2. RELIABILITY/PRECISION AND ERRORS OF MEASUREMENT

BACKGROUND

A test, broadly defined, is a set of tasks or stimuli designed to elicit responses that provide a sample of an examinee's behavior or performance in a specified domain. Coupled with the test is a scoring procedure that enables the scorer to evaluate the behavior or work samples and generate a score. In interpreting and using test scores, it is important to have some indication of their reliability.

The term *reliability* has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency). To maintain a link to the traditional notions of reliability while avoiding the ambiguity inherent in using a single, familiar term to refer to a wide range of concepts and indices, we use the term *reliability/precision* to denote the more general notion of consistency of the scores across instances of the testing procedure, and the term *reliability coefficient* to refer to the reliability coefficients of classical test theory.

The reliability/precision of measurement is always important. However, the need for precision increases as the consequences of decisions and interpretations grow in importance. If a test score leads to a decision that is not easily reversed, such as rejection or admission of a candidate to a professional school, or a score-based clinical judgment (e.g., in a legal context) that a serious cognitive injury was sustained, a higher degree of reliability/precision is warranted. If a decision can and will be corroborated by information from other sources or if an erroneous initial decision can be easily corrected, scores with more modest reliability/precision may suffice.

Interpretations of test scores generally depend on assumptions that individuals and groups exhibit some degree of consistency in their scores across independent administrations of the testing procedure. However, different samples of performance from the same person are rarely identical. An individual's performances, products, and responses to sets of tasks or test questions vary in quality or character from one sample of tasks to another and from one occasion to another, even under strictly controlled conditions. Different raters may award different scores to a specific performance. All of these sources of variation are reflected in the examinees' scores, which will vary across instances of a measurement procedure.

The reliability/precision of the scores depends on how much the scores vary across replications of the testing procedure, and analyses of reliability/precision depend on the kinds of variability allowed in the testing procedure (e.g., over tasks, contexts, raters) and the proposed interpretation of the test scores. For example, if the interpretation of the scores assumes that the construct being assessed does not vary over occasions, the variability over occasions is a potential source of measurement error. If the test tasks vary over alternate forms of the test, and the observed performances are treated as a sample from a domain of similar tasks, the random variability in scores from one form to another would be considered error. If raters are used to assign scores to responses, the variability in scores over qualified raters is a source of error. Variations in a test taker's scores that are not consistent with the definition of the construct being assessed are attributed to errors of measurement.

A very basic way to evaluate the consistency of scores involves an analysis of the variation in each test taker's scores across replications of the testing procedure. The test is administered and then, after a brief period during which the examinee's standing on the variable being measured would not be expected to change, the test (or a distinct but equivalent form of the test) is administered a second time; it is assumed that the first administration has no influence on the second administration. Given that the attribute being measured is assumed to remain the same for each test taker over the two administrations and that the test administrations are independent of each other, more variation across the two administrations indicates more error in the test scores and therefore lower reliability/precision.

The impact of such measurement errors can be summarized in a number of ways, but typically, in educational and psychological measurement, it is conceptualized in terms of the standard deviation in the scores for a person over replications of the testing procedure. In most testing contexts, it is not possible to replicate the testing procedure repeatedly, and therefore it is not possible to estimate the standard error for each person's score via repeated measurement. Instead, using model-based assumptions, the average error of measurement is estimated over some population, and this average is referred to as the *standard error of measurement* (SEM). The SEM is an indicator of a lack of consistency in the scores generated by the testing procedure for some population. A relatively large SEM indicates relatively low reliability/precision. The *conditional standard error of measurement* for a score level is the standard error of measurement at that score level.

To say that a score includes error implies that there is a hypothetical error-free value that characterizes the variable being assessed. In classical test theory this error-free value is referred to as the person's *true score* for the test procedure. It is conceptualized as the hypothetical average score over an infinite set of replications of the testing procedure. In statistical terms, a person's true score is an unknown parameter, or constant, and the observed score for the person is a random variable that fluctuates around the true score for the person.

Generalizability theory provides a different framework for estimating reliability/precision. While classical test theory assumes a single distribution for the errors in a test taker's scores, generalizability theory seeks to evaluate the contributions of different sources of error (e.g., items, occasions, raters) to the overall error. The *universe score* for a person is defined as the expected value over a universe of all possible replications of the testing procedure for the test taker. The universe score of generalizability theory plays a role that is similar to the role of true scores in classical test theory.

Item response theory (IRT) addresses the basic issue of reliability/precision using information functions, which indicate the precision with which observed task/item performances can be used to estimate the value of a latent trait for each test taker. Using IRT, indices analogous to traditional reliability coefficients can be estimated from the item information functions and distributions of the latent trait in some population.

In practice, the reliability/precision of the scores is typically evaluated in terms of various coefficients, including reliability coefficients, generalizability coefficients, and IRT information functions, depending on the focus of the analysis and the measurement model being used. The coefficients tend to have high values when the variability associated with the error is small compared with the observed variation in the scores (or score differences) to be estimated.

Implications for Validity

Although reliability/precision is discussed here as an independent characteristic of test scores, it should be recognized that the level of reliability/precision of scores has implications for validity. Reliability/precision of data ultimately bears on the generalizability or dependability of the scores and/or the consistency of classifications of individuals derived from the scores. To the extent that scores are not consistent across replications of the testing procedure (i.e., to the extent that they reflect random errors of measurement), their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited.

Specifications for Replications of the Testing Procedure

As indicated earlier, the general notion of reliability/precision is defined in terms of consistency over replications of the testing procedure. Reliability/precision is high if the scores for each person are consistent over replications of the testing procedure and is low if the scores are not consistent over replications. Therefore, in evaluating reliability/precision, it is important to be clear about what constitutes a replication of the testing procedure.

Replications involve independent administrations of the testing procedure, such that the attribute being measured would not be expected to change. For example, in assessing an attribute that is not expected to change over an extended period of time (e.g., in measuring a trait), scores generated on two successive days (using different test forms if appropriate) would be considered replications. For a state variable (e.g., mood or hunger), where fairly rapid changes are common, scores generated on two successive days would not be considered replications; the scores obtained on each occasion would be interpreted in terms of the value of the state variable on that occasion. For many tests of knowledge or skill, the administration of alternate forms of a test with different samples of items would be considered replications of the test; for survey instruments and some personality measures, it is expected that the same questions will be used every time the test is administered, and any substantial change in wording would constitute a different test form.

Standardized tests present the same or very similar test materials to all test takers, maintain close adherence to stipulated procedures for test administration, and employ prescribed scoring rules that can be applied with a high degree of consistency. Administering the same questions or commonly scaled questions to all test takers under the same conditions promotes fairness and facilitates comparisons of scores across individuals. Conditions of observation that are fixed or standardized for the testing procedure remain the same across replications. However, some aspects of any standardized testing procedure will be allowed to vary. The time and place of testing, as well as the persons administering the test, are generally allowed to vary to some extent. The particular tasks included in the test may be allowed to vary (as samples from a common content domain), and the persons who score the results can vary over some set of qualified scorers.

Alternate forms (or *parallel forms*) of a standardized test are designed to have the same general distribution of content and item formats (as described, for example, in detailed test specifications), the same administrative procedures, and at least approximately the same score means and standard deviations in some specified population or populations. Alternate forms of a test are considered interchangeable, in the sense that they are built to the same specifications, and are interpreted as measures of the same construct.

In classical test theory, strictly parallel tests are assumed to measure the same construct and to yield scores that have the same means and standard deviations in the populations of interest and have the same correlations with all other variables. A classical reliability coefficient is defined in terms of the correlation between scores from strictly parallel forms of the test, but it is estimated in terms of the correlation between alternate forms of the test that may not quite be strictly parallel.

Different approaches to the estimation of reliability/precision can be implemented to fit different data-collection designs and different interpretations and uses of scores. In some cases, it may be feasible to estimate the variability over replications directly (e.g., by having a number of qualified raters evaluate a sample of test performances for each test taker). In other cases, it may be necessary to use less direct estimates of the reliability coefficient. For example, internal-consistency estimates of reliability (e.g., split halves coefficient, KR–20, coefficient alpha) use the observed extent of agreement between different parts of one test to estimate the reliability associated with form-to-form variability. For the split-halves method, scores on two more-or-less parallel halves of the test (e.g., odd-numbered items and even-numbered items) are correlated, and the resulting half-test reliability coefficient is statistically adjusted to estimate reliability for the full-length test. However, when a test is designed to reflect rate of work, internal-consistency estimates of reliability (particularly by the odd-even method) are likely to yield inflated estimates of reliability for highly speeded tests.

In some cases, it may be reasonable to assume that a potential source of variability is likely to be negligible or that the user will be able to infer adequate reliability from other types of evidence. For example, if test scores are used mainly to predict some criterion scores and the test does an acceptable job in predicting the criterion, it can be inferred that the test scores are reliable/precise enough for their intended use.

The definition of what constitutes a standardized test or measurement procedure has broadened significantly over the last few decades. Various kinds of performance assessments, simulations, and portfolio-based assessments have been developed to provide measures of constructs that might otherwise be difficult to assess. Each step toward greater flexibility in the assessment procedures enlarges the scope of the variations allowed in replications of the testing procedure, and therefore tends to increase the measurement error. However, some of these sacrifices in reliability/precision may reduce construct irrelevance or construct underrepresentation and thereby improve the validity of the intended interpretations of the scores. For example, performance assessments that depend on ratings of extended responses tend to have lower reliability than more structured assessments (e.g., multiple-choice or short-answer tests), but they can sometimes provide more direct measures of the attribute of interest.

Random errors of measurement are viewed as unpredictable fluctuations in scores. They are conceptually distinguished from *systematic errors*, which may also affect the performances of individuals or groups but in a consistent rather than a random manner. For example, an incorrect answer key would contribute systematic error, as would differences in the difficulty of test forms that have not been adequately equated or linked; examinees who take one form may receive higher scores on average than if they had taken the other form. Such systematic errors would not generally be included in the standard error of measurement, and they are not regarded as contributing to a lack of reliability/precision. Rather, systematic errors constitute construct-irrelevant factors that reduce validity but not reliability/precision.

Important sources of random error may be grouped in two broad categories: those rooted within the test takers and those external to them. Fluctuations in the level of an examinee's motivation, interest, or attention and the inconsistent application of skills are clearly internal sources that may lead to random error. Variations in testing conditions (e.g., time of day, level of distractions) and variations in scoring due to scorer subjectivity are examples of external sources that may lead to random error. The importance of any particular source of variation depends on the specific conditions under which the measures are taken, how performances are scored, and the interpretations derived from the scores.

Some changes in scores from one occasion to another are not regarded as error (random or systematic), because they result, in part, from changes in the construct being measured (e.g., due to learning or maturation that has occurred between the initial and final measures). In such cases, the changes in performance would constitute the phenomenon of interest and would not be considered errors of measurement.

Measurement error reduces the usefulness of test scores. It limits the extent to which test results can be generalized beyond the particulars of a given replication of the testing procedure. It reduces the confidence that can be placed in the results from any single measurement and therefore the reliability/precision of the scores. Because random measurement errors are unpredictable, they cannot be removed from observed scores. However, their aggregate magnitude can be summarized in several ways, as discussed below, and they can be controlled to some extent (e.g., by standardization or by averaging over multiple scores).

The standard error of measurement, as such, provides an indication of the expected level of random error over score points and replications for a specific population. In many cases, it is useful to have estimates of the standard errors for individual examinees (or for examinees with scores in certain score ranges). These conditional standard errors are difficult to estimate directly, but can be estimated indirectly. For example, the test information functions based on IRT models can be used to estimate standard errors for different values of a latent ability parameter and/or for different observed scores. In using any of these model-based estimates of conditional standard errors, it is important that the model assumptions be consistent with the data.

Evaluating Reliability/Precision

The ideal approach to the evaluation of reliability/precision would require many independent replications of the testing procedure on a large sample of test takers. The range of differences allowed in replications of the testing procedure and the proposed interpretation of the scores provide a framework for investigating reliability/precision.

For most testing programs, scores are expected to generalize over alternate forms of the test, occasions (within some period), testing contexts, and raters (if judgment is required in scoring). To the extent that the impact of any of these sources of variability is expected to be substantial, the variability should be estimated in some way. It is not necessary that the different sources of variance be estimated separately. The overall reliability/precision, given error variance due to the sampling of forms, occasions, and raters, can be estimated through a test-retest study involving different forms administered on different occasions and scored by different raters.

The interpretation of reliability/precision analyses depends on the population being tested. For example, reliability or generalizability coefficients derived from scores of a nationally representative sample may differ significantly from those obtained from a more homogeneous sample drawn from one gender, one ethnic group, or one community. Therefore, to the extent feasible (i.e., if sample sizes are large enough), reliability/precision should be estimated separately for all relevant subgroups (e.g., defined in terms of race/ethnicity, gender, language proficiency) in the population. (Also see chap. 3, “Fairness in Testing.”)

Reliability/Generalizability Coefficients

In classical test theory, the consistency of test scores is evaluated mainly in terms of reliability coefficients, defined in terms of the correlation between scores derived from replications of the testing procedure on a sample of test takers. Three broad categories of reliability coefficients are recognized: (a) coefficients derived from the administration of alternate forms in independent testing sessions (*alternate-form coefficients*); (b) coefficients obtained by administration of the same form on separate occasions (*test-retest coefficients*); and (c) coefficients based on the relationships/interactions among scores derived from individual items or subsets of the items within a test, all data accruing from a single administration (*internal-consistency coefficients*). In addition, where test scoring involves a high level of judgment, indices of scorer consistency are commonly obtained. In formal treatments of classical test theory, reliability can be defined as the ratio of true-score variance to observed score variance, but it is estimated in terms of reliability coefficients of the kinds mentioned above.

In generalizability theory, these different reliability analyses are treated as special cases of a more general framework for estimating error variance in terms of the variance components associated with different sources of error. A *generalizability coefficient* is defined as the ratio of universe score variance to observed score variance. Unlike traditional approaches to the study of reliability, generalizability theory encourages the researcher to specify and estimate components of true score variance, error score variance, and observed score variance, and to calculate coefficients based on these estimates. Estimation is typically accomplished by the application of analysis-of-variance techniques. The separate numerical estimates of the components of variance (e.g., variance components for items, occasions, and raters, and for the interactions among these potential sources of error) can be used to evaluate the contribution of each source of error to the overall measurement error; the variance-component estimates can be helpful in identifying an effective strategy for controlling overall error variance.

Different reliability (and generalizability) coefficients may appear to be interchangeable, but the different coefficients convey different information. A coefficient may encompass one or more sources of error. For example, a coefficient may reflect error due to scorer inconsistencies but not reflect the variation over an examinee's performances or products. A coefficient may reflect only the internal consistency of item responses within an instrument and fail to reflect measurement error associated with day-to-day changes in examinee performance.

It should not be inferred, however, that alternate-form or test-retest coefficients based on test administrations several days or weeks apart are always preferable to internal-consistency coefficients. In cases where we can assume that scores are not likely to change, based on past experience and/or theoretical considerations, it may be reasonable to assume invariance over occasions (without conducting a test-retest study). Another limitation of test-retest coefficients is that, when the same form of the test is used, the correlation between the first and second scores could be inflated by the test taker's recall of initial responses.

The test information function, an important result of IRT, summarizes how well the test discriminates among individuals at various levels of ability on the trait being assessed. Under the IRT conceptualization for dichotomously scored items, the *item characteristic curve* or *item response function* is used as a model to represent the increasing proportion of correct responses to an item at increasing levels of the ability or trait being measured. Given appropriate data, the parameters of the characteristic curve for each item in a test can be estimated. The test information function can then be calculated from the parameter estimates for the set of items in the test and can be used to derive coefficients with interpretations similar to reliability coefficients.

The information function may be viewed as a mathematical statement of the precision of measurement at each level of the given trait. The IRT information function is based on the results obtained on a specific occasion or in a specific context, and therefore it does not provide an indication of generalizability over occasions or contexts.

Coefficients (e.g., reliability, generalizability, and IRT-based coefficients) have two major advantages over standard errors. First, as indicated above, they can be used to estimate standard errors (overall and/or conditional) in cases where it would not be possible to do so directly. Second, coefficients (e.g., reliability and generalizability coefficients), which are defined in terms of ratios of variances for scores on the same scale, are invariant over linear transformations of the score scale and can be useful in comparing different testing procedures based on different scales. However, such comparisons are rarely straightforward, because they can depend on the variability of the groups on which the coefficients are based, the techniques used to obtain the coefficients, the sources of error reflected in the coefficients, and the lengths and contents of the instruments being compared.

Factors Affecting Reliability/Precision

A number of factors can have significant effects on reliability/precision, and in some cases, these factors can lead to misinterpretations of the results, if not taken into account.

First, any evaluation of reliability/precision applies to a particular assessment procedure and is likely to change if the procedure is changed in any substantial way. In general, if the assessment is shortened (e.g., by decreasing the number of items or tasks), the reliability is likely to decrease; and if the assessment is lengthened with comparable tasks or items, the reliability is likely to increase. In fact, lengthening the assessment, and thereby increasing the size of the sample of tasks/items (or raters or occasions) being employed, is an effective and commonly used method for improving reliability/precision.

Second, if the variability associated with raters is estimated for a select group of raters who have been especially well trained (and were perhaps involved in the development of the procedures), but raters are not as well trained in some operational contexts, the error associated with rater variability in these operational settings may be much higher than is indicated by the reported interrater reliability coefficients. Similarly, if raters are still refining their performance in the early days of an extended scoring window, the error associated with rater variability may be greater for examinees testing early in the window than for examinees who test later.

Reliability/precision can also depend on the population for which the procedure is being used. In particular, if variability in the construct of interest in the population for which scores are being generated is substantially different from what it is in the population for which reliability/precision was evaluated, the reliability/precision can be quite different in the two populations. When the variability in the construct being measured is low, reliability and generalizability coefficients tend to be small, and when the variability in the construct being measured is higher, the coefficients tend to be larger. Standard errors of measurement are less dependent than reliability and generalizability coefficients on the variability in the sample of test takers.

In addition, reliability/precision can vary from one population to another, even if the variability in the construct of interest in the two populations is the same. The reliability can vary from one population to another because particular sources of error (rater effects, familiarity with formats and instructions, etc.) have more impact in one population than they do in the other. In general, if any aspects of the assessment procedures or the population being assessed are changed in an operational setting, the reliability/precision may change.

Standard Errors of Measurement

The standard error of measurement can be used to generate confidence intervals around reported scores. It is therefore generally more informative than a reliability or generalizability coefficient, once a measurement procedure has been adopted and the interpretation of scores has become the user's primary concern.

Estimates of the standard errors at different score levels (that is, conditional standard errors) are usually a valuable supplement to the single statistic for all score levels combined. Conditional standard errors of measurement can be much more informative than a single average standard error for a population. If decisions are based on test scores and these decisions are concentrated in one area or a few areas of the score scale, then the conditional errors in those areas are of special interest.

Like reliability and generalizability coefficients, standard errors may reflect variation from many sources of error or only a few. A more comprehensive standard error (i.e., one that includes the most relevant sources of error, given the definition of the testing procedure and the proposed interpretation) tends to be more informative than a less comprehensive standard error. However, practical constraints often preclude the kinds of studies that would yield information on all potential sources of error, and in such cases, it is most informative to evaluate the sources of error that are likely to have the greatest impact.

Interpretations of test scores may be broadly categorized as *relative* or *absolute.* Relative interpretations convey the standing of an individual or group within a reference population. Absolute interpretations relate the status of an individual or group to defined performance standards. The standard error is not the same for the two types of interpretations. Any source of error that is the same for all individuals does not contribute to the relative error but may contribute to the absolute error.

Traditional norm-referenced reliability coefficients were developed to evaluate the precision with which test scores estimate the relative standing of examinees on some scale, and they evaluate reliability/precision in terms of the ratio of true-score variance to observed-score variance. As the range of uses of test scores has expanded and the contexts of use have been extended (e.g., diagnostic categorization, the evaluation of educational programs), the range of indices that are used to evaluate reliability/precision has also grown to include indices for various kinds of change scores and difference scores, indices of decision consistency, and indices appropriate for evaluating the precision of group means.

Some indices of precision, especially standard errors and conditional standard errors, also depend on the scale in which they are reported. An index stated in terms of raw scores or the trait-level estimates of IRT may convey a very different perception of the error if restated in terms of scale scores. For example, for the raw-score scale, the conditional standard error may appear to be high at one score level and low at another, but when the conditional standard errors are restated in units of scale scores, quite different trends in comparative precision may emerge.

Decision Consistency

Where the purpose of measurement is classification, some measurement errors are more serious than others. Test takers who are far above or far below the cut score established for pass/fail or for eligibility for a special program can have considerable error in their observed scores without any effect on their classification decisions. Errors of measurement for examinees whose true scores are close to the cut score are more likely to lead to classification errors. The choice of techniques used to quantify reliability/precision should take these circumstances into account. This can be done by reporting the conditional standard error in the vicinity of the cut score or the decision-consistency/accuracy indices (e.g., percentage of correct decisions, Cohen's kappa), which vary as functions of both score reliability/precision and the location of the cut score.

Decision consistency refers to the extent to which the observed classifications of examinees would be the same across replications of the testing procedure. *Decision accuracy* refers to the extent to which observed classifications of examinees based on the results of a single replication would agree with their true classification status. Statistical methods are available to calculate indices for both decision consistency and decision accuracy. These methods evaluate the consistency or accuracy of classifications rather than the consistency in scores per se. Note that the degree of consistency or agreement in examinee classification is specific to the cut score employed and its location within the score distribution.

Reliability/Precision of Group Means

Estimates of mean (or average) scores of groups (or proportions in certain categories) involve sources of error that are different from those that operate at the individual level. Such estimates are often used as measures of program effectiveness (and, under some educational accountability systems, may be used to evaluate the effectiveness of schools and teachers).

In evaluating group performance by estimating the mean performance or mean improvement in performance for samples from the group, the variation due to the sampling of persons can be a major source of error, especially if the sample sizes are small. To the extent that different samples from the group of interest (e.g., all students who use certain educational materials) yield different results, conclusions about the expected outcome over all students in the group (including those who might join the group in the future) are uncertain. For large samples, the variability due to the sampling of persons in the estimates of the group means may be quite small. However, in cases where the samples of persons are not very large (e.g., in evaluating the mean achievement of students in a single classroom or the average expressed satisfaction of samples of clients in a clinical program), the error associated with the sampling of persons may be a major component of overall error. It can be a significant source of error in inferences about programs even if there is a high degree of precision in individual test scores.

Standard errors for individual scores are not appropriate measures of the precision of group averages. A more appropriate statistic is the standard error for the estimates of the group means.

Documenting Reliability/Precision

Typically, developers and distributors of tests have primary responsibility for obtaining and reporting evidence for reliability/precision (e.g., appropriate standard errors, reliability or generalizability coefficients, or test information functions). The test user must have such data to make an informed choice among alternative measurement approaches and will generally be unable to conduct adequate reliability/precision studies prior to operational use of an instrument.

In some instances, however, local users of a test or assessment procedure must accept at least partial responsibility for documenting the precision of measurement. This obligation holds when one of the primary purposes of measurement is to classify students using locally developed performance standards, or to rank examinees within the local population. It also holds when users must rely on local scorers who are trained to use the scoring rubrics provided by the test developer. In such settings, local factors may materially affect the magnitude of error variance and observed score variance. Therefore, the reliability/precision of scores may differ appreciably from that reported by the developer.

Reported evaluations of reliability/precision should identify the potential sources of error for the testing program, given the proposed uses of the scores. These potential sources of error can then be evaluated in terms of previously reported research, new empirical studies, or analyses of the reasons for assuming that a potential source of error is likely to be negligible and therefore can be ignored.

The reporting of indices of reliability/precision alone—with little detail regarding the methods used to estimate the indices reported, the nature of the group from which the data were derived, and the conditions under which the data were obtained—constitutes inadequate documentation. General statements to the effect that a test is “reliable” or that it is “sufficiently reliable to permit interpretations of individual scores” are rarely, if ever, acceptable. It is the user who must take responsibility for determining whether scores are sufficiently trustworthy to justify anticipated uses and interpretations for particular uses. Nevertheless, test constructors and publishers are obligated to provide sufficient data to make informed judgments possible.

If scores are to be used for classification, indices of decision consistency are useful in addition to estimates of the reliability/precision of the scores. If group means are likely to play a substantial role in the use of the scores, the reliability/precision of these mean scores should be reported.

As the foregoing comments emphasize, there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations, nor is the test developer limited to a single approach for any instrument. The choice of estimation techniques and the minimum acceptable level for any index remain a matter of professional judgment.

STANDARDS FOR RELIABILITY/PRECISION

The standards in this chapter begin with an overarching standard (numbered 2.0), which is designed to convey the central intent or primary focus of the chapter. The overarching standard may also be viewed as the guiding principle of the chapter, and is applicable to all tests and test users. All subsequent standards have been separated into eight thematic clusters labeled as follows:

1. Specifications for Replications of the Testing Procedure
2. Evaluating Reliability/Precision
3. Reliability/Generalizability Coefficients
4. Factors Affecting Reliability/Precision
5. Standard Errors of Measurement
6. Decision Consistency
7. Reliability/Precision of Group Means
8. Documenting Reliability/Precision

Standard 2.0

Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.

Comment: The form of the evidence (reliability or generalizability coefficient, information function, conditional standard error, index of decision consistency) for reliability/precision should be appropriate for the intended uses of the scores, the population involved, and the psychometric models used to derive the scores. A higher degree of reliability/precision is required for score uses that have more significant consequences for test takers. Conversely, a lower degree may be acceptable where a decision based on the test score is reversible or dependent on corroboration from other sources of information.

Cluster 1. Specifications for Replications of the Testing Procedure

Standard 2.1

The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation.

Comment: For any testing program, some aspects of the testing procedure (e.g., time limits and availability of resources such as books, calculators, and computers) are likely to be fixed, and some aspects will be allowed to vary from one administration to another (e.g., specific tasks or stimuli, testing contexts, raters, and, possibly, occasions). Any test administration that maintains fixed conditions and involves acceptable samples of the conditions that are allowed to vary would be considered a legitimate replication of the testing procedure. As a first step in evaluating the reliability/precision of the scores obtained with a testing procedure, it is important to identify the range of conditions of various kinds that are allowed to vary, and over which scores are to be generalized.

Standard 2.2

The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores.

Comment: The evidence for reliability/precision should be consistent with the design of the testing procedures and with the proposed interpretations for use of the test scores. For example, if the test can be taken on any of a range of occasions, and the interpretation presumes that the scores are invariant over these occasions, then any variability in scores over these occasions is a potential source of error. If the tasks or stimuli are allowed to vary over alternate forms of the test, and the observed performances are treated as a sample from a domain of similar tasks, the variability in scores from one form to another would be considered error. If raters are used to assign scores to responses, the variability in scores over qualified raters is a source of error. Different sources of error can be evaluated in a single coefficient or standard error, or they can be evaluated separately, but they should all be addressed in some way. Reports of reliability/precision should specify the potential sources of error included in the analyses.

Cluster 2. Evaluating Reliability/Precision

Standard 2.3

For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

Comment: It is not sufficient to report estimates of reliabilities and standard errors of measurement only for total scores when subscores are also interpreted. The form-to-form and day-to-day consistency of total scores on a test may be acceptably high, yet subscores may have unacceptably low reliability, depending on how they are defined and used. Users should be supplied with reliability data for all scores to be interpreted, and these data should be detailed enough to enable the users to judge whether the scores are precise enough for the intended interpretations for use. Composites formed from selected subtests within a test battery are frequently proposed for predictive and diagnostic purposes. Users need information about the reliability of such composites.

Standard 2.4

When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/precision data, including standard errors, should be provided for such differences.

Comment: Observed score differences are used for a variety of purposes. Achievement gains are frequently of interest for groups as well as individuals. In some cases, the reliability/precision of change scores can be much lower than the reliabilities of the separate scores involved. Differences between verbal and performance scores on tests of intelligence and scholastic ability are often employed in the diagnosis of cognitive impairment and learning problems. Psychodiagnostic inferences are frequently drawn from the differences between subtest scores. Aptitude and achievement batteries, interest inventories, and personality assessments are commonly used to identify and quantify the relative strengths and weaknesses, or the pattern of trait levels, of a test taker. When the interpretation of test scores centers on the peaks and valleys in the examinee's test score profile, the reliability of score differences is critical.

Standard 2.5

Reliability estimation procedures should be consistent with the structure of the test.

Comment: A single total score can be computed on tests that are multidimensional. The total score on a test that is substantially multidimensional should be treated as a composite score. If an internal-consistency estimate of total score reliability is obtained by the split-halves procedure, the halves should be comparable in content and statistical characteristics.

In adaptive testing procedures, the set of tasks included in the test and the sequencing of tasks are tailored to the test taker, using model-based algorithms. In this context, reliability/precision can be estimated using simulations based on the model. For adaptive testing, model-based conditional standard errors may be particularly useful and appropriate in evaluating the technical adequacy of the procedure.

Cluster 3. Reliability/Generalizability Coefficients

Standard 2.6

A reliability or generalizability coefficient (or standard error) that addresses one kind of variability should not be interpreted as interchangeable with indices that address other kinds of variability, unless their definitions of measurement error can be considered equivalent.

Comment: Internal-consistency, alternate-form, and test-retest coefficients should not be considered equivalent, as each incorporates a unique definition of measurement error. Error variances derived via item response theory are generally not equivalent to error variances estimated via other approaches. Test developers should state the sources of error that are reflected in, and those that are ignored by, the reported reliability or generalizability coefficients.

Standard 2.7

When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products.

Comment: Task-to-task variations in the quality of an examinee's performance and rater-to-rater inconsistencies in scoring represent independent sources of measurement error. Reports of reliability/precision studies should make clear which of these sources are reflected in the data. Generalizability studies and variance component analyses can be helpful in estimating the error variances arising from each source of error. These analyses can provide separate error variance estimates for tasks, for judges, and for occasions within the time period of trait stability. Information should be provided on the qualifications and training of the judges used in reliability studies. Interrater or interobserver agreement may be particularly important for ratings and observational data that involve subtle discriminations. It should be noted, however, that when raters evaluate positively correlated characteristics, a favorable or unfavorable assessment of one trait may color their opinions of other traits. Moreover, high interrater consistency does not imply high examinee consistency from task to task. Therefore, interrater agreement does not guarantee high reliability of examinee scores.

Cluster 4. Factors Affecting Reliability/Precision

Standard 2.8

When constructed-response tests are scored locally, reliability/precision data should be gathered and reported for the local scoring when adequate-size samples are available.

Comment: For example, many statewide testing programs depend on local scoring of essays, constructed-response exercises, and performance tasks. Reliability/precision analyses can indicate that additional training of scorers is needed and, hence, should be an integral part of program monitoring. Reliability/precision data should be released only when sufficient to yield statistically sound results and consistent with applicable privacy obligations.

Standard 2.9

When a test is available in both long and short versions, evidence for reliability/precision should be reported for scores on each version, preferably based on independent administration(s) of each version with independent samples of test takers.

Comment: The reliability/precision of scores on each version is best evaluated through an independent administration of each, using the designated time limits. Psychometric models can be used to estimate the reliability/precision of a shorter (or longer) version of an existing test, based on data from an administration of the existing test. However, these models generally make assumptions that may not be met (e.g., that the items in the existing test and the items to be added or dropped are all randomly sampled from a single domain). Context effects are commonplace in tests of maximum performance, and the short version of a standardized test often comprises a nonrandom sample of items from the full-length version. As a result, the predicted value of the reliability/precision may not provide a very good estimate of the actual value, and therefore, where feasible, the reliability/precision of both forms should be evaluated directly and independently.

Standard 2.10

When significant variations are permitted in tests or test administration procedures, separate reliability/precision analyses should be provided for scores produced under each major variation if adequate sample sizes are available.

Comment: To make a test accessible to all examinees, test publishers or users might authorize, or might be legally required to authorize, accommodations or modifications in the procedures that are specified for the administration of a test. For example, audio or large print versions may be used for test takers who are visually impaired. Any alteration in standard testing materials or procedures may have an impact on the reliability/precision of the resulting scores, and therefore, to the extent feasible, the reliability/precision should be examined for all versions of the test and testing procedures.

Standard 2.11

Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.

Comment: Reporting estimates of reliability/precision for relevant subgroups is useful in many contexts, but it is especially important if the interpretation of scores involves within-group inferences (e.g., in terms of subgroup norms). For example, test users who work with a specific linguistic and cultural subgroup or with individuals who have a particular disability would benefit from an estimate of the standard error for the subgroup. Likewise, evidence that preschool children tend to respond to test stimuli in a less consistent fashion than do older children would be helpful to test users interpreting scores across age groups.

When considering the reliability/precision of test scores for relevant subgroups, it is useful to evaluate and report the standard error of measurement as well as any coefficients that are estimated. Reliability and generalizability coefficients can differ substantially when subgroups have different variances on the construct being assessed. Differences in within-group variability tend to have less impact on the standard error of measurement.

Standard 2.12

If a test is proposed for use in several grades or over a range of ages, and if separate norms are provided for each grade or each age range, reliability/precision data should be provided for each age or grade-level subgroup, not just for all grades or ages combined.

Comment: A reliability or generalizability coefficient based on a sample of examinees spanning several grades or a broad range of ages in which average scores are steadily increasing will generally give a spuriously inflated impression of reliability/precision. When a test *is* intended to discriminate within age or grade populations, reliability or generalizability coefficients and standard errors should be reported separately for each subgroup.

Cluster 5. Standard Errors of Measurement

Standard 2.13

The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score.

Comment: The standard error of measurement (overall or conditional) that is reported should be consistent with the scales that are used in reporting scores. Standard errors in scale-score units for the scales used to report scores and/or to make decisions are particularly helpful to the typical test user. The data on examinee performance should be consistent with the assumptions built into any statistical models used to generate scale scores and to estimate the standard errors for these scores.

Standard 2.14

When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

Comment: Estimation of conditional standard errors is usually feasible with the sample sizes that are used for analyses of reliability/precision. If it is assumed that the standard error *is* constant over a broad range of score levels, the rationale for this assumption should be presented. The model on which the computation of the conditional standard errors is based should be specified.

Standard 2.15

When there is credible evidence for expecting that conditional standard errors of measurement or test information functions will differ substantially for various subgroups, investigation of the extent and impact of such differences should be undertaken and reported as soon as is feasible.

Comment: If differences are found, they should be clearly indicated in the appropriate documentation. In addition, if substantial differences do exist, the test content and scoring models should be examined to see if there are legally acceptable alternatives that do not result in such differences.

Cluster 6. Decision Consistency

Standard 2.16

When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

Comment: When a test score or composite score is used to make classification decisions (e.g., pass/fail, achievement levels), the standard error of measurement at or near the cut scores has important implications for the trustworthiness of these decisions. However, the standard error cannot be translated into the expected percentage of consistent or accurate decisions without strong assumptions about the distributions of measurement errors and true scores. Although decision consistency is typically estimated from the administration of a single form, it can and should be estimated directly through the use of a test-retest approach, if consistent with the requirements of test security, and if the assumption of no change in the construct is met and adequate samples are available.

Cluster 7. Reliability/Precision of Group Means

Standard 2.17

When average test scores for groups are the focus of the proposed interpretation of the test results, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases the standard error of the group mean should be reported, because it reflects variability due to sampling of examinees as well as variability due to individual measurement error.

Comment: The overall levels of performance in various groups tend to be the focus in program evaluation and in accountability systems, and the groups that are of interest include all students/clients who could participate in the program over some period. Therefore, the students in a particular class or school at the current time, the current clients of a social service agency, and analogous groups exposed to a program of interest typically constitute a sample in a longitudinal sense. Presumably, comparable groups from the same population will recur in future years, given static conditions. The factors leading to uncertainty in conclusions about program effectiveness arise from the sampling of persons as well as from individual measurement error.

Standard 2.18

When the purpose of testing is to measure the performance of groups rather than individuals, subsets of items can be assigned randomly to different subsamples of examinees. Data are aggregated across subsamples and item subsets to obtain a measure of group performance. When such procedures are used for program evaluation or population descriptions, reliability/precision analyses must take the sampling scheme into account.

Comment: This type of measurement program is termed *matrix sampling.* It is designed to reduce the time demanded of individual examinees and yet to increase the total number of items on which data can be obtained. This testing approach provides the same type of information about group performances that would be obtained if all examinees had taken all of the items. Reliability/precision statistics should reflect the sampling plan used with respect to examinees and items.

Cluster 8. Documenting Reliability/Precision

Standard 2.19

Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.

Comment: Information on the method of data collection, sample sizes, means, standard deviations, and demographic characteristics of the groups tested helps users judge the extent to which reported data apply to their own examinee populations. If the test-retest or alternate-form approach is used, the interval between administrations should be indicated.

Because there are many ways of estimating reliability/precision, and each is influenced by different sources of measurement error, it is unacceptable to say simply, "The reliability/precision of scores on test X is .90." A better statement would be, "The reliability coefficient of .90 reported for scores on test X was obtained by correlating scores from forms A and B, administered on successive days. The data were based on a sample of 400 10th-grade students from five middle-class suburban schools in New York State. The demographic breakdown of this group was as follows: . . ." In some cases, for example, when small sample sizes or particularly sensitive data are involved, applicable legal restrictions governing privacy may limit the level of information that should be disclosed.

Standard 2.20

If reliability coefficients are adjusted for restriction of range or variability, the adjustment procedure and both the adjusted and unadjusted coefficients should be reported. The standard deviations of the group actually tested and of the target population, as well as the rationale for the adjustment, should be presented.

Comment: Application of a correction for restriction in variability presumes that the available sample is not representative (in terms of variability) of the test-taker population to which users might be expected to generalize. The rationale for the correction should consider the appropriateness of such a generalization. Adjustment formulas that presume constancy in the standard error across score levels should not be used unless constancy can be defended.