RELIABILITY ISSUES AND EVIDENCE

Introduction

This paper discusses assessment *reliability* with emphasis on the benefits of careful assessment design and administration when used for measuring students with disabilities. Our discussion centers on the concept of measurement error, specifically in the context of (a) the process for collecting responses, (b) the scores assigned to observed responses, (c) the decisions made based on these scores, and (d) the reliability of an assessment system. The importance of reliability pivots around the need for assurances that assessments are designed and used in ways that minimize unstable response patterns and corresponding individual and collective examinee scores. Reliable measurement is also a necessary condition for measurement of validity—although it is not the only condition. Without reliability, it is impossible to determine whether an assessment accurately measures student achievement. The challenge that must be addressed is to offer flexible assessments that can be adapted to different student needs.

Perhaps the most psychometrically technical aspect of assessment, reliability is generally described in terms of score consistency. The *Standards for Educational and Psychological Testing* define reliability as "the consistency of [such] measurements when the testing procedure is repeated on a population of individuals or groups" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 25). Reliability typically refers to the measurement error that is introduced into the "entire measurement process" (p. 27), limits the degree to which generalizations can be made beyond the specific testing event, and quantifies the confidence that can be held in the value assigned to any performance. "Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores" (p. 31). Specifically, for the purposes of this paper, we are concerned about the reliability (dependability, replicability, etc.) of behavior, scores, and inferences, as well as accounting for types of error.

Error can be classified into two types: (a) systematic and (b) unsystematic (random). Systematic error addresses validity issues; unsystematic error address reliability issues. Reliability is related to measurement error, which "almost always refers to the random component of error" (Feldt & Brennan, 1989, p. 105). Because large-scale assessments involve so many steps for

development, implementation, and analysis, unsystematic error enters into the process in many different ways. Obviously, the use of performance assessments and testing accommodations can introduce a host of additional sources of error beyond the student and item. The design and development of items and tasks may introduce unsystematic error; for example, performance tasks, while considered comparable, may render alternate forms nonequivalent. Unsystematic error can result from varied assessment implementation by different teachers, and in different classrooms with different students. Finally, the scoring process itself may introduce unsystematic error (e.g., scoring via raters).

Calculating an index of reliability requires quantifying the measurement error associated with (a) observed behaviors and (b) their associated numeric scores. The situation becomes complex when observed behaviors depend on the sampling of items and the manner in which items "elicit" observed behavior. This is true irrespective of whether an item format uses a selected response (SR) or constructed response (CR) format. Furthermore, assigning numeric values to observed behaviors—that is scoring and scaling—affects the reliability of the measurement system. Scoring issues pertain to whether the score is very specific (e.g., using a scale from 1–500) or very general (e.g., with three score levels, as in conventional classification standards of "does not meet," "meets," or "exceeds"). Ultimately, we need some indication that careful assessment design (item sampling, administration, and scoring) diminishes error.

States have designed a range of approaches to assessment so that students can freely participate. Yet this range also may result in the introduction of unsystematic error and the potential for an array of random "nuisance" factors that may threaten the psychometric reliability of assessments. Because a state's assessment comprises this wide variety of approaches, and given their implementation with diverse populations, multiple types of evidence are required to ensure that reliable measures are obtained.

This discussion begins with first providing a conceptual definition of reliability, then identifies sources of error, and finally describes evidence that focuses on measurement reliability and measurement designs to attenuate measurement error. Note: Error scores, parallel forms, reliability coefficients, and standard error of measurement are the most important concepts in defining reliability. The sources of error often arise from procedural components of design and delivery of a large-scale assessment; the impact of that error is then documented through statistical analysis. This combination of procedural and statistical evidence forms the first line defense for developing of a validity argument. All tests must be reliable; their reliability,

however, does not guarantee the validity of inferences from the results. It would be impractical to thoroughly cover most of the relevant technical psychometric issues pertaining to measurement reliability in this paper. Readers wishing for more technical documentation should refer to the references provided throughout.

Definition of Reliability

It is difficult to appraise the presence of reliability without definitions. Most important are the concepts of error scores, parallel forms, reliability coefficients, and standard error of measurement.

Error Scores

One of the most traditional conceptualizations is in terms of the *true score*: "a personal parameter that remains constant over the time required to take at least several measurements" and "the limit approached by the average of observed scores as the number of these observed scores increases" (Feldt & Brennan, 1989, p. 106). Unfortunately, it is impossible to know a person's true score; it must be estimated from the *observed score*, which provides imperfect information. Therefore, in addition to the observed score, an *error score* must be theorized. A very simple concept of observed score, true score, and error score is captured in Equation 1.

observed score = true score + error score. (1)

The observed score is composed of two components: (a) the true score and (b) the error score. Both the true score and error score are unobserved and must be estimated.

The concept of error score is at the heart of reliability. The goal of good measurement design is to minimize the error component. Note: In the simple model above (Equation 1), error is thought to occur randomly. The importance of random error may be recognized if an assessment is used repeatedly to measure the same individual. The observed score would not be the same on each repeated assessment. In fact, scores are more or less variable, depending on the reliability of the assessment instrument. The best estimate of an examinee's true score is the average of observed scores obtained from repeated measures. The variability around the mean is the theoretical concept of error, also called error variance. As noted earlier, measurement error can occur in the form of either systematic bias, which deals with construct validity, or random error, which deals with reliability. Random error can never be eliminated completely.

Parallel Forms

A formal concept of error is developed largely around assumptions pertaining to parallel forms. To estimate error scores, it is not advisable to administer the same assessment repeatedly to the same examinee. It is more effective to use parallel forms of the assessment. Parallel forms are assessments comprising different tasks, but the tasks are designed so that they can be assumed to be randomly sampled from the same domain of comparable difficulty. The correlation $r_{x_1x_2}$ of scores from any two parallel forms, x_1 and x_2, are highly correlated only if the assessment is highly reliable. The concept of correlated parallel forms lets us continue the definition of psychometric reliability. Equation 2 describes $r_{x_1x_2}$ in terms of observed score variances v_{x_1} and v_{x_2}, and their covariance $v_{x_1x_2}$.

$$r_{x_1x_2} = v_{x_1x_2} \Big/ sd_{x_1} sd_{x_2} \tag{2}$$

Equation 2 can be written in terms of true score and observed score variance (Feldt & Brennan, 1989; Chatterji, 2003). Equation 3 shows that the observed correlation of two parallel forms provides information for estimating assessment reliability. Substituting Equation 1 in Equation 3, Equation 4 shows that observed score variance is composed of true score and error score variance. As error score diminishes, the ratio of true score and observed score variance approaches a value of 1. So, if the correlation of parallel forms, $r_{x_1x_2}$, approaches one, then the error variance must be small. Conversely, if $r_{x_1x_2}$ is small, the error variance must be large.

$$r_{x_1x_2} = v_{true} \Big/ v_{observed} \tag{3}$$

$$r_{x_1x_2} = v_{true} \Big/ (v_{true} + v_{error}) \tag{4}$$

While the assumption of parallel forms (items or tests) is generally necessary psychometrically, it is extremely difficult to accomplish. Sources of error are identified below. Use of nonequivalent (nonparallel) forms is identified as one of the most important and difficult to control.

Standard Error of Measurement and Information

Another conceptualization of measurement accuracy is developed in terms of the standard error of measurement (SEM). As described above, the concept of random error around the true score results from administering repeated parallel forms. The SEM of a measure is essentially the average deviation of error scores around the true score. As with reliability, SEM (σ_e) can be estimated in terms of correlated observations x_1 and x_2. According to Equation 5, as the correlation of parallel forms increases, the standard error of measurement diminishes toward zero.

$$\sigma_e = \sigma_x \sqrt{r_{x_1 x_2}} \sqrt{1 - r_{x_1 x_2}} \tag{5}$$

It is important to keep in mind that these measures are estimations. Theoretically, each time the assessment is administered a different measure is likely to be obtained. The degree of difference depends on the reliability or error in measurement.

The preceding SEM estimation is classical, in contrast to an alternative Item Response Theory (IRT) perspective. In IRT, items are calibrated with respect to difficulty and discrimination among many other possible item characteristics. Using the item calibrations, it is possible to estimate the amount of *information* provided by a test and its items. Furthermore, the amount of information depends on the ability of the examinees. For instance, a difficult item administered to someone with low ability will not generate meaningful, informative results. Response to an easy item provides much more information about someone with low ability. As a rule, items are most informative when responded to by a person with an ability level comparable to the level of the item. (Note: In IRT, item difficulty and person ability are on the same scale, which makes it possible to match items to persons.) A test that is not too easy or too difficult for the respondent is a highly informative test.

Sources of Random Error

Random error arises from student variables, task sampling, item calibration, and scaling, as well as other sources. These different sources affect the process at different times in the development and implementation of large-scale assessments; therefore, they need to be documented and monitored throughout the process. The effect can then be minimized to provide more stable and dependable estimates of students' performance. The documentation would provide appropriate procedural evidence to allow the formulation of a validity argument.

With appropriate analyses, statistical evidence would be used to complement the procedural evidence. However, as noted in the Standards (American Educational Research Association et al., 1999), various forms of reliability estimates are possible, and they need to address specifically the source of error for which they are targeted. For example, if raters are used in the scoring process, then interjudge reliability needs to be documented; with alternate forms, this type needs to be noted; when change over time is being documented, test-retest reliability needs to be established; finally, internal consistency provides evidence of reliability of items and tasks.

An alternate assessment poses numerous challenges that are associated with measurement error. Some sources of random error pertain to examinee characteristics, item and test design, administration, and scoring protocols. State large-scale assessments typically use both SR and CR items or tasks, either with or without accommodations. CR is used in performance measures that require a rubric (subjectively scored) or performance measures that require observation of student performance, completion of performance tasks, or collection of student work samples. The opportunities for measurement error are likely to expand with increased flexibility. As a consequence, assessment design and reliability estimates need to take into account the multiple factors that can attenuate measurement accuracy. The challenge with isolating and controlling sources of measurement error is complicated by the relationships among error sources, as described below.

Student Variables

Students come into school situations from a variety of home environments, all of which can affect their performance in school. For example, students come to school hungry, tired, or fatigued, and so forth. As they interact with classroom tasks and receive feedback, students come to have expectations of success or failure, reflecting motivation and self-efficacy that may interact differentially with the kinds of tasks they are given. All of these conative factors may influence the results of large-scale assessments in unsystematic (i.e., random) ways (McGrew, Johnson, Cosio, & Evans, 2003).

In addition, for students with disabilities, a number of personal and behavioral characteristics may also unsystematically influence performance. For example, with some disabilities (e.g., attention deficit-hyperactivity disorders), medications are used; depending upon the dosage or uptake, performance on large-scale tests may be inconsistent. Even without the use of medications, students with disabilities may exhibit behavioral tendencies that distract them from

attending to tasks (tendencies of perseveration, distractibility, inattentiveness, etc.). The administration of the test may be nonstandardized and therefore may influence students unevenly (e.g., it may negatively affect some and act neutrally for others). Whenever such behavior or conditions influence students' performance unsystematically, reliability is weakened, as is the overall claim of validity (the claim that the outcomes reflect what the student knows and can do). Therefore, the inference of proficiency is less certain. A careful analysis of the context and the student is needed, however, as some variations in personal state (health, attention deficit-hyperactivity disorders) would be regarded as sources of systematic error. For example, the *Standards* note that test anxieties that can be "recognized in an examinee" are considered "systematic errors" and "are not generally regarded as an element that contributes to unreliability" (American Educational Research Association et al., 1999 p. 26).

As a consequence, participation in large-scale assessment systems is not only a matter of scheduling students to take a test at the end of the year. Rather, the assessment needs to be considered an important part of the school's annual cycle of activities. The large-scale assessment program needs to take into account such behavioral factors when collecting performances from students. Although tests may be given only once during the year (typically in the spring), plans for test administration should be introduced early in the year to allow students and teachers a fair opportunity to participate.

As an example of testing conditions reflecting ongoing classroom conditions, many states require teachers to use the same accommodations in testing that have been part of the accommodations used in the classroom. If these accommodations are not implemented in a standardized manner as part of the teaching or testing, unsystematic variance may be introduced. Furthermore, a teacher who is watchful during the year may be able to better understand critical student behaviors and recommend specific accommodations for testing at the end of the year.

Task Sampling

Samples of performance tasks must be prepared so that they are parallel in format and difficulty. That is, the tasks are ideally comparable to the extent that a student would not perform differently with one or another because they are both of equal difficulty. The sample of tasks is apt to be more or less variable with respect to difficulty and representation of the performance domain. Using multiple forms, individuals can be assessed over time or compared to another. The extent to which tasks differ is of obvious consequence because, with more variation, the

change over time or comparisons over multiple individuals is less trustworthy. Score variability that is attributable to task differences needs to be identified with carefully controlled studies in which parallel tasks and forms are used.

With portfolio assessments, it is often the teacher who selects the student work to be included in a collection or portfolio. Although selection criteria may be specified in both the test administration manual and in training, teacher judgment is ultimately involved. Consequently, the portfolio or collection of work may represent the grade level content broadly or narrowly. Choices of what is included or excluded in the collection can therefore affect the adequacy of the evidence in representing what a student knows and can do. From the perspective of repeatability, another collection, assembled at a different time or by a different teacher, may or may not support inferences drawn from the original collection. Therefore, it is critical to consider task sampling in the context of the assessment approach. It is generally easier to establish parallel forms when dealing with brief constructed tasks; when using performance collections and portfolios, it may be more difficult to establish comparability of tasks and forms.

Item Calibration

Assessment developers increasingly recognize the value of item calibration, since assessment items are not necessarily equivalent (Thissen & Wainer, 2001). Whether CR or SR, assessment items provide differential amounts of information depending on the respondent's true ability. Item calibrations are estimates of item characteristics such as item difficulty or sensitivity (van der Linden & Hambleton, 1997). With accurate item calibrations, estimation of true scores becomes considerably more accurate. That is, calibration helps minimize standard error of estimation.

Calibration accuracy pertains directly to measurement reliability. The value of item calibrations for ability estimation depends on the appropriate choice of IRT model and proper calibration procedures. The technicalities involved in these decisions are far beyond the scope of this paper, but the importance of good calibration should be noted. First, the calibration process requires adequate sampling of examinee response patterns. Ideally, a range of abilities is represented in the calibration sample. Second, an appropriate IRT model must be applied. For instance, alternate assessments rely heavily on performance tasks. Usually, observed performance is scored polytomously, that is, more than correct/incorrect. This method of scoring requires the use of a rating scale, partial credit, or graded response model. Numerous other

possible models are described in the literature (van der Linden & Hambleton, 1997; Boomsma, van Duijn & Snijders, 2001).

Another aspect of IRT item calibration that can influence reliability is the unidimensionality of the assessment—the degree to which an assessment measures a single construct (an ideal condition). In reality, this outcome is very difficult to achieve on rigidly constructed measures. Also, when using alternate assessments, the challenge increases dramatically as flexible CR tasks are applied and risk the involvement of multiple ability factors and other variable factors such as time constraints, rater severity, and so forth. Multidimensional IRT models can be used to accurately calibrate performance tasks, thereby yielding more reliable ability estimation. Local item dependency is generally attributable to multidimensional problems. Good assessment development must identify and provide corrections for this situation.

Finally, when developing measures for diverse populations, it is important to understand whether assessment tasks function identically across the populations. Ideally, tasks should perform equivalently across populations, although one population may have higher mean ability than another population. This type of analysis helps maintain quality control over assessment development. For example, Yovanoff and Tindal (in press) were able to understand how well the Oregon Early Reading Extended Assessment tasks functioned irrespective of whether students were in special or general education. In this study, a range of constructed response tasks was placed on the same scale and used as the first benchmark of the state test to provide students of all abilities a sufficient range of difficulties, leading to appropriate assessment. These tasks included letter naming and letter sounding as well as word, sentence, and passage reading.

Scaling and Equating

If items are calibrated, they can be fitted onto a scale and then used to monitor change over time or compare students of differing abilities. In this process, assessments need to be rescaled or equated with an external measure. State measurement programs perform this function when scores on alternate assessments are placed on the same scale as the general assessment scores. As noted above, the Oregon Early Reading Extended Assessment (Yovanoff and Tindal, in press) was scaled with the general Oregon Statewide Assessment. Using the appropriate IRT model and necessary research design, Yovanoff and Tindal demonstrated that the early reading performance tasks functioned appropriately for students who were otherwise unable to be measured accurately with the general benchmark on statewide assessment.

In the end, this type of scaling and measurement calibration makes the assessment both more accurate and more informative. Equating standard errors is extremely important for appraising the accuracy of score equivalents (Kolen & Brennan, 1995). Whenever assessments are equated, the standard error should be reported for both the overall population and individual students. Standard error is conditional on the student's score. If the scale contains too few items that are appropriate for a student's ability, the SEM is greater, and it is more difficult to accurately locate the student on the scale.

Scoring Process

Irrespective of any errors made in collecting assessment data or as estimated with reliability coefficients, different or unique errors can also be made when making judgments. This type of random error refers to ratings and classifications made for students, such as pass/fail or below basic, basic, proficient, and advanced. In this instance, the focus is less on the actual score consistency than on the consistency of judgments about states of mastery. Two types of judgments can contain error: (a) at the score level, the focus is on rubrics (or partial correct responses); (b) at the classification level, the focus is not only on the final decision to classify a student's performance but also on the standard-setting process itself. The analysis, therefore, needs to consider both the individual judgments made for a student as well as the overall process for making classification decisions.

Although score errors need to be addressed, classification errors are far too serious, are more difficult to detect, and require more resources to resolve. Furthermore, whereas score error is usually minimized at the cut score, judgment error is most problematic at the cut score. According to the *Standards*:

> Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mis-measured without serious consequences. Mis-measurement of examinees whose true scores are close to the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical score. (American Educational Research Association et al., 1999, p. 3)

Even with the conditional SEM (for an individual score) reported at the cut score, classification judgments can be problematic. For example, Hollenbeck and Tindal (1999) reported that, although judges were in considerable agreement at the exact or adjacent score values, they were in the greatest disagreement with respect to judgments of proficiency (at the cut score). In this study, judges agreed about writing quality (using a 6-point scale) when it was judged very low (rated 1 or 2) or very high (rated 5 or 6); they disagreed, however, when writing scores were in the middle (at ratings of 3 or 4, which includes the cut score of 4 for passing). As a consequence, the state educational agency began reporting "conditional" proficiency (in essence noting that disagreement occurred at the cut score) to acknowledge this type of error.

Reliability at the classification level involves attention to proper selection of content experts as well as training and feedback. The *Standards* are very clear:

When subjective judgment enters into test scoring, evidence should be provided on both inter-rater 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, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products. (American Educational Research Association et al., 1999, p. 34)

The distinction between the reliability of the score and the judgment sometimes becomes blurred when both are estimated at the same time. A hybrid model, taking into account both the reliability of the scoring process and the decisions made from the score, is created in the case in which a single judgment is based on all evidence. Oregon's juried assessment represents such a combined judgment that is primarily of the classification but includes evidence of individual scores also (e.g., performance on classroom tests or achievement on the state test when it has been modified) (Yovanoff & Tindal, in press). In analyzing reliability for this system, the key is to ensure both that the judgment is reliable and that the achievement is judged against grade level standards.

If Oregon's approach with the juried assessment is typical of what states develop to meet the "alternate assessment judged against grade level content standards," there is likely to be heavy reliance on two categories of judgment: one that addresses the *sufficiency* of evidence to make a determination and another that addresses *proficiency* based on the collection as a whole. To achieve reliability in these judgments, Oregon relies on systematic procedures and structured

criteria for making the determination: One indirectly increases reliability by identifying what was thought to be random error, and identifying portions of it as having systematic causes. These causes can then be addressed through systematic procedures, thereby decreasing the amount of random error. Oregon's method was considered by the state's national Technical Advisory Committee (TAC) and considered a reasoned and prudent approach to avoiding false negative judgments (e.g., denying proficiency when it was deserved). Some committee members conjectured that the juried assessment might actually present a higher standard of achievement than the general multiple-choice assessments. Their view was noteworthy in that they considered the approach on its merits and weighed against both the *Standards* for testing and the consequences to the student. In the end, the TAC considered it a promising strategy and determined that they had nothing better to offer that would meet the same demands for integrity and fairness.

The Juried Assessment administration manual describes them as being completed under nonstandard conditions—either through the Collection of Evidence process or through a Collection to Jury a Modification to answer the question: "Does the evidence provided by the student meet the Oregon content and performance standards for a particular subject?" (Oregon Department of Education, 2005, p. 5). Training in scoring is provided to ensure reliability; in addition, collections that are rated above the standard must be verified through a secondary source. The requirements for a collection of evidence in Oregon's Juried Assessment are clarified in subject specific documents on the Web site along with the administration manual. A description of the juried process was extracted from the manual for inclusion here. The Juried Assessment is designed for students literate in a language other than English, with either physical disabilities preventing participation in the writing assessment or any other disability affecting the student's ability to read and write. According to the *Juried Assessment Guidelines*:

The Moderation Panel would consider the evidence and determine whether test results using the translation in the first example, the word prediction software in the second, or the auditory methods in the third, are reliable and valid in addressing a specific standard. If the panel determines that the change does not affect the validity of the test score for this student, the student's score would then be considered for meeting that standard.

Juried Modifications are approved one student and one assessment at a time. A panel of experts makes the final determination. It is believed that there may be a student with a significant learning disability who uses assistive technology, screen readers, and recorded text

to perform the task of understanding text and interpreting “meaning”. The panel might approve this modification as an accommodation for the particular student after reviewing the student's case if:

- The student is skilled in using the read aloud adaptations
- The measure of comprehension reflects the student’s own knowledge and understanding
- The student achieves the same standards for interpreting text required of all students

If approved, the student would be permitted to use the “read aloud” modification with the Reading/Literature Knowledge and Skills assessment and have the opportunity to “meet” (e.g. be determined “proficient” on the standard). There is the possibility that the decision could be made after testing was completed if there was sufficient documentation of the process to assure that it was the student’s own work (Oregon Department of Education, 2005, p. 5).

Alignment

Another source of unsystematic error in the data collection process is introduced in the participation of students in alternate assessments when conducting an alignment analysis between grade level content standards and portfolio assessment approaches. When states use portfolios as part of the alternate assessment that are defined by student need (e.g., a fixed set of entries are not specified a priori, but teachers select unique entries for each student), the process of alignment between the alternate assessment and the standards cannot take place without sampling students. If the sampling of standards is isomorphic with the sampling of students, any statements about alignment on content coverage, breadth of knowledge, and depth of knowledge are primarily a function of those who participated. In this source of error (more like survey sampling error), stable statements about alignment are difficult to make. Because each student samples only a prespecified set of standards by design, alignment at the student level is inherently skewed. As a result, the process needs to sample a sufficient number of students to determine the coverage of standards being addressed at the system level. At this level, sampling of students needs to be considered using not only ages but also disabilities, geographic region, and type of program to make any inferences about alignment.

Assessment Administration

A final source of error relates to assessment administration. One reason for using standardized procedures in large-scale assessment systems is to minimize the error from external sources.

Testing personnel (most often teachers), however, can introduce error (unsystematic variance) through the way that they administer or score the test. Ironically, few states have training systems for test administration. Educators assume that the conditions as noted in the test booklets are the same as those enacted in the classroom. Significant deficits are evident in teacher knowledge concerning high-stakes testing. Most teachers' knowledge about testing and measurement comes from "trial-and-error learning in the classroom" (Wise, Lukin, & Roos, 1991, p. 39). This problem, however, is rarely addressed through any in-service programs, even though these authors attributed the lack of assessment knowledge to teacher certification agencies at the state level (i.e., states do not require assessment/measurement courses for initial teacher certification).

This kind of unsystematic error is best addressed by state educational agencies (SEAs) through rigorous training and monitoring throughout administration of the large-scale assessment system. "Measurements derived from observations of behavior or evaluations of products are especially sensitive to a variety of error factors. These include evaluator biases and idiosyncrasies, scoring subjectivity, and intra-examinee factors that cause variation from one performance or product to another (American Educational Research Association et al., 1999, p. 29). Evidence can be both procedural and empirical for documenting the reliability associated with (a) test administration, and (b) response rating. Because random measurement errors are inconsistent and unpredictable, they cannot be removed from observed scores. However, their aggregate magnitude can be summarized in several ways…." (American Educational Research Association et al., 1999, p. 27).

Options for Participation in Large-Scale Assessments

Participation methods present a somewhat different challenge regarding sources of error. The estimate of reliability for the first method, taking the general assessment without accommodations, is likely to be a characteristic that has already been addressed in most states' technical reports. There is still some question about which students participated in the assessment or, more importantly, which subgroups did not participate, and the possible affect participation might have had on estimates of reliability.

Participation in the general assessment with accommodations increasingly has been studied (Sireci, Li, & Scarpati, 2003), and the use of accommodations increasingly has been utilized (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005), although in both instances, little is definite about the influence of reliability on either participation or the use of accommodations

(see section on standard error of measures). Nevertheless, the 1999 *Standards* are quite clear: "When significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each major variation if adequate sample sizes are available" (American Educational Research Association et al., 1999, p. 36).

As a consequence, little is known about estimates of reliability in any testing program in which changes are made (either as accommodations or for the remaining three methods: alternate assessments judged against grade level, modified, or alternate achievement standards). These latter methods present particular challenges because the quantity of data resulting from the method is likely to be limited in scope. Therefore, making inferences beyond an instance of testing is difficult. These are (a) alternate assessment judged against grade level and (b) alternate assessment judged against alternate achievement standards. It is possible that the alternate assessment based on modified achievement standards can follow the course of research followed by accommodations with sufficient numbers and adequate standardization to allow generalizations.

Ironically, standardization may be the antithesis of the solutions for controlling measurement error. Again, there is no direct control of random error, but by identifying systematic sources of what had been considered random error, total error (and therefore estimates of measurement error) can be reduced. By forcing tests to be taken in the same way across all students, both internal and external sources of error may be exacerbated [maintained] rather than controlled. For example, a student with hyperactive tendencies (and who therefore takes medications to control sources of error that are internal to the student due to inattentiveness) given a test in one session (to control sources of error external to the student due to time and setting) may actually need to have accommodations made in order to make appropriate inferences about performance that are not influenced by construct-irrelevant variance. When accommodations are made, reliability-related evidence is needed to support the consistency of administration and scoring across replications (time, items, and raters).

Generalizability Theory and Differentiated Error

So far, all error has been undifferentiated and treated as one source. However, this discussion has noted many plausible sources of this undesirable error. Using generalizability theory, this single "error" term is decomposed into various *facets*, or factors, that influence performance (Brennan, 2001). The most consistently studied facets are judges, tasks, and occasions. Using

carefully planned research designs, assessment developers can better understand to what extent the assessment facets influence the reliability of observations.

Generalizability studies (G-study) are used to differentiate error and identify how much of the examinee score is attributable to, for instance, lack of rater agreement or task variability. This information is extremely valuable as it casts light on where assessments need adjustment. Obviously, alternate assessment can benefit from this effort and identify exact sources of error. Once the error term is partitioned into specific sources, the assessment development research can proceed to estimate measurement reliability.

A typical finding of G-studies is that the primary source of variance is the task itself. For example, assessments of science using experiments or a paper and pencil test result in very different estimates of performance. Likewise, comparisons of other CR and SR formats may result in different performance estimates, primarily due to format rather than content. The influence of raters and occasions has typically not been found to be as influential as format.

Using Decision Studies, multiple assessment scenarios can be constructed along with the corresponding reliability estimate. For instance, with G-study information, it is possible to know how much reliability will improve if more raters were used or responses to more tasks were obtained. Typically, using more than five to seven raters does not substantially improve estimates of performance. The implications of this relation for time and money considerations are very clear. If longer examinee test times are not possible, then the assessment would not include more tasks. Instead, the addition of another rater could be considered to the extent that it would bring reliability up to an acceptable standard.

Types of Reliability

This paper has identified some of the primary sources of random error that can jeopardize the quality of alternate assessments. These sources need to be documented and monitored to provide procedural evidence that any measures of behavior are replicable. Corresponding to many of these error sources is an appropriate type of reliability. By analyzing performance, appropriate statistical evidence can be assembled to support a validity argument. As noted earlier, however, neither type of evidence is sufficient to support any claims or inferences. Further documentation is needed to document and provide validity evidence.

Conventional reliability indices such as Cronbach's alpha and Kuder-Richardson formulas, KR20 and KR21, are based generally on the concepts of observed score variance, true score, and error score variance (Feldt & Brennan, 1989). According to the formal definitions above (Equation 3 and Equation 4), reliability is considered as the ratio of true score variance to observed score variance. Ideally, an assessment will diminish error and maintain an observed score that is largely composed of true score. Generally, as error score variance diminishes, the correlation of observed and true scores approaches the maximum value of one. The correlation of parallel forms is conceptually identical to the correlation of true score and observed score, and it is one reliability index of popular interest. Of course, this depends on the equivalence of the two assessments as required by the definition of parallel forms. Conventional reliability indices and estimates of standard error allow understanding of the stability (consistency) of the score within the distribution and further calculate confidence intervals around the true score.

In the typical large-scale assessment, four types of reliability coefficients are considered, each associated with different sources of error: (a) test-retest, (b) parallel forms, (c) internal consistency, and (d) inter-rater agreement. Use test-retest if error is believed to be due to occasion or time; use parallel form if error is thought to be due to the form used; use internal consistency if error is believed to have been introduced by the specific sample of items, tasks, or behaviors; and use inter-rater agreement if there is any reason to question the judgment or rating of performance.

In alternate assessments, the approach taken more or less determines the type of reliability that is most important. And, as noted in the 1999 *Standards*, "a reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent" (American Educational Research Association et al., 1999, p. 32).

Teacher indirect observations (judgments) when using rating scales need to address inter-judge primarily to ensure the same ratings would be given by anyone else. It also may be important for the ratings to be made at two different times to ensure that the behavior is stable. When ratings are completed with subscores reported, internal consistency may be a necessary dimension to ensure that the individual judgments hang together within a subscore. Portfolios include a compilation of work samples usually collected in the natural environment (of the school, community, work, or family) and are judged on some dimension (such as generalization, independence, accuracy, etc.). Because a judgment is being made, inter-judge agreements are

critical. Because, however, the collection of work samples often includes slightly different forms, it may be important to consider the variance that accrues from these parallel forms. Finally, it is possible that internal consistency is needed to support the claim that various work samples are all related equally to the total score. For performance events, the most critical dimensions of dependability are the internal consistency (for brief tasks in which many items comprise a task) or parallel form (a particularly important facet in generalizability theory); test-retest may be important with extended tasks. Finally, performance task collections need to emphasize the same types of reliability as used with portfolios.

Table 1 and Table 2 summarize much the foregoing discussion. Table 1 provides a simple summary of appropriate reliability indices for various assessment designs.

Table 1

Assessment Methods and Appropriateness of Reliability Indices

	Reliability Indices				
Assessment Method	**Test-Retest**	**Parallel Form**	**Internal Consistency**	**Inter-Judge**	**IRT Calibration/ Standard Error**
Selected Response	Very Appropriate	Very Appropriate	Very Appropriate	Very Appropriate	Very Appropriate
Checklists and Rating Scales	Appropriate	Inappropriate		Very Appropriate	Appropriate
Portfolio	Inappropriate	Appropriate		Very Appropriate	
Performance Event		Appropriate	Very Appropriate	Not Applicable	Appropriate
Performance Tasks	Appropriate	Inappropriate		Very Appropriate	Appropriate

For most assessment approaches listed in Table 1, SEM or conditional standard error of measurement (CSEM) should be provided. The value of SEM estimates is that they enable the computation of confidence bands around examinee ability and item characteristic estimations. Table 2 lists some of the standards associated with reporting SEM and CSEM (American Educational Research Association et al., 1999).

Table 2

Standards Relevant to the Standard Error of Measurement

Standard 2.1
"For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported" (p. 31).
Standard 2.2
"The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale unite and in units of each derived score recommended for use in test interpretation" (p. 31).
Standard 2.14
"Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score" (p 35).

According to Nitko (1996), these different methods for obtaining reliability coefficients generate different results; the composition and variability of the group affects the reliability coefficient because the method is based on correlations. Better reliability results from more items, behaviors, or products; objective scoring generally results in more reliable results; and the amount of error associated with an examinees performance depends (is conditional) on their score level. Fleming, Taylor, and Carran (2004) actually compare two methods for determining inter-rater reliability, highlighting the effects when correlation coefficients are used versus percent agreement and conclude that judges can have judgments that are highly correlated and in disagreement.

Haertel (1999) also points out that several sources of error are not considered at the group statistics level and may actually become important when making statements about aggregate level outcomes. Student absences definitely influence performance at the score level and likely provide very unsystematic variance at the distribution level. Invalid or below chance scores influence the mean of a group with students having varying levels of attemptedness in completing an item. Invalid identification of students at the aggregate level influences the scores that are included in the distribution. Finally, a number of other mistakes in the calculation of group statistics have an obvious influence on the reliability of results.

While these forms of reliability are traditionally reported in the general education assessment, they have rarely been considered for alternate assessments. Rather, most reported evaluations of reliability for alternate assessments have focused primarily on inter-rater agreement (Browder

et al., 2003). Crawford and McDonald (2003) analyzed inter-rater agreement by counting the numeric difference between teachers' scores and the scores of a trained external rater. Each indicator was scored for each student assessment in the study for level of agreement using score as an exact match between the teacher and trained rater, and as a difference of 1 on a 5-point scale. They also examined scores that differed by 2 points, scores that differed by 3 points, score that differed by 4 points, and indicators for which no score was recorded by at least one of the raters. Totals were calculated and percentage of agreement for all indicators recorded was recorded.

When alternate assessments are judged against alternate achievement standards, it may be necessary to address local item dependency when student scores can reflect both "real" student performance and the level of assistance provided by the individual administering the assessment. Tests must also be reliable to be valid measures of a student's performance. An increase in the number of items (tasks) to be assessed might make activities more homogenous (Dunbar, Koretz, & Hoover, 1991) and, therefore, influence the degree of reliability (Fahey, Filbin, & Connolly, 2004). For example, the Colorado Student Assessment Program Alternate (CSAPA) was built by employing standardized assessment items and materials. In addition, numerous adaptations were accepted during the administration of the CSAPA. Content scaffolding, systematic instructional strategies, and the scoring of indicators depended on the teacher. To gauge the impact of these factors on score variance, the correlation between individual items (internal consistency) was calculated using Cronbach's alpha. The reliability coefficients spanned from .96 to .99.

Rudner and Schafer (2001) reported that most large-scale assessment coefficients ranged between .80 and .90. Using their yardstick the coefficients for the CSAPA suggested very strong internal consistency. They observed, however, that it was difficult to be sure that the strong coefficients were indicative of the homogeneity of the items and not a result of integrating skill performance and level of prompting. In addition, the strong coefficients might suggest a restriction in the range of items, since the majority of students taking the CSAPA scored in the upper two performance categories. They concluded that the area warranted further investigation to assess effects of combining accuracy and degree of support on reliability measures.

Reliability of an Assessment System

An assessment system may comprise a number of alternatives, all with the goal of reaching a conclusion for the student. It is a good idea to see how the various components supplement each other and how reliable the system is toward the end. An assessment system is a combination of various assessment approaches that provide a full range of participation methods. Two definitions of "system" apply here:

- A complex whole formed from related parts: a combination of related parts organized into a complex whole, a social system (MSN Encarta, n.d.).
- A set of things working together as a mechanism or interconnecting network. (Compact Oxford English Dictionary of Current English, Third Edition, 2005).

An example of an assessment system is described on the Washington Department of Education Web site:

> The Washington State Assessment System (WSAS) is composed of three broad programs: statewide standardized testing; classroom-based assessments; and assessment staff development. The statewide testing program focuses on the Essential Academic Learning Requirements (EALRs), which are Washington's content standards, and provides broad achievement indicators for the state, districts, schools, and individual students. . . . The Washington Assessment of Student Learning (WASL) currently is comprised of a series of criterion-reference tests in reading, writing, mathematics, and science. These standards based assessments incorporate three item types: selected response (multiple-choice); short constructed response; and extended constructed response. Performance standards for the assessments in reading, writing, mathematics, and science have been set using an item mapping technique designed after that developed by researchers at CTB/ McGraw-Hill. (Office of the Superintendent of Public Instruction, Washington Department of Education, 2004)

In this example, one aspect of the system contains three components: standardized testing, classroom-based assessments, and staff development. Within this broader system, the statewide assessment of achievement (the WASL) contains three formats across four subjects. In this paper, the concept of assessment system is extended beyond the subject areas and assessment formats to include assessment methods. A number of issues appeared in this expanded view:

- Does one component of the system support the inferences of another n the conjunctive/compensatory sense? That is, do the separate decisions within each component have to be present at some level to be counted toward the whole, or can some components of the system be used to compensate for others when coming to judgment?
- Must the technical adequacy among the elements of the system be comparable—whether made up of selected response, short constructed response, and extended constructed response, as in the Washington model, or made up of the general assessment with or without accommodations and the three forms of alternate assessment (based on grade level, modified achievement, or alternate achievement standards) as presented in the decision framework?
- Regarding this system's reliability, how does the weakest link—that is, least reliable or least replicable component—affect the entire system? Do estimates of reliability within one component compromise confidence in the other components? How are the reliabilities of constructed responses considered along with the reliabilities of selected responses (given that they are often lower)? How is the value of each component weighted in relation to the time spent in collecting the information and the resulting score?
- How do we assess fairness when some components of the system warrant greater confidence than others in the inferences that can be made?
- When determining Adequate Yearly Progress of individual student proficiency, do judgments based on components of the system benefit or disadvantage some subgroups more than others?

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Boomsma, A., Van Duijn, M. A. J., & Snijders, T. A. B. (2001*). Essays on item response theory.* New York: Springer-Verlag.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Browder, D. M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children, 70*(1), 45–61.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston: Pearson Education.

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005). 2003 state policies on assessment participation and accommodations for students with disabilities (Synthesis Report 56). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved October 20, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html

Compact Oxford English Dictionary of Current English, Third Edition. (2005, June). Retrieved July 2, 2005, from http://www.askoxford.com/concise_oed.

Crawford, L., & McDonald, M. (2003). *A reliability study of the 2002 administration of the third grade Colorado State Assessment Program Alternate (CSAPA) assessment in reading and writing.* Denver, CO: Colorado Department of Education Report.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Fahey, K., Filbin, J., & Connolly, T. (2004). *Colorado's performance-based alternate assessment: Historical perspective and lessons learned technical report.* Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Fleming, J. A., Taylor, J. M., & Carran, D. (2004). A comparison of two methods of determining interrater reliability. *Assessment for Effective Intervention*, *29*(2), 39–51.

Haertel, E. H. (October, 1999). *Reliability: How to quantify error and how to communicate about it.* Handout at the Interactive Lecture Series. Dover, NH: National Center for the Improvement of Educational Assessment.

Hollenbeck, K., & Tindal, G. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment*, *6*(1), 23–40.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

McGrew, K. S., Johnson, D. R., Cosio, A., & Evans, J. (2003). *Increasing the chance of no child being left behind: Beyond cognitive and achievement abilities*. Unpublished manuscript, University of Minnesota.

MSN Encarta. (n.d.).

Nitko, J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs, NJ: Merrill.

Office of the Superintendent of Public Instruction, Washington Department of Education. (2004, March). *Assessment.* Olympia, WA: Author. Retrieved October 24, 2005, from http://www.k12.wa.us/assessment/default.aspx

Oregon Department of Education. (2005). *Juried assessment 2004–2005: Guidelines for meeting Oregon's standards using the juried assessment process & guidelines for jurying a modification.* Salem, OR: Author. Retrieved June 15, 2005, from http://www.ode.state.or.us/teachlearn/testing/admin/juried/juriedassmtmanual0405.pdf

Rudner, L. M., & Schafer, W. D. (2001). *Reliability* (ERIC Digest). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED458213). Retrieved January 26, 2004, from http://www.ericfacility.net/databases/ERIC_Digests/ed458213.html

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literatur*e (Center for Educational Assessment Research Report No. 485). Amherst: School of Education, University of Massachusetts.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Hillsdale, NJ: Erlbaum.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Washington State Department of Education. Alternate Assessment System. Retrieved December 5, 2005, from http://www.k12.wa.us/assessment/altassess.aspx.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37–42.

Yovanoff, P., & Tindal, G. (in press). Scaling early reading alternate assessments with statewide measures. *Exceptional Children*.

This document was produced under U.S. Department of Education Contract No. EDO4CO0025/0002 with the American Institutes for Research. Renee Bradley served as the contracting officer's representative. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this report or on Web sites referred to in this report is intended or should be inferred.