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

egree of consistency or
assification is specific to
and its location within

## of Group Means

verage) scores of groups
ain categories) involve
lifferent from those that
level. Such estimates are
of program effectiveness
ional accountability sys-
luate the effectiveness of

rformance by estimating
r mean improvement in
from the group, the vari-
ng of persons can be a
specially if the sample
nt that different samples
t (e.g., all students who
naterials) yield different
t the expected outcome
group (including those
p in the future) are un-
s, the variability due to
in the estimates of the
uite small. However, in
of persons are not very
he mean achievement of
room or the average ex-
amples of clients in a
ror associated with the
be a major component
e a significant source of
programs even if there is
in individual test scores.
ndividual scores are not
he precision of group av-
e statistic is the standard
the group means.

## lity/Precision

distributors of tests have
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 v
of the test, and the obs
treated as a sample fro
tasks, the variability in s
another would be consi
used to assign scores to r
in scores over qualified r
Different sources of erro
single coefficient or star
be evaluated separately,
addressed in some way. R
cision should specify tl
error included in the ana

### Cluster 2. Evaluating Reliability/Precision

### Standard 2.3

**For each total score, sul
of scores that is to be i
relevant indices of relia
be reported.**

**Comment:** It is not suffi
ofreliabilities and standar
only for total scores whe
terpreted. The form-to-fo
sistency of total scores on
high, yet subscores may
reliability, depending on
and used. Users should be
data for all scores to be
data should be detailed
users to judge whether
enough for the intended
Composites formed from
a test battery are frequentl
and diagnostic purposes.
about the reliability of su

### Standard 2.4

**When a test score inte
differences between two**

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

lity. Information should
fications and training of
lity studies. Interrater or
may be particularly im-
ɔservational data that in-
ons. It should be noted,
s evaluate positively cor-
favorable or unfavorable
ıay color their opinions
ıigh interrater consistency
minee consistency from
ıterrater agreement does
ıility of examinee scores.

## fecting

ıse tests are scored locally,
should be gathered and
coring when adequate-
:.

, many statewide testing
al scoring of essays, con-
s, and performance tasks.
yses can indicate that ad-
rs is needed and, hence,
of program monitoring.
ı should be released only
statistically sound results
cable privacy obligations.

in both long and short
iability/precision should
each version, preferably
dministration(s) of each
ıt samples of test takers.

ty/precision of scores on
luated through an inde-
each, using the designated
c models can be used to
recision 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 s
or school at the current
a social service agenc
exposed to a program of
a sample in a longitudir
parable groups from the
in future years, given st
leading to uncertainty in
effectiveness arise from
well as from individual

### Standard 2.18

**When the purpose of
performance of group
subsets of items can be
ferent subsamples of exa
across subsamples and
measure of group perfc
cedures are used for pr
ulation descriptions, rel
must take the sampling**

Comment: This type o
termed *matrix samplin*
the time demanded of
yet to increase the to
which data can be obtai
provides the same ty
group performances th
examinees had taken all c
cision statistics should
used with respect to exa

## Cluster 8. Documer Reliability/Precisio

### Standard 2.19

**Each method of quant
cision of scores should
expressed in terms of
the method. The samp
select test takers for reli
and the descriptive sta
subject to privacy oblig
should be reported.**