# Reliability of Alternate Assessments

Gerald Tindal
Paul Yovanoff
Patricia Almond

University of Oregon

## *Reliability of Alternate Assessments*

The purpose of this chapter is to define and describe reliability as it pertains to the consistency or stability of scores assigned to students. This consistency and stability is usually considered in the context of multiple replications of a test. When testing students, we want the scores that result from our test administration to consistently reflect student ability or skill; only then can we trust their accuracy. In fact, if we cannot trust score consistency or stability, we cannot make valid interpretations. This is another way of relating reliability with validity (the interpretation of results and the decision-making process). Reliability sets the upper limit for validity. "Although reliability is discussed here as an independent characteristic of test scores, it should be recognized that the level of reliability of [any] score has implications for the validity of score interpretations. Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores." (AERA, APA, & NCME, 1999, p. 31).

"The hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score for the procedure is called measurement error" (AERA, APA, & NCME, 1999, p. 25) and "it limits the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process" (p. 27). Two general types of variation are present in all measurements: (a) systematic, and (b) random. Systematic variation may be explainable and thought to be operating on all objects (persons) being measured. Random variation often is unexplainable and to the extent that it is present, then measurement reliability is compromised. It is the random variation that is the focus of our reliability analysis and is usually attributed to error (which can be parsed further in generalizability theory).

The following topics are addressed in this chapter: (a) considering students and measurement approaches as sources of error, (b) documenting types of reliability coefficients to report (internal consistency, alternate or parallel forms, test – re-test, and inter-judge), (c) documenting consistency to make reliable decisions (particularly as it pertains to calculating the standard error of measure [SEM] that is then used to estimate the accuracy of a score or classification), and finally, (d) providing an example using a performance task from the Oregon Extended Assessment. These five topics weave together in moving from an analysis of the manner in which data are collected, which defines potential sources of error, to the documentation of this information (expressed as a reliability coefficient) that is then used to analyze decision-making accuracy.

## *Students and Measurement Approaches as Sources of Error*

It is assumed that error comes from a variety of sources given that measurement often is only one score obtained at one point in time with a fixed sample of items or tasks. Generally, this error can come from students or the test itself (how it is constructed, administered, or scored). The *Standards* refer to these two sources as "rooted within the examinees or those external to them" (AERA, APA, NCME, p. 26). For example, students can introduce error from variations in their level of fatigue, motivation, and interest, as well as level and type of access skills that facilitate or impede performance (sensory impairments or disabilities). Whatever the source of this random error, it represents "the difference between any observed score and corresponding true

score for each examinee . . . random error can be large or small, positive or negative" (Haladyna & Downing, 2004, p. 18). The test itself also can be a source of error. Any test is a finite collection of items and the degree to which they have been appropriately sampled and comparably formatted and administered may introduce a source of error that results in inconsistent performance estimates. This sampling, formatting, and administering may result in inconsistently difficult or easy items. Finally, the manner in which the test is scored can introduce unsystematic error (e.g., the training of judges is not adequate, the scoring keys are inconsistent, or student responses are miscoded).

For most large-scale assessments, measurement error is related to the measurement approach. In tests for students with significant disabilities, three approaches are commonly used (even though most large-scale tests use multiple-choice tests): (a) portfolios (or collections of evidence), (b) performance tasks, and (c) observations with rating scales. For this reason, the type of reliability that is documented usually needs to be specifically related to the type of measurement approach to help identify the potential sources of error that might be present. As noted in the *Standards* (AERA, APA, & NCME, 1999), the form of reliability needs to be specific to the measurement approach and the reliability coefficient needs to reflect the appropriate source of error. Two standards apply:

> *Standard 2.4*. "Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method" (p. 32).
>
> *Standards 2.5*. "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" (p. 32).

*Error from the Student*

At this time, little research documents sources of error from students in most large-scale alternate assessments. Though traditional measurement books describe this source of error as important, it is difficult to document for students, particularly those with the most significant disabilities. In general, unsystematic error occurs when students' attention fluctuates, their interest wanes while taking a test, or their preparation for multiple-day examinations is different (in terms of sleep, nutrition, and motivation). This type of error needs to be distinguished from "the systematic factors that may differentially affect the performance of individual test takers [which] are not as easily detected or overridden as those affecting groups . . . the individual systematic errors are not generally regarded as an element that contributes to unreliability. Rather they constitute a source of construct-irrelevant variance and thus may detract from validity" (AERA, APA, & NCME, p. 26).

*Error from the Measurement Approach (Construction, Administration, and Scoring)*

When considering the test or measure as a source of error, reliability analyses need to begin by taking into account the type of measurement (approach) being implemented. Each of these approaches has the potential to introduce various sources of random error.

In a portfolio or any collection of evidence, error is most likely to arise from an uneven sampling of evidence (some work samples are difficult while others are easy) or from scoring (rating) student work samples (some work samples may be rated with no specific reference while other work samples reflect discrete counts of items completed correctly or incorrectly). Therefore, the number of "events" may differ between portfolios and performance assessments. For example, portfolios may include more documents (given that they are easier to collect and are done so over a longer interval) while performance assessments contain a limited number of samples collected over a more circumscribed time frame. In addition, administration of either portfolios or performance tasks may not be optimal, though collections of evidence may be more flexible in the way that the test is "administered."

Many of these differences between portfolios and performance assessments apply equally well to observations (whether they reflect counts of behaviors or ratings/judgments). Yet, because observations are done in the field and are conducted in the presence of the student performing a task), other (and unique) issues may serve as sources of random error. For example, the difficulty of the task may influence performance (as in both portfolio and performance assessments), as well as the directions and support provided to the student, which can directly affect the students' performance. Although this source of error may be present in both portfolios and performances, it usually is not possible to address because it is not observed. If the score of the observation is based on an interval schedule (reflecting frequencies in which a behavior is coded), reliability may be a function of the interval size as well as the definitions of the behavior. In summary, different sources of random error may come from either the student or the test (development, administration, or scoring) and each measurement approach presents different ways that this error appears.

### *Types of Reliability Coefficients to Report*

"The critical information on reliability includes the identification of the major sources of error, summary statistics bearing on the size of such errors, and the degree of generalizability of scores across alternate forms, scorers, administrations, or other relevant dimensions" (AERA, APA, NCME, 1999, p. 27) and a clear description of the examinee population to which the reliability data apply. Four types of reliability coefficients traditionally can be been reported and ideally are selected according to the measurement approach and the (potential) sources of error:

> *Internal consistency* summarizes the manner in which items are correlated within a test: how well each item correlates with the total test (or the degree to which alternate forms can be created internally by comparing odd and even items or the first half of the test with the second half of the test, reflecting two strategies for dividing a test). Internal consistency can be summarized by Cronbach's alpha or KR20. In the split half strategy, a simple correlation coefficient is calculated between the two halves (which then needs to be adjusted using the Spearman Brown Prophecy formula to determine what the coefficient would have been if the full test had been given).
>
> *Alternate (parallel) form* reliability provides an index of consistency across two or more forms of a test and is critical if multiple forms exist. We use the term alternate or parallel forms to reflect two versions of the same test being administered. Of course, we would

want to randomly assign the order of forms so that we would not confound *form* with the *order of administration* (e.g., Form B is always given first and therefore is lower or higher because of a fatigue or novelty effect, respectively). A simple correlation coefficient is calculated as the reliability index.

*Test-retest* reliability focuses on the sameness of score from one time to another when the exact same assessment is given over a short time interval. Of course, we would not want the interval of separation to be too great so that learning or other factors interfere with the score value being estimated. This reliability is usually summarized as a correlation coefficient in which students are compared in their rank orders on these two occasions.

*Interjudge* (or inter-rater) reliability addresses the degree to which different judges evaluate or rate performance consistently. This type of reliability is usually summarized as percent of agreement and usually is important only if the test score reflects a subjective judgment; if the test is scored using a selected response with only one correct answer, we are not usually very interested in this form of reliability. In many states, this agreement is either exact or "off-by-one". Of course, the latter may actually miss the whole point if two scores are collapsed as only "off-by-one," but they are at the cut score (in effect, the two judges disagree about the whether or not the student "meets" or "does not meet" the standard).

With each measurement approach susceptible to different sources of error, different reliability coefficients need to be considered. Internal consistency reliability is probably the most critical across all measurement approaches. Though it typically is the most frequently used type in technical manuals of general education tests, it rarely is presented in technical manuals from alternate assessments. Alternate form reliability also is not typically reported for alternate assessments, primarily because only one form is administered; yet, it may be important if changes are made from year to year. Test-retest is almost never considered in alternate assessments, presumably because of the population of students. Finally, what predominates is inter-rater or interjudge reliability, perhaps because of the popularity of portfolios as the dominant measurement approach with alternate assessments.

## *Beyond Documentation of Consistency: Making Reliable Decisions*

Although documentation of reliability coefficients is important in quantifying specific aspects of consistency given the potential for various sources of error associated with a measurement approach, this step is rarely the last or most important one. Rather, it is the impact of consistency on estimates of true scores and classification of students that needs to be considered as the final step in understanding reliability. It is the impact of consistency on accuracy that is important; with consistent scores, estimates of performance can be more accurate.

### *Reliability Used to Calculate Standard Error of Measurement and Estimate True Scores*

Although reliability coefficients communicate information on the consistency of scores, it is important to take a further step and focus on the impact on interpretations. As noted in the *Standards* (1999), "the standard error of measurement is generally more relevant than the reliability coefficient once a measurement procedure has been adopted and interpretation of

scores has become the user's primary concern. (p. 29). And, like the reliability coefficients, the same ambiguities for interpretation of SEM appear.

Classical test score theory states that observed scores are comprised of true score and error score. With procedures available to determine the degree of error, it should be possible to estimate the true score. This estimate of measurement error around the true score results in a confidence interval in which the true score is likely to appear. If we could eliminate all the error in a test, then the observed score would be equal to the true score. However, because this is impossible, we need to estimate the error and then use our calculation to predict the true score. In this estimation, we can focus on either the average error for all scores in the distribution (standard error of measurement) or the error associated with one specific score value in the distribution (conditional standard error of measure).

How should the standard error of measurement be used in making interpretations from state tests, or any tests for that matter? Probably the best way to define the standard error of measure is to see how it is calculated: $\sigma_e = \sigma_x \sqrt{r_{x_1 x_2}} \sqrt{1 - r_{x_1 x_2}}$ (1)
In this formula, the SEM is a function of the variance of the test ($\sigma_x$) and the correlation between parallel forms ($\sqrt{r_{x_1 x_2}}$). In this formula of SEM, it should be very clear that, as the correlation between these parallel forms increases it (the SEM) decreases and eventually (theoretically) becomes zero. In the *Standards* (1999), are found several references to reliability and SEM:

> *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 unit and in units of each derived score recommended for use in test interpretation" (p. 31).

*Reliability Used to Estimate the Accuracy of a Score or Classification of Performance*
This emphasis on SEM is particularly critical when considering achievement levels that have been demarcated into groups using cut scores (e.g., exceeds, meets, does not meet, and far below meets). Whereas, the "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 standards" (AERA, APA, & NCME, 1999, p. 29). In standards-based assessments, it is the absolute interpretation that counts.

> *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).

The problem, however, is that these two interpretations are interdependent in that the greatest precision for an absolute decision is needed not at the extremes of a relative distribution but somewhere within the middle at the cut score.

In summary, reliability is extremely important in standards-based testing in three ways. First, with large-scale testing, the system is so complex that error can enter from a number of sources. With item development being so integral to a range of standards at multiple grade levels and with so many teachers and students taking part in the testing program, we need to be confident that error is not entering the system from students when they arrive at the test or from the test itself (construction, administration, or scoring). Second, given the high stakes associated with current standards-based assessments (i.e., graduation or some other important decision being made on the basis of test scores), "the need for precision increases as the consequences of decisions and interpretations grow in importance" (p. 30). We need this error to be minimal at the cut score. Otherwise, we would be making false decisions in either of two ways: (a) we would be failing students who really should be passing or (b) we would be passing students who really should be failing. Generally, in the public schools, greater concern is with the former (considered a false negative). Finally, if we wish to hedge our interpretations, we should compute the standard error of measure to define an interval within which we would be confident that the true score would be located. This confidence interval usually is expressed with a lower and upper bound for either of two levels of confidence: 68% or 95%.

One caveat should be noted about the discussion of reliability in this chapter based on a classical definition of reliability. We focused on replicated forms, which is somewhat different than one based on item response theory (IRT) where items are calibrated on two dimensions: (a) difficulty and (b) discrimination. In this view, items and tests have varying amounts of information, which in turn is a function of student ability. The amount of measurement error associated with a test depends on the student's ability level. If a student of very low ability is administered an extremely difficult item, we do not gain much information; the same is true if a student of high ability is administered a very easy item. To avoid this situation, then, tests usually have a range of items so students of varying ability have items available to answer. By scaling item difficulty and person ability on the same scale, we can learn much about the information provided by both the items and the test.

This discussion of sources of error that need to be uniquely considered in relation to the measurement approach can now be operationalized with an illustrative alternate assessment based on performance tasks. In this example, the performance event is first described and then various reliability coefficients are presented. Eventually, had we administered all of the performance events in this alternate assessment, it would be possible to compute an overall reliability coefficent which could then be used to calculate the standard error of measurement and a confidence interval within which a true score would be located.

*An Example from a Performance Assessment from Oregon Extended Assessment*

In the following study, the data come from a statewide alternate assessment administered in the 2004-05 school year.

*Sources of error*. The task involves a student reading from a list of words. The manner in which the test is administered could be a major source of error. In Figure 1 below, notice that a source of error could be the sample of words (and reflect low internal consistency or have different alternate forms) or the manner in which it is administered (e.g., pointing can be used). Finally, because of partial scoring (0, 1, and 2), error could enter the results.

*Figure 1. Sample Reading Word Task*[a] *(with 8 words printed on flash cards)*

READING WORDS:
Present the cards in the order shown in the left hand column of the table below. Place the words in a stack on the table in front of the student and say, **"Read each word as I show you the card."** Continue presenting words. Prompt the student after a delay with no response.

POINTING to WORDS:
If the student cannot identify the words using expressive communication (speech, sign language, or communication device), follow these directions: Randomly place all of the words face up on the table and say, **"Point to the word after I say it."** Continue saying words in the order listed in the table below. Prompt student after a delay with no response. Record the student's points in the table. If the student responds incorrectly, record his or her response.

| Points for Reading: Word completely correct = 2<br>ANY correct sound = 1<br>Incorrect = 0 | Points for Pointing: Correctly pointed to word = 2<br>Incorrectly pointed to word = 0 |
|---|---|

[a] *Participation (n=463): 2 Modified, 17 Not Administered Inappropriate, 444 Standard / 18 used Pointing responses*

This test was given to the following students in March and April. See Table 1.

*Table 1. Description of the Grade 5 Population*

| Disability | Frequency | % | Valid % | Cumulative % |
|---|---|---|---|---|
| Mental Retardation | 149 | 32.2 | 33.6 | 33.6 |
| Hearing Impairment | 6 | 1.3 | 1.4 | 35.0 |
| Vision Impairment | 1 | .2 | .2 | 35.2 |
| Speech-Language | 54 | 11.7 | 12.2 | 47.4 |
| Emotional Disturbance | 9 | 1.9 | 2.0 | 49.4 |
| Orthopedic Impairment | 10 | 2.2 | 2.3 | 51.7 |
| Traumatic Brain Injury | 6 | 1.3 | 1.4 | 53.0 |
| Other Health Impairment | 47 | 10.2 | 10.6 | 63.7 |
| Autism | 55 | 11.9 | 12.4 | 76.1 |
| Severe Learning Disability | 106 | 22.9 | 23.9 | 100.0 |
| Total | 443 | 95.7 | 100.0 | |
| Missing | 20 | 4.3 | | |
| Total | 463 | 100.0 | | |

The data file reflects 8 items (words) with partial scoring (0, 1, or 2) that was split into 2 halves (first half and second half). This could have been divided into odd and even items. In either case, the focus of this reliability is internal consistency. See Table 2.

*Table 2. Extended Reading (XR) Data File: Administration and Format for First Three Records of 463 for Items 1-8 and First-Second Half*

| ADMIN | Format | XR_1 | XR_2 | XR_3 | XR_4 | XR_5 | XR_6 | XR_7 | XR_8 | First | Sec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| STD | Naming | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 8 | 8 |
| STD | Naming | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| STD | Pointing | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 8 | 8 |

The results of the test reflected the following frequencies of different scores. See Table 3.

*Table 3. Descriptive Statistics of Extended Reading (XR) for Items 1-8*

| Results | XR_1 | XR_2 | XR_3 | XR_4 | XR_5 | XR_6 | XR_7 | XR_8 |
|---|---|---|---|---|---|---|---|---|
| No. Blanks | 32 | 29 | 28 | 30 | 31 | 31 | 33 | 31 |
| No. Scored 0 | 174 | 29 | 77 | 36 | 58 | 98 | 103 | 192 |
| No. Scored 1 | 125 | 19 | 127 | 34 | 50 | 107 | 144 | 126 |
| No. Scored 2 | 132 | 386 | 231 | 363 | 324 | 227 | 183 | 114 |
| Total Count | 463 | 463 | 463 | 463 | 463 | 463 | 463 | 463 |

| Sum | 389 | 791 | 589 | 760 | 698 | 561 | 510 | 354 |
|---|---|---|---|---|---|---|---|---|

The following item level statistics were computed, showing the average, standard deviation (amount of variation), and number of cases or students. See Table 4.

*Table 4. Item Level Data*

| | *Mean* | *Std Dev* | *Cases* |
|---|---|---|---|
| XR_1 | .9038 | .8388 | 426.0 |
| XR_2 | 1.8310 | .5130 | 426.0 |
| XR_3 | 1.3592 | .7609 | 426.0 |
| XR_4 | 1.7582 | .5903 | 426.0 |
| XR_5 | 1.6080 | .7154 | 426.0 |
| XR_6 | 1.2981 | .8162 | 426.0 |
| XR_7 | 1.1831 | .7966 | 426.0 |
| XR_8 | .8192 | .8190 | 426.0 |
| Scale | 10.7606 | 4.0513 | |

The following statistics were calculated for the total task (all 8 words), showing the average across all 8 items, the average minimum and maximum scores across the 8 words, and ratio of these two (max:min), and the variance. See Table 5.

*Table 5. Average Performance on the Task*

| Item Means | Mean | Min | Max | Range | Max/Min | Variance |
|---|---|---|---|---|---|---|
| | 1.3451 | .8192 | 1.8310 | 1.0117 | 2.2350 | .1394 |

Finally, various reliability coefficients were computed based on the item level data noted above. See Table 6 for split half (odd-even or first-second half which was adjusted because of the reduced number of items used in its calculation), Cronbach's alpha (the average relation between each item and the total), and parallel form (considering each half as a form).

*Table 6. Reliability Coefficients*

*Split Half*
- Correlation between forms = .70
- Equal-length Spearman-Brown = .83
- Guttman Split-half = .81
- Unequal-length Spearman-Brown = .83

*Cronbach's Alpha*
- Part 1 (4 items) = .67
- Part 2 (4 items) = .77

*Parallel Form*
- Estimated reliability of scale = .84
- Unbiased estimate of reliability = .84

The results indicate that reading words is a reliable performance assessment with little difference in the manner in which consistency is noted: (a) split half, (b) Cronbach's alpha, or (c) parallel forms.

*References*

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

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.

Haladyna, T., & Downing, S. (2002). Construct irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practices*, *23*(1), 17-27.

Quenemoen, R., Thompson, S. & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria* (Synthesis Report 50). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 3, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html

Thompson, S., & Thurlow, M. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.