classified beyond that expected by chance. While two other indices developed by Carver (1970) might be considered in this review, they are omitted because they are insensitive to the consistency of individual classification decisions and only reflect whether the group percentage of masters stays the same.

\( p_o \) Versus \( \kappa \)

The choice between \( p_o \) and \( \kappa \) as indices of classification consistency depends upon how consistency is defined in a given decision context and how the properties of the statistics are evaluated. This comparison will concentrate on the appropriateness of the indices for tests used for individual decisions at the classroom level, individual certification decisions at the school level (e.g., minimum competency testing) and program evaluation decisions at the district level.

Advantages of \( p_o \). The \( p_o \) index measures the overall consistency of mastery-nonmastery classifications. Subkoviak (1980, p. 152) noted two factors that contribute to that consistency: the mastery-nonmastery composition of the group tested and the measurement precision or accuracy of the test itself. The index is sensitive to the selected cutting score, test length, and score variability. However, the position of the cutting score tends to have a much more profound effect upon the magnitude of the index than either of the other two characteristics. Higher values of \( p_o \) are associated with cutting scores at the tails of a unimodal score distribution and lower values occur with cutting scores near the mean (Subkoviak, 1980). This trend may not follow for bimodal distributions. The index values also increase as the number of test items increases and the score variance increases. In practice, however, it is possible to obtain \( \hat{p}_o \) values of .75 or higher with

\[ \text{All estimates of indices cited throughout the paper and the tables will be distinguished from the parameters with the symbol "d".} \]
Table 2
Evaluation of Threshold Loss Function Indices
(Listed in order of increasing overall complexity)

<table>
<thead>
<tr>
<th>Index</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Administrations</th>
<th>Test Forms Assumption</th>
<th>Distribution Assumptions</th>
<th>Correction for Chance Agreement</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_0$</td>
<td>Proportion of individuals consistently classified as masters and nonmasters on repeated measurements of one form or parallel forms</td>
<td>Hambleton and Novick (1973)</td>
<td>2</td>
<td>1 Form or Parallel</td>
<td>None</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Proportion of individuals consistently classified as above beyond that expected by chance</td>
<td>Swaminathan, Hambleton, and Algina (1974) based on Cohen (1960)</td>
<td>2</td>
<td>1 Form or Parallel</td>
<td>None</td>
<td>Yes *</td>
<td></td>
</tr>
</tbody>
</table>

*The $\kappa$ index is corrected for chance according to the marginal frequencies of the particular contingency table data set under consideration. This constraint may be viewed as a disadvantage of using $\kappa$ for tests intended for individual decisions at the classroom level.

Advantages: Provides unbiased estimate of $p_0$. Hand calculable and easily interpretable.
Disadvantages: Requires two test administrations. When only one test form is used, testing effect can spuriously deflate estimates. $p_0$ has comparatively large standard errors for classroom size samples (Subkoviak, 1978, 1980).

Advantages: Hand calculable; computer programs are also available (e.g., Antonak, 1977; Berk & Campbell, 1976; Wixon, 1979). $\kappa$ and $\phi$ coefficients are virtually identical (Marshall & Serlin, 1979; Reid & Roberts, 1978).
Disadvantages: Yields biased estimate of $\kappa$. Sensitivity to test length and variance may restrict range of values. Requires two test administrations. Testing effect based on test-retest design can spuriously deflate estimates. Dependence of index values on cutting score, marginal frequencies, test length, and score variance complicates index interpretation.
### Table 2 (continued)

<table>
<thead>
<tr>
<th>Index</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Adminis-trations</th>
<th>Test Forms Assumption</th>
<th>Distribution Assumptions</th>
<th>Correction for Chance Agreement</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta}_0 )</td>
<td>Same as ( \theta_0 ) defined previously</td>
<td>Marshall (1976)</td>
<td>1</td>
<td>Parallel</td>
<td>Binomial distribution of observed scores over repeated measurements (assumes 0-1 items, statistical independence of items, and equal item difficulties)</td>
<td>No</td>
<td>Average ( \hat{\theta}_0 ) across all possible split-halves of a single test, extended to a full-length (double-length test) estimate with Spearman-Brown type prophecy formula. Advantages: Requires only one test administration. Comparatively small standard errors for classroom-size samples (Subkoviak, 1978, 1980). Easily interpretable. Disadvantages: Yields biased estimates for short tests (Subkoviak, 1978, 1980). Provides no estimate of ( \kappa ). Computationally complex with no readily available computer program.</td>
</tr>
<tr>
<td>( \hat{p}_0 )</td>
<td>Same as ( p_0 ) and ( \kappa ) defined previously</td>
<td>Subkoviak (1976; 1980)</td>
<td>1</td>
<td>Parallel</td>
<td>Binomial or compound binomial distribution of observed scores over repeated measurements</td>
<td>No; Yes</td>
<td>Observed proportion correct scores and KR-20 coefficient are used to obtain linear regression approximations to the compound binomial ( \hat{p}_0 ). Relies on the estimation of the true score for each person. Consistency index for each person is averaged over the population to produce ( \hat{p}_0 ). Advantages? Same as ( \hat{\theta}_0 ). Disadvantages: Yields biased estimates of ( \hat{p}_0 ) for short tests (Algina &amp; Noe, 1978; Subkoviak, 1978). Sensitivity of ( \kappa ) to test length and variance may restrict range of values. Computationally complex with no readily available computer program.</td>
</tr>
</tbody>
</table>
Table 2 (continued)

<table>
<thead>
<tr>
<th>Index</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Administrations</th>
<th>Test Forms Assumption</th>
<th>Distribution Assumptions</th>
<th>Correction for Chance Agreement</th>
<th>Comments</th>
</tr>
</thead>
</table>
| \( \hat{p}_0 \) \( \hat{e} \) | Same as \( \hat{p} \) and \( \hat{e} \) defined previously | Huynh (1976) | 1                      | Parallel\(^a\)        | Binomial distribution of observed scores over repeated measurements Beta distribution of true scores | No; Yes | Indices computed from the univariate and bivariate distributions. Under similar distribution assumptions for double length tests \( p \) equals \( \beta(p_0) \) (Marshall & Serlin, 1979). 
Advantages: Requires only one test administration. Comparatively small standard errors for classroom-size samples (Subkoviak, 1978, 1980). \( p \) is easily interpretable. Based on the mathematically elegant Keats and Lord (1962) model. Provides most accurate estimates of \( p \) for unimodal distributions. Violation of equal item difficulty assumption seems to have negligible effect on estimates (Huynh & Saunders, 1979; Subkoviak, 1978). Tabled values of \( p \) and \( \hat{e} \) for tests with 5 to 10 items per objective as well as a computer program are currently available (Huynh, 1979; Subkoviak, 1980). A simple normal approximation method that is hand calculable has also been recommended (Peng & Subkoviak, 1980). 
Disadvantages: Yields biased (slightly conservative) estimates of \( p \) and \( \hat{e} \) for short tests (Huynh & Saunders, 1979; Subkoviak, 1978). Sensitivity of \( \hat{e} \) to test length and variance may restrict range of values. Most conceptually and mathematically complex approach. |

\(^a\)Huynh (1976, pp. 254-55, 1977, p. 1) defined two equivalent test forms as two independent random samples of items drawn from a specified item universe. While these test forms could be randomly parallel, the author indicates that they possess the properties of classically parallel forms essential for the assumptions of the beta-binomial model (Huynh, personal communication, November 1979).
<table>
<thead>
<tr>
<th>Index</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Administrations</th>
<th>Test Forms Assumption</th>
<th>Distribution Assumptions</th>
<th>Correction for Chance Agreement</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM</td>
<td>Kappanax (maximum value of $x$)</td>
<td>Huynh (1977)</td>
<td>1</td>
<td>Parallel$^a$</td>
<td>Same as Huynh's (1976)$^b$</td>
<td>Yes</td>
<td>Upperbound of $\tilde{x}$ is usually reached at a cutting score near the mean, although $KM$ is not a function of the cutting score. It reflects the situation in which a test functions best in terms of decision consistency. Advantages: Same as Huynh's (1976) $x$. Disadvantages: Same as Huynh's (1976) $\tilde{x}$. Interpretation and utility of $KM$ by itself or in conjunction with $\tilde{x}$ are unclear.</td>
</tr>
</tbody>
</table>

$^a$Huynh (1976, pp. 254-55, 1977, p. 1) defined two equivalent test forms as two independent random samples of items drawn from a specified item universe. While these test forms could be randomly parallel, the author indicates that they possess the properties of classically parallel forms essential for the assumptions of the beta-binomial model (Huynh, personal communication, November 1979).
These properties are especially important in view of the structure and uses of most teacher-made criterion-referenced tests.

Although these advantages indicate why $p_o$ might be the preferred index of agreement for tests intended for individualized instructional decisions at the classroom level, the disadvantages associated with the uses of $\kappa$ will demonstrate more dramatically why $\kappa$ may not be the preferred index for such tests. The rejection of $p_o$ in favor of Cohen's (1960) $\kappa$ coefficient was advanced originally by Swaminathan, Hambleton, and Algina (1974, p. 264). They felt that $\kappa$ was somewhat more appropriate since it took into account the agreement that could be expected by chance alone. The $\kappa$ index measures the test's contribution to the overall proportion of consistent classifications, that is, test consistency. It is calculated by subtracting from $p_o$ the proportion of consistency expected from the particular mastery-nonmastery composition of the group (Subkoviak, 1980). The properties of $\kappa$ that stem from this "correction for chance agreement" are what make this index problematic.

**Disadvantages of $\kappa$.** There are three properties of the $\kappa$ index that render it undesirable for criterion-referenced tests used for classroom decision making: the correction for chance is constrained by the marginal frequencies and $\kappa$'s sensitivity to test length and test score variability may restrict the range of values. All three properties complicate the interpretation of $\kappa$.

First, the correction for chance agreement is constrained by the marginal frequencies of the particular $2 \times 2$ contingency table data set under consideration. An index of +1.00 can be obtained only when the marginals on both forms or measurements are equal. Livingston and Wingersky's (1979) arguments on this issue are most persuasive:

Applying such a correction to a pass/fail contingency table is equivalent to assuming that the proportion of examinees passing the test could not have been anything but what it happened to be. For example, if 87% of the examinees passed the test, kappa will "correct for chance" under the assumption that "chance" would result in exactly 87% of the examinees passing the test. This assumption makes sense when the pass/fail cutoff is chosen on the basis of the scores to which it will be applied, so as to pass a specified proportion of the examinees. It does not make sense when the pass/fail cutoff represents an absolute standard that is to be applied individually to each examinee. (p. 250)

One condition under which the correction could be justified is where the selection of the cutting score is determined, in part, by the consequences (losses) of passing or failing a particular proportion or segment of the population. This cutting score is relative rather than absolute. Exemplary of this type of cutting score are the performance standards for many minimum competency tests that are set by district or state level decision makers. When these tests are used for individual certification decisions, the cutting score is usually adjusted according to the political, economic, social, and/or instructional consequences of not passing or certifying a certain proportion of the students in the school district.

Second, the values of $\kappa$ increase with test length. Since this is a characteristic of most reliability statistics, it should not be viewed necessarily as a disadvantage of $\kappa$. It is the consequences of this sensitivity that are troublesome. In criterion-referenced testing an item sample or subtest designed to assess mastery of an instructional objective generally consists of between 3 and 10 items on tests developed for classroom usage and of rarely...
CRITERION-REFERENCED TEST RELIABILITY

more than 20 items on tests developed for district or state level usage. For the majority of test applications at these levels, then, the values of \( \kappa \) may be restricted in range due to short subtest lengths. One would expect short tests would provide less reliable information and, hence, yield lower indices. This restriction would not be of concern in large-scale minimum competency testing programs where the items are often aggregated across objectives to apply a single cutting score to the total test and, subsequently, to estimate classification consistency based on that score.

Third, the values of \( \kappa \) increase with score variability. A restriction in variance that would be accompanied by a restriction in the range of \( \kappa \) will occur frequently in criterion-referenced testing due to short subtests as discussed above and to the homogeneous composition of the group that is tested. In many cases the test results tend to yield either a large proportion of masters or nonmasters. This is especially evident when classroom test scores are used for individual placement, formative and summative decisions. The former two types of decision are typically diagnostic in nature and made prior to instruction. At that point in instruction, the results usually indicate disproportionately large numbers of nonmasters. Conversely, summative decisions which are made at the conclusion of instruction are often based on results with disproportionately large numbers of masters.

In addition to these properties of \( \kappa \), there are other characteristics worthy of consideration: (a) \( \kappa \)'s sensitivity to the cutting score is just the inverse of \( p_o \), such that higher values of \( \kappa \) correspond to cutting scores near the mean score and lower values correspond to cutting scores at the tails of the distribution (Huynh & Saunders, 1979; Subkoviak, 1980), (b) \( \kappa \) is a biased estimator (Harris & Pearlman, 1977, p. 35), (c) one of the single administration approaches (Huynh, 1976) provides less accurate estimates of \( \kappa \) (about 10% error) than of \( p_o \), and (d) the interpretation and usefulness of the \( \kappa \) index in practice is not completely clear (See, for example, Safrit & Stamm, 1979).

Recommendations. The foregoing comparison of \( p_o \) and \( \kappa \) as indices of classification consistency suggests that \( p_o \) should be the index of agreement for criterion-referenced tests where an absolute cutting score is chosen and for other tests that may contain short subtests and/or yield low score variance. While \( \kappa \) may be a legitimate index of agreement for tests where relative cutting scores are set according to the consequences of passing or failing a particular proportion of the students, the problems associated with \( \kappa \) render it less useful than \( p_o \). Caution should be observed in its use and interpretation. Given the dependence of both indices on the cutting score, test length, score variability, and the cell frequencies of the 2 x 2 contingency table, it is recommended that the test maker report all of this information along with the agreement index to facilitate perceptive and correct index interpretations (See Livingston & Wingersky, 1979; Millman, 1974a).

One- Versus Two-Administration Approaches

Once an initial decision has been made to estimate \( p_o \) or \( \kappa \), a second decision about whether to use a one- or two-administration approach must be made. Among the six classification consistency approaches, the Hambleton and Novick (1973) two-administration method appears to have the greatest utility for classroom test construction. For tests developed at the district and state levels and by publishers either the Hambleton and Novick (1973) or Swaminathan et al. (1974) two-administration method or the Huynh (1976) single-administration method should be used.

Hambleton and Novick method. The Hambleton and Novick method provides the
only unbiased estimate of $p_o$. It is also the easiest method to understand, to compute and to interpret. Unfortunately, there are two problems that must be confronted: $\hat{p}_o$ has comparatively large standard errors (6 to 8%) for classroom-size samples ($n = 30$) (Subkoviak, 1980) and two test administrations are required.

With regard to the accuracy of the $\hat{p}_o$ estimate, teachers should be urged to interpret the index values with caution. Since the preliminary evidence indicates that the error decreases as the cutting score increases on short subtests, a consideration of about 6% error in evaluating classification consistency based on mastery standards at or above 80% seems advisable.

It is usually quite difficult and very time consuming for teachers to build two classically parallel tests for a given set of objectives. Therefore, it is recommended that the $\hat{p}_o$ estimate be computed from repeated measures of one test form. For many content domains the traditional weakness of testing effect associated with the test-retest design will be negligible particularly if the first testing is done after considerable instruction and retest data is collected as a review measure. When a significant testing effect is suspected, as when some teaching occurs to correct errors, the teacher should be cognizant of its influence on the decision-consistency estimates. Since knowledge gained from the initial testing tends to spuriously inflate the scores on the second testing, only the status of nonmasters on testing one will be altered on testing two. If several nonmasters become masters due to testing effect, $\hat{p}_o$ will be spuriously deflated. This conservative bias should be taken into account in interpreting the $\hat{p}_o$ index for teacher-made criterion-referenced tests.

**Hambleton-Novick/Swaminathan et al. method.** Depending on the selection of $p_o$ or $\kappa$, the Hambleton and Novick (1973) or Swaminathan et al. (1974) method should be employed by test makers at the district and state level and by test publishers. If parallel test forms are constructed for test security or other reasons, the two-form estimate is recommended over the single-form estimate to be discussed next. The two-form estimate is more accurate and is simpler to compute and to interpret.

**Huynh method.** When it is not possible to develop parallel forms, the more sophisticated test maker should seriously consider the Huynh procedure. Subkoviak (1980) has recommended this approach on statistical as well as practical grounds. It provides relatively precise though conservatively biased estimates of $p_o$ and $\kappa$. However, since the estimates are based on only one test administration, they do not reflect all sources of instability across repeated testings. Huynh’s (1976) method, as well as the other single-administration approaches, assume that a student’s knowledge and all of the test administration conditions remain unchanged from one testing to the next. This would never be upheld in practice. Consequently, the method may overestimate decision consistency even though the estimates of the parameters are conservatively biased.

Huynh’s (1976) beta-binomial model is one of the most conceptually, mathematically, and computationally complex threshold loss function approaches. Despite this complexity, it has several distinct advantages compared to the alternatives (See Table 2). In addition, a simple normal approximation of the beta-binomial distributions was recently tested by Peng and Subkoviak (1980). They found this approximation to provide relatively accurate estimates of $p_o$ and $\kappa$. This finding should be particularly meaningful to test makers without computer facilities since the approximation procedure can be hand calculated.
SQUARED-ERROR LOSS AGREEMENT INDICES

In contrast to the preceding loss function which focuses on the consistency of classifications, the squared-error loss approach to measurement error deals with the consistency of measurements or scores. It is based on the squared deviations of individual scores from the cutting score. This builds in a sensitivity to degrees of mastery and nonmastery along the score continuum, not just the qualitative master-nonmaster classification assumed by the threshold loss function. The losses associated with false mastery and false nonmastery errors are not assumed to be equally serious. The consequences of misclassifying students who are far above or below the cutting score are considered more serious than the consequences of misclassifying those who are close to the cutting score. The sensitivity to degree of mastery also means that a squared-error loss agreement index reflects the magnitude of all errors of measurement including those that do not lead to misclassification (Brennan & Kane, 1977a, p. 287). This is one fundamental problem with this loss function.

\( k^2(X,T) \) Versus \( \Phi(\lambda) \)

An evaluation of the two agreement indices in this category is shown in Table 3. The indices are quite similar in regard to what they are measuring, although the actual formulations are very different. The primary distinctions between the indices pertain to their assumptions about test forms (classically parallel or randomly parallel) and, based on these assumptions, their definitions of error variance or squared-error loss.

Brennan (1980) defined two types of error variance in terms of analysis of variance components: \( \sigma^2(\delta) \), person \( \times \) item interaction and \( \sigma^2(\Delta) \), item effect plus person \( \times \) item interaction. The former, which is derived from the assumption of classically parallel test forms, is part of Livingston's (1972a) \( k^2(X,T) \) index; the latter, which is derived from the assumption of randomly parallel test forms, is incorporated in Brennan's (1980) \( \Phi(\lambda) \) index.

There are several technical features common to both indices that should be mentioned: (a) no distribution assumptions are necessary, (b) as agreement indices, they are uncorrected for chance agreement (See Kane & Brennan, 1980), (c) index values can change, independent of their standard errors of measurement, (d) index values increase as the cutting score is set farther and farther from the mean (similar to values of \( p_0 \)), (e) index values increase as test length (number of items sampled) increases due to decreases in error variance, and (f) when there is no true or universe score variance, the indices can still have positive values as long as the true score mean does not equal the cutting score. It is also interesting to note that when the cutting score equals the sample mean, \( k^2(X,T) \) reduces to the KR-20 coefficient and \( \Phi(\lambda - \bar{X}) \) is identical to the KR-21 coefficient.

Criticisms of \( k^2(X,T) \). Criticisms of some of these technical features of Livingston's (1972a) index have been expressed by Harris (1972), Shavelson, Block, and Ravitch (1972), and Hambleton and Novick (1973). Since the criticisms may also be directed at Brennan's (1980) index, the responses that follow apply to both \( k^2 \) and \( \Phi(\lambda) \).

Harris (1972) argued that "although Livingston's reliability coefficient is (generally) larger than the conventional one, the standard error of measurement is the same, and consequently this larger coefficient does not imply a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value" (p. 29). Brennan and Kane (1977a, p. 286) felt that this criticism was primarily a matter of
Table 3
Evaluation of Squared-Error Loss Function Indices
(Both indices are rated equal in overall complexity)

<table>
<thead>
<tr>
<th>Index</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Administrations</th>
<th>Test Forms Assumption</th>
<th>Distribution Assumptions</th>
<th>Correction for Chance Agreement</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k^2(X,T)$</td>
<td>Ratio of true score to expected observed score mean squared deviations from the cutting score--error variance $\sigma^2(\hat{X})$ based on person x item interaction</td>
<td>Livingston (1972a, 1977)</td>
<td>1</td>
<td>Parallel</td>
<td>None</td>
<td>No</td>
<td>A generalization of the classical reliability coefficient, where the expected mean squared deviation about the cutting score is substituted for the variance about the mean score. When the cutting score equals the mean, $k^2$ reduces to the KR-20 coefficient. Advantages: Requires only one test administration, hand calculable. Disadvantages: Dependence of index values on cutting score, test length, and the estimated true or universe score variance and error variance complicates index interpretation.</td>
</tr>
<tr>
<td>$\phi(\lambda)$</td>
<td>Ratio of universe score to expected observed score mean squared deviations from the cutting score--error variance $\sigma^2(\lambda)$ based on item effect plus person x item interaction</td>
<td>Brennan (1980); Brennan and Kane (1977a, 1977b)</td>
<td>1</td>
<td>Randomly parallel</td>
<td>None</td>
<td>No</td>
<td>Measures the dependability of mastery-nomastery decisions based on the testing procedure. Derived from generalizability theory (Cronbach et al., 1972). When the cutting score ($\lambda$) equals the mean, $\phi(\lambda)$ is identical to the KR-21 coefficient (Brennan, 1977). Advantages: Same as $k^2$. Computer programs especially useful for more complex analysis of various designs and the estimation of variance components are available (e.g., Brennan, 1979; Erlich &amp; Shavelson, 1976). Estimation procedures are very flexible to accommodate different test structures and score types. Disadvantages: Same as $k^2$.</td>
</tr>
</tbody>
</table>

*Brennan (1980, pp. 204-213) also defined another index of dependability $\phi$ that is independent of any cutting score. It estimates the contribution of the testing procedure to the dependability of the decisions, over what would be expected on the basis of chance agreement. Unlike the above indices, the $\phi$ index is most useful as a domain score estimation statistic (See Table 4).*
“form rather than substance.” Livingston’s $k^2$ index and the standard error of measurement are different statistics that provide different information for different purposes. The $k^2$ index as well as the $\Phi(\lambda)$ index measures the consistency of scores in relation to the cutting score on repeated testings; the standard errors of measurement $\sigma(\delta)$, which is identical to that derived from classical reliability theory, and $\sigma(\Delta)$, respectively, measure the inconsistency or average imprecision of an individual’s scores independent of the cutting score on repeated testings. These standard errors could be used as domain score estimation statistics for setting up confidence intervals around individual scores. Those statistics do not diminish the meaningfulness and utility of $k^2$ and $\Phi(\lambda)$. As Livingston (1972b) pointed out in his reply to Harris, “... reliability is not a characteristic of a single score, but of a group of scores ... the larger criterion-referenced reliability coefficient does imply a more dependable overall determination of whether each true score falls above or below the criterion level, when this decision is to be made for every individual score in the distribution” (p. 31).

Shavelson et al. (1972) cited three problems with $k^2$: (a) it is “a function of the criterion as well as a function of individuals’ responses to items ... [consequently, it] is not directly related to the repeatability of the measure” (p. 134), (b) “this theory is applicable to situations where there may be no differences among true scores of individuals ... this seems unnecessary” (pp. 134–135), and (c) $k^2$ “does not estimate the ratio of true score variance over observed score variance alone; $k^2(X,T)$ cannot be used or interpreted like conventional reliabilities ... we suggest that this statistic be given some other name than ‘reliability’ ” (p. 135).

The first charge pertains to the sensitivity of the indices to the cutting score, specified previously as technical feature (d). Livingston (1972c) argued that “It is entirely appropriate that [$k^2$] ... should depend heavily on the difference between the group mean score and the criterion score ... [this] difference ... can be considered ‘true variance’ for criterion-referenced purposes” (p. 139); Brennan (1979) contended that “As an examinee’s observed score gets farther away from the cutting score, it becomes less likely that the examinee will be misclassified. ... Since ‘most’ examinee scores are concentrated around the mean, it seems intuitively reasonable that $\Phi(\lambda)$ should increase as the cutting score moves away from the mean” (p. 38). Both researchers perceive the characteristic as desirable rather than as a limitation. However, it is essential that the test maker report the cutting score upon which a particular index value is based along with other information so that the relationship is considered in interpreting the magnitude of $k^2$ and $\Phi(\lambda)$. Brennan (1980, p. 228) also suggests that it might even be better to report a “curve” of index values as a function of several possible cutting scores.

Furthermore, the Shavelson et al. (1972) assertion regarding the concept of repeatability does not appear to follow from the dependence of the index on the cutting score. Both $k^2$ and $\Phi(\lambda)$ are defined in terms of the repeatability of squared deviations of scores from the cutting score rather than from the mean score.

Shavelson et al.’s second charge focuses on the circumstances whereby the indices may yield positive values when there may be no true score variance, specified previously as technical feature (f). Livingston (1972a, p. 19) noted that even if a group of students randomly guessed at all items and the resultant set of scores produced a high index, that index would be meaningful; the test user could be quite confident that all of the students had true scores below the cutting score (Livingston, 1972c, p. 139). Brennan (1980)
offered an explanation for the situation where the error variance is positive and all students have identical high universe scores compared to the cutting score. He indicated that "In this case, $\Phi(\lambda)$ will be positive, reflecting the fact that it is relatively easy to determine correctly whether or not examinees are above the cut-off score" (p. 203).

The third charge relates to the appropriateness of the term "reliability" for describing and interpreting the indices. Livingston's (1972c) position on this issue is clear, "[$k^2$] deserves to be called a reliability coefficient because it represents the ratio of 'true' to 'observed' mean squared deviations from the criterion score, and the criterion-referenced correlation between alternate forms of the same test, and the squared criterion-referenced correlation between true scores and observed scores" (p. 139). Brennan and Kane (1977a, p. 285), however, have rejected the term "reliability coefficient" in favor of another term "dependability coefficient" for three reasons: (a) $\Phi(\lambda)$ is defined in terms of expected squared deviations rather than variances, (b) $\Phi(\lambda)$ entails definitions of error and parallel tests not usually associated with reliability coefficients, and (c) the magnitude of $\Phi(\lambda)$ depends upon a constant ($\lambda$), while the magnitude of the error variance is independent of $\lambda$. At the beginning of this paper a rationale was given for the term "agreement index" which is similar, in part, to Brennan and Kane's first reason.

Unlike the foregoing criticisms by Harris (1972) and Shavelson et al. (1972) that concentrate on technical characteristics of the indices, Hambleton and Novick's (1973) comments address the more fundamental issue of whether the squared-error loss approach to estimating "reliability" should be employed at all in criterion-referenced measurement. They stated emphatically that "Livingston misses the point for much of criterion-referenced testing ... the problem is one of deciding whether a student's true performance level is above or below some cutting score ... Livingston's choice of a loss function ... is wrong" (p. 168). Recently Hambleton et al. (1978, p. 17) admitted that their position is based on "conjecture" and only time will tell if they are correct. Given the numerous decision applications and interpretations of criterion-referenced test scores in practice at the outset of the 1980s, Hambleton and Novick's (1973) earlier comments would appear to have become outdated.

Advantages of $k^2(X,T)$ and $\Phi(\lambda)$. On the positive side, both indices have certain advantages that deserve serious attention by test makers at the classroom, district, and state levels. First, the estimates $k^2$ and $\Phi(\lambda)$ require only one test administration. Second, the indices have particular utility for criterion-referenced tests used by teachers to place students into instructional treatments. Once an initial assessment of mastery-nonmastery status has been completed based on the cutting score, the information typically of greatest value to teachers prior to instruction is the degree of mastery or nonmastery of each student along the score continuum. The detailed diagnosis of how much and what content a student does or does not know can assist the teacher in planning and assigning individualized prescriptions. Third, from a practical standpoint, the ease with which both indices could be hand calculated is comparable to that of the KR-20. The formula and computational examples for $\Phi(\lambda)$ prepared especially for practitioners by Brennan (1980, pp. 211–212) should contribute greatly to that ease. For test makers with computer resources, computer programs with extensive documentation are also available (e.g., Brennan, 1979; Erlich & Shavelson, 1976).

In addition to the above advantages, Brennan's (1980) approach has the advantage of flexibility in the analysis of different test structures corresponding to different test score uses. Brennan's application of generalizability theory to criterion-referenced measure-
ment makes it possible to estimate $\Phi(\lambda)$ and universe score variance and error variance components so that they reflect the various facets of interest in a given situation.

**Recommendations.** With regard to the interpretation of the indices, it is strongly recommended that the cutting score, test length and estimated error variance [either $\hat{\sigma}^2(\delta)$ or $\hat{\sigma}^2(\Delta)$] be reported to test users. This information should furnish an adequate base for index interpretation given the number and diversity of factors that affect the values of $k^2$ and $\Phi(\Delta)$. Brennan (1980) also urges test makers to report estimates of universe score variance and the squared deviation of the grand mean from the cutting score and, whenever possible, to estimate $\Phi(\lambda)$ in terms of variance components. This emphasis on variance components is consistent with the guidelines set forth in *Standards for Educational & Psychological Tests* (APA/AERA/NCME, 1974).

The choice between $k^2(X,T_x)$ and $\Phi(\lambda)$ is reduced to the simple decision of whether the assumption of classically parallel test forms or randomly parallel test forms, respectively, is appropriate for the test that was constructed. The advantages and disadvantages of the squared-error loss function should also be considered in that choice.

The guidelines presented previously for determining the appropriate loss function dealt with the type of decision to be made with the scores and the losses associated with the two types of decision error. There are some contexts where the threshold loss or squared-error loss function may be preferable and other contexts where neither may be preferable. Shepard (1980) suggested an "ideal index" that implies another loss function; it would reflect only false mastery and false nonmastery classification errors and it would weight those errors by degree. Unfortunately, at present such an index and loss function do not exist.

**DOMAIN SCORE ESTIMATION STATISTICS**

Unlike the agreement indices reviewed in the previous sections, the domain score estimation statistics are concerned generally with estimating the stability of an individual's score or proportion correct in the item domain independent of any mastery standard. Five statistics in this category are described in this section. All of them have meaning for individual decision making at the classroom or school level. Program evaluation decisions would require estimates of the average domain score and the standard error of that average. These estimates vary as a function of the multiple matrix sampling design that is used in the school district. A discussion of the different estimates corresponding to the different matrix sampling designs is beyond the scope of this review. Interested readers should consult Sirotnik's (1974) chapter which is oriented especially toward the practicing evaluator. More technical treatments of the topic include Lord and Novick's (1968, Chap. 11) chapter and Shoemaker's (1973) monograph.

The domain score estimation statistics are evaluated in Table 4. They are organized into two subcategories—individual specific and group specific. The characteristics of these subcategories will be examined first.

**Individual Specific Versus Group Specific**

The individual specific statistics are defined, computed and interpreted separately for each individual. They consist of two estimates of standard error that can be used to set up a confidence interval around each individual's observed proportion correct score. The group specific statistics represent averages of particular individual specific statistics over persons. Two of those statistics consist of estimates of the standard error of measurement
Table 4
Evaluation of Domain Score Estimation Statistics
(Statistics within each category are listed in order of increasing overall complexity)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Assessments</th>
<th>Test Forms</th>
<th>Assumption</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Specific</td>
<td>Standard error of a proportion—binomial formula for sampling error</td>
<td>Berk (1980b); Cochran (1963, pp. 49-59); Millman (1974a, 1974b)</td>
<td>1</td>
<td>Randomly parallel</td>
<td>Advantages: Provides unbiased estimate of error based on infinite or finite item domain; latter might require finite population correction (fpc) if item sample is large relative to item domain. Hand calculable and easily interpretable; tabulated values are also widely available (e.g., Marascuilo, 1971). Standard error remains the same for all tests of the same length irrespective of item format (Lord, 1957). Error can be used to set up a confidence interval or Uncertainty Band (UB) around an individual’s P or the P standard for mastery. Disadvantages: Computed using only ( \hat{p} ) and the number of items; estimate does not depend on the content, format, or statistical characteristics of the items. Therefore, estimates of error for a given test length will be lower for extreme values of ( \hat{p} ) and higher for moderate values. Hand computation is tedious because the error must be computed for each individual ( \hat{p} ).</td>
<td></td>
</tr>
<tr>
<td>S.E. ( \text{meas.}(x) )</td>
<td>Standard error of measurement—binomial formula for sampling error</td>
<td>Lord (1955, 1957, 1959)</td>
<td>1</td>
<td>Randomly parallel</td>
<td>Equivalent to the unbiased estimate of ( \sigma ) based on an infinite item domain (See Cochran, 1963, p. 51). Advantages: Same as ( \sigma ). Disadvantages: Same as ( \sigma ).</td>
<td></td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
<th>Source</th>
<th>No. of Administra-</th>
<th>Test Forms</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{a}$</td>
<td>Standard error of measurement--average of individual specific errors of measurement $\sigma (x_{meas})$</td>
<td>Lord and Novick (1968, pp. 60, 153-160) (See also Brennan, 1980)</td>
<td>1</td>
<td>Parallel</td>
<td>Conventional estimate of error derived from classical reliability theory. Equivalent to $\sigma (\theta)$ in the context of generalizability theory, which is based on person x item interaction. Advantages: Hand calculable and easily interpretable. Error can be used to set up a confidence interval around an individual's $P$. Disadvantages: Yields biased estimate of error (See Lord &amp; Novick, 1968, p. 192). Due to the fluctuation of the individual specific standard errors with different values of $P$, the average of those errors will tend to be lowest when there is a large number of extremely high and/or extremely low scores in the distribution (e.g., skewed or bimodal distribution); that low estimate can be very misleading.</td>
</tr>
<tr>
<td>$\sigma (A)$</td>
<td>Standard error of measurement--average of individual specific errors of measurement $S.E. (x_{meas})$</td>
<td>Brennan (1980)</td>
<td>1</td>
<td>Randomly parallel</td>
<td>Incorporates the effect due to item sampling along with the person x item interaction into the estimate. Derived from generalizability theory (Cronbach et al., 1972). Advantages: Same as $\sigma_{a}$. For the simple matrix sampling of persons and items, $\sigma (A)$ is equivalent to the square root of Lord and Novick's (1968, p. 251) unbiased estimate $\sigma (\theta)$; therefore, it can be computed easily from the mean and variance using formula (11.9.4). Disadvantages: Same as $\sigma_{a}$, except for first statement.</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Ratio of universe score variance to expected observed score variance--error variance $\sigma^{2} (A)$ based on item effect plus person x item interaction</td>
<td>Brennan (1980)</td>
<td>1</td>
<td>Randomly parallel</td>
<td>Measures the contribution of the testing procedure to the dependability of the decisions. Derived from generalizability theory (Cronbach et al., 1972). $\phi$ has an upper limit equal to the KR-20 coefficient and a lower limit equal to the KR-21 coefficient (Brennan, 1980, p. 213). Advantages: Hand calculable; computer programs are also available (e.g., Brennan, 1979). Estimation procedures are very flexible to accommodate different test structures and score uses. Disadvantages: Dependence of index values on test length and the estimated universe score and error variance complicates index interpretation. Specific application of the index in practice is unclear.</td>
</tr>
</tbody>
</table>
based on classically parallel and randomly parallel test forms' assumptions. The third statistic is an agreement index devised exclusively for domain-referenced interpretations.

Distribution assumptions. The four standard errors \( e_p, \sigma_{E_{\text{meas.}}}(x_a), \sigma_E, \) and \( \sigma(\Delta) \) can be estimated without any distribution assumptions. However, some assumptions are necessary when those errors are used to provide interval estimates of the domain score. A distribution assumption is essential for each standard error of measurement in this application so that a probability level can be assigned to the established confidence interval (e.g., 68%, 95%). Therefore, it is recommended that the binomial distribution be assumed for the two individual specific standard errors, the normal distribution be assumed for the traditional \( \sigma_E \), and either the binomial or compound binomial distribution may be assumed for \( \sigma(\Delta) \).

Point estimates of domain score. Although not included in Table 4, there are two point estimates of the domain score that have been proposed, \( \hat{P} \) and \( \hat{\gamma} \). While they do not estimate the degree of error in the measurement process as the other statistics do, a brief description of their properties seems warranted since the interval estimates or standard errors examined in the next section are based on point estimates.

The simplest point estimate is \( \hat{P} \), the observed proportion of items answered correctly by an individual. This estimate of the domain score \( P \) is unbiased and highly stable for full-length tests. For short subtests that are keyed to instructional objectives the estimate can be very unstable (Hambleton, Hutton, & Swaminathan, 1976).

The other point estimate is \( \hat{\gamma} \) (Hambleton et al., 1976). It involves a quasi-Bayesian approach that extends the binomial model underlying \( \sigma_{E_{\text{meas.}}}(x_a) \) (See Table 4) by incorporating direct (\( \hat{P} \)), collateral (group performance), and prior information (past test performance) into the estimate of \( P \). This approach yields more precise values than \( \hat{P} \) and several alternative classical and Bayesian estimates (Jackson, 1972; Lewis, Wang, & Novick, 1975; Novick, Lewis, & Jackson, 1973) even for subtests with as few as eight items (See Hambleton et al., 1976, for details). The trade-offs the test maker needs to consider to obtain this increase in precision over \( \hat{P} \) are the conceptual and computational complexity of the model and some distribution assumptions that may be rather tenuous in many practical applications.

Interval estimates of domain score (standard errors). Four of the five approaches in Table 4 consist of different types of standard error that can be used to set up a confidence interval or band within which the domain score \( P \) lies. All of the estimates of error measure the degree of imprecision or inconsistency of an individual's scores on repeated testings.

The two individual specific estimates \( \hat{e}_p \) and \( \sigma_{E_{\text{meas.}}}(x_a) \) are identical when the item domain is assumed to be infinite. The former is defined in the language of survey sampling theory where an individual's test performance is expressed as a "\( P \) value" or the percentage or proportion of items answered correctly (\( p \)); the latter is defined in the language of psychometric theory where an individual's test performance is expressed as a raw score \( (x_a) \). Both are based on the binomial formula for the "... sampling error of the number of white balls in a sample drawn at random from an urn containing a large number of white and black balls" (Lord, 1957, p. 511). The equivalence of the two unbiased estimates is readily apparent from their computational formulas:

\[
\hat{e}_p = \sqrt{\frac{pq}{n-1}}
\]  

(1)
and

\[ S.E_{\text{meas.}}(x_a) = \sqrt{\frac{x_a(n - x_a)}{n - 1}} \]  \hspace{1cm} (2)

The formulas recommended by Berk (1980b) and Millman (1974b) include the finite population correction (fpc) that is appropriate for finite item domains. It is shown below:

\[ \hat{c}_p = \sqrt{\frac{(N - n)(pq)}{n - 1}} \]  \hspace{1cm} (3)

In most situations, however, Formula 1 can be employed for infinite as well as finite domains as long as the projected or real domain size \((N)\) is large relative to the sample size \((n)\).

The application of this standard error of measurement in the context of an interval estimate can focus on either an individual's \(P\) or the \(P\) standard for mastery. Establishing a confidence interval or uncertainty band (Millman, 1974b) around the cutting score is one method for relating the standard error to mastery-nonmastery decisions. Individuals with scores that lie above the error region and those with scores that lie below may be classified as masters and nonmasters, respectively. If a 95% confidence interval is used, the probability of correct mastery and nonmastery decisions will be at least 95% (See Millman, 1974b, for some examples). This concept can also be linked to the issue of test length. The standard error could be specified a priori to guide the determination of the number of items to be randomly sampled from a finite item domain (See Berk, 1980b).

Probably the most distinctive feature of the individual specific estimate is that for any given value of \(P\), the error remains the same for all tests of the same length irrespective of item content, format, or statistical characteristics (Lord, 1957, 1959; Swineford, 1959). Unfortunately, due to the fact that the estimate is a function of only \(\hat{P}\) and the number of items, short subtests, say 10 items or less, measuring particular objectives will unpredictably yield sizable errors indicating the instability of \(\hat{P}\).

The two group specific estimates \(\hat{\sigma}_E\) and \(\hat{\sigma}(\Delta)\) are averages of the individual specific errors of measurement \(\hat{\sigma}(E_{ga})\) and \(S.E_{\text{meas.}}(x_a)\), respectively. These relationships have been proven by Lord and Novick (1968) and Brennan and Kane (1977b). Since the single average or "blanket estimate" of the standard error of measurement is applied to each individual score, the ideal level of precision is attained only when each individual has the same error. This seldom occurs in practice. Consequently, \(\hat{\sigma}_E\) and \(\hat{\sigma}(\Delta)\) should be considered as rough approximations of their corresponding individual specific errors.

These approximations are affected by test length and by the composition of the group tested or form of the score distribution. Based on the previous discussion, if the individual specific estimates of error are markedly influenced by the number of items, so are the averages. Therefore, the most stable estimates of group specific error are obtained with lengthy subtests or the total test. With regard to the score distribution, the best approximations seem to require homogeneous group performance. If the distribution contains a large number of extremely high and/or extremely low scores plus some moderate scores as in the case of a skewed or bimodal distribution, the estimate of error would be comparatively low. This result is misleading because that estimate, when applied to each of the scores, would tend to overestimate the standard errors for the
extreme scores and underestimate the errors for the moderate scores. One way to minimize and possibly eliminate this bias is to homogenize the data base. That can be accomplished, using Lord and Novick’s (1968, p. 155) suggestion, by partitioning a given score distribution (whether it is normal, bimodal, or another shape) into sections that identify subgroups of relatively similar ability and then computing $\hat{\sigma}_E$ or $\hat{\sigma}(\Delta)$ for each subgroup. This strategy would assure that each estimated subgroup standard error of measurement will be generally uniform for all points in that part of the distribution to which it is applied.

Selection of a standard error of measurement. The choice among the individual specific standard error of measurement S.E.\_\text{meas.} \( (x_a) \) and the two group specific standard errors $\hat{\sigma}_E$ and $\hat{\sigma}(\Delta)$ is contingent upon how “error” is defined in a particular test application. If the imprecision of an individual’s scores is measured over a hypothetical set of repeated testings with classically parallel test forms and the test is primarily used to rank order individuals, the appropriate error would be $\hat{\sigma}_E$. In this case the sampling of items does not introduce error into the estimation of an individual’s domain score relative to the average domain score for the group (Brennan & Kane, 1977a, p. 283). On the other hand, if the imprecision is measured over a hypothetical set of repeated testings with randomly parallel test forms and the test is used to determine an individual’s level of performance in the domain irrespective of the performance of other students, either S.E.\_\text{meas.} \( (x_a) \) or $\hat{\sigma}(\Delta)$ would be appropriate. In that case the sampling of items does introduce error into the estimator of an individual’s domain score. The final decision then rests with whether the individual specific or “blanket estimate” should be employed. Although the estimate of error for each individual $P$ has several advantages which were mentioned previously, it is also usually subject to excessively large sampling fluctuations (Lord, 1955). Therefore, when the standard error is applied to the cutting score, the individual specific estimate S.E.\_\text{meas.} \( (x_a) \) should be computed; otherwise, preference should be given to the estimate $\hat{\sigma}(\Delta)$ averaged over individuals.

Domain-referenced agreement index. The last group specific statistic, $\Phi$, measures the contribution of the testing procedure to the dependability of the decisions that are made. In contrast to the other domain score estimation statistics, $\Phi$ can be viewed as an agreement index corrected for chance. Developed within the framework of generalizability theory, the index has many of the features of its squared-error loss function counterpart $\Phi(\lambda)$. However, since it is not defined in relation to the cutting score ($\lambda$), it furnishes different information and answers a different question. While $\Phi$ is recommended as a general-purpose index of dependability for domain-referenced interpretations (Brennan, 1980, p. 204), the specific utility of the index in practice needs further clarification.

Recommendations. All of the foregoing statistics can be computed by hand from just one test administration. This advantage plus their ease of interpretation and application suggest that they could be used by teachers to estimate the precision of an individual’s scores on repeated testings with classically parallel or randomly parallel test forms. A teacher need only select one of the estimates of standard error [S.E.\_\text{meas.} \( (x_a) \), $\hat{\sigma}_E$, or $\hat{\sigma}(\Delta)$] or the agreement index $\Phi$.

There are two problems, however, that appear to restrict the usefulness of those statistics for individual decision making at the classroom or school level. First, as noted in the preceding descriptions of the statistics, a relatively large number of items per objective or the total number of items on the test must be used to obtain stable estimates.
This requirement is contrary to the characteristics of most teacher-made and professionally-developed criterion-referenced tests. Seldom are there more than 10 items per objective on these tests and the use of the total score for classroom level decisions about individuals is inconsistent with the intent and structure of the test. (Note: Total score decision making about an individual’s level of competence in the domain is justified for individual certification decisions in the context of minimum competency testing programs.) Second, the actual use of test scores for decision making typically requires some reference point or standard—relative or absolute. While the estimate of a student’s domain score supplies information about how much a student knows or his/her level of skill in a given content domain, it is difficult to make placement, formative, or summative decisions based on that information alone. Once a score is transformed into a normative score (e.g., percentile, stanine) for norm-referenced test score interpretations or a score is referenced to a mastery standard for criterion-referenced test score interpretations, individual decision making becomes considerably easier. The aforementioned statistics must be assessed in light of these practices. Consequently, the meaning of the various errors of measurement might be enhanced if they are related to the cutting score as in the application of the uncertainty band concept. The $ \Phi $ index has already received attention from this perspective in the form of the $ \Phi(\lambda) $ squared-error loss agreement index.

These issues are not generalizable to estimates of the average domain score. The precision and utility of “averages” in the evaluation of competing instructional programs using appropriate units of analysis (individual, class, school, or district) have been demonstrated (See, for example, Sirotnik, 1974). When criterion-referenced tests are employed in program evaluation, the use of various matrix sampling designs can provide more dependable estimates of average domain scores without any loss in the dependability of estimates of individual domain scores (See Brennan, 1980, p. 217).

CONCLUSIONS

The first choice of a “reliability” category for a particular application is based on the test forms assumption (classically parallel or randomly parallel), whether or not a cutting score is set, and, most importantly, the intended test score interpretation, type of decision, and seriousness of losses associated with the decision errors. These latter characteristics are the primary determinants of the appropriate loss function. When mastery-nonmastery classifications are to be made and the losses related to classification errors are assumed to be equally serious regardless of how far they are from the cutoff score, the threshold loss function should be used; when degrees of mastery-nonmastery along the score continuum are of interest and the losses related to misclassified students who are far from the cutting score are assumed to be the most serious, the squared-error loss function should be used. Other characteristics may indicate that neither of those loss functions are appropriate. When level of competency in the domain (with or without a mastery-nonmastery distinction) is to be assessed, domain score estimation statistics should be considered.

The evaluation of the “reliability” indices and approaches within each category (See Tables 2, 3, and 4) suggested the following conclusions:

1. Threshold loss function indices—Hambleton and Novick’s (1973) two-administration method for estimating $ p_o $ is the easiest to understand, to compute, and to interpret; it appears to have the greatest utility for classroom test construction and
decision making. In addition to this method, Swaminathan et al.'s (1974) two-administrations method for estimating $\kappa$ and Huynh's (1976) single-administration method for estimating $p_0$ and $\kappa$ merit the attention of test publishers and test makers at the district and state levels.

2. Squared-error loss function indices—Livingston's (1972a) and Brennan's (1980) indices provide meaningful information about the consistency of scores in relation to the cutting score, particularly for placement tests. The choice between the indices depends upon the assumption about test forms.

3. Domain score estimation statistics—The various estimates seem to have limited usefulness for classroom decisions unless they are linked to the cutting score, as exemplified by Millman's (1974b) approach. Estimates of the average domain score appear to have greater utility in the context of program evaluation decisions in which case a standard error statistic must be selected to reflect both the imprecision of measurement and sampling error of subjects.

REFERENCES


ANTONAK, R. F. A computer program to compute measures of response agreement for nominal scale data obtained from two judges. *Behavior Research Methods & Instrumentation*, 1977, **9**, 533.


BERK, R. A. A comparison of six content domain specification strategies for criterion-referenced tests. *Educational Technology*, 1980, **20** (9), 49–52. (a)


LIVINGSTON, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, 9, 31. (b)

LIVINGSTON, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion referenced testing: Comments on reliability." *Journal of Educational Measurement*, 1972, 9, 139–140. (c)


LORD, F. M. Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 1957, 17, 510–521.


REID, J. B., & ROBERTS, D. M. *A Monte Carlo comparison of phi and kappa as measures of


SWINEFORD, F. Note on “Tests of the same length do have the same standard error of measurement.” Educational and Psychological Measurement, 1959, 19, 241–242.


AUTHOR

BERK, RONALD A. Address: Division of Education, The Johns Hopkins University, Baltimore, MD 21218. Title: Associate Professor. Degrees: B.A. American University; M.Ed., Ph.D. University of Maryland. Specialization: Psychometric Theory, Research Methodology.