

CHAPTER 16: RELIABILITY

Reliability is the most critical technical characteristic of any measurement system. The reliability of the scores resulting from an assessment should be demonstrated before issues such as validity, fairness, and interpretability can be discussed. Reliability is an expression of how well an assessment measures actual learning. Because the TAKS, SDAA II, RPTE, EOC, TAKS-Alt and TAAS exit level assessments can provide only estimates of achievement levels, their scores contain a certain amount of error; test reliability measurements quantify this error. There are many different methods for estimating test reliability. For a thorough discussion of test reliability, see *Introduction to Classical and Modern Test Theory* (Crocker & Algina, 1986).

Internal Consistency Estimates

Test reliability is an indication of the consistency of the assessment. TAKS, SDAA II, RPTE, EOC, and TAAS exit level test reliability data are based on internal consistency measures. These include, in particular on the Kuder-Richardson Formula 20 (KR20) for tests involving dichotomously scored (multiple-choice) items and the stratified coefficient alpha for TAKS tests involving a combination of dichotomous and polytomous (short-answer and extended response) items. Most internal consistency reliabilities are in the high .80s to low .90s range (1.0 being perfectly reliable), with reliabilities for TAKS assessments ranging from .83 to .93, for SDAA II assessments ranging from .71 to .86, and for RPTE assessments ranging from .93 to .94. The reliability for the Algebra I EOC Assessment was .92. (Note: SDAA II tests were lengthened in 2004–2005 to increase reliabilities. However, reliabilities may still be lower on some SDAA II tests such as those for grades K–2 because they are shorter to reduce the burden on this population of students.)

Appendix C presents reliability estimates for all content areas and objectives, for all students as well as for major demographic groups. Included in this appendix are summary statistics (N-count, mean, standard deviation, number of items) and related statistics such as the standard error of measurement and mean p-value.

Procedures Used

The KR20 is a mathematical expression of the classical test theory definition of test reliability. This definition expresses test reliability as the ratio of true score variance to observed score variance (test performance); it is generally expressed symbolically as the following:

$$P_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2},$$

where the reliability, $P_{XX'}$, of test X is a function of the ratio between true score variance, σ_T^2 , and observed score variance, σ_X^2 . Observed score variance is defined as the combination of true score variance and error variance, σ_E^2 . As error variance is reduced, reliability increases

(that is, students' observed scores are more reflective of students' true scores or actual proficiencies). The internal consistency estimate of this reliability can be mathematically represented as

$$KR20 = \left[\frac{k}{k-1} \right] \left[\frac{\sigma_x^2 - \sum_{i=1}^k p_i (1-p_i)}{\sigma_x^2} \right],$$

where $KR20$ is a lower-bound estimate of the true reliability, k is the number of items in test X , σ_x^2 is the observed score variance of test X , and p_i is the proportion of students who got item i correct (that is, the item p -value). This formula is used when test items are scored dichotomously.

Coefficient alpha (also known as Cronbach's alpha) is an extension of $KR20$ to cases where items are scored polytomously (into more than two categories) and is computed as follows:

$$\alpha = \left[\frac{k}{k-1} \right] \left[1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_x^2} \right],$$

where α is a lower-bound estimate of the true reliability, k is the number of items in test X , σ_x^2 is the observed score variance of test X , and σ_i^2 is the variance of item i .

The stratified coefficient alpha is a further extension of coefficient alpha used when a mixture of item types appears on the same test. In computing the stratified coefficient alpha as an estimate of reliability, each item type component (multiple-choice, open-ended, or essay) is treated as a subtest. A separate measure of internal-consistency reliability is computed for each component and combined as follows:

$$Strat\alpha = 1 - \frac{\sum_{j=1}^c \sigma_{x_j}^2 (1 - \alpha_j)}{\sigma_x^2},$$

where c is the number of item type components, α_j is the estimate of reliability for each item type component, $\sigma_{x_j}^2$ is the observed score variance for each item type component, and σ_x^2 is the observed score variance for the total score. For components consisting of multiple-choice and open-ended (short answer) items, a standard coefficient alpha (see above) is used as the estimate of component reliability. The correlation between ratings of the first two raters is used as the estimate of component reliability for essay prompts.

Although many options are available for estimating reliability of tests with a mixture of item types, the stratified coefficient alpha was deemed most appropriate for TAKS. For a more detailed research report showing the comparison of stratified coefficient alpha to other mixed-model reliability estimates, see "Determining An Appropriate Index of Reliability" in the

2007 Texas Education Agency Technical Report Series which can be found at <http://www.tea.state.tx.us/student.assessment/resources/techdig07/index.html>.

For SDAA II writing tests, KR20 estimates of reliability are reported for the multiple-choice portion of the tests, and correlations between ratings of the first two raters are reported for the essay prompts. Unlike TAKS writing, where these two measures are combined, the measures are kept separate for SDAA II writing to more accurately reflect the separation of the essay prompt from the multiple-choice items in the scoring tables. For SDAA II Instructional Level 9 reading and SDAA II Instructional Level 10 English language arts (ELA), a stratified coefficient alpha is computed that combines the reliability estimates from the multiple-choice and open-ended items. In addition, for SDAA II Instructional Level 10 ELA, the correlation between the ratings of the first two raters is reported for the essay prompt.

Classical Standard Error of Measurement

The classical standard error of measurement (SEM) is calculated using both the standard deviation and the reliability of test scores; SEM represents the amount of variance in a score resulting from factors other than achievement. The standard error of measurement assumes that underlying traits such as academic achievement cannot be measured precisely without a perfectly precise measuring instrument. For example, factors such as chance error, differential testing conditions, and imperfect test reliability can cause a student's observed score (the score achieved on a test) to fluctuate above or below his or her true score (the true proficiency of the student). The SEM is calculated as

$$\text{SEM} = \sigma_x \sqrt{1 - r},$$

where r is the reliability estimate (for example, a KR20, coefficient alpha, or stratified alpha) and σ_x is the standard deviation of test X .

It is important to note that the classical SEM index provides only an estimate of the average test score error for all students regardless of their individual proficiency levels. It is generally accepted (see, for example, Peterson, Kolen, & Hoover, 1989) that the SEM varies across the range of student proficiencies and that individual score levels on any particular test could potentially have different degrees of measurement error. For this reason, it is useful to report not only a test-level SEM estimate but individual score-level estimates as well. Individual score-level SEMs are commonly referred to as conditional standard errors of measurement (CSEMs).

Conditional Standard Error of Measurement (CSEM)

The CSEM provides an estimate of reliability that is conditional on the proficiency estimate. In other words, the CSEM provides a reliability estimate, or error estimate, at each score point. Because there is typically more information about students with scores in the middle of the score distribution, the CSEM is usually smallest, and scores are more reliable at that score level.

Item response theory methods for estimating both individual score-level CSEM and test-level SEM were used because test- and item-level difficulties for TAKS, SDAA II, RPTE, EOC, and TAAS exit level tests are calibrated using the Rasch measurement model. The standard error of each test is calculated as the average conditional standard error across all students. SDAA II tests report CSEM in terms of raw score units whereas TAKS, RPTE, EOC, and TAAS exit level report CSEM in terms of scale score units.

For SDAA II tests, SEM estimates were calculated using the following steps, which average CSEMs across the entire test (Lord, 1980). First, the error variance for a given raw score level was defined as the sum of the conditional Bernoulli variance of each item (item i) at the given raw score level of k . It was calculated as

$$\sigma_{e|\xi_k}^2 = \sum_{i=1}^n P_i(\theta_k) Q_i(\theta_k),$$

where the error variance, $\sigma_{e|\xi_k}^2$, at a raw score level of k is equal to the sum of the conditional Bernoulli variance of a student with a proficiency level θ_k on the total test of n items.

Under the assumptions of the Rasch measurement model, each raw score level, k , is associated with only one corresponding proficiency estimate, θ_k . For every value of θ_k a specific probability of responding correctly to each item i , P_i , can be defined as a function of

$$P_i(\theta_k) = \frac{1}{1 + e^{-(\theta_k - b_i)}},$$

where b_i is the difficulty of item i . Further, for every value θ_k , the probability of incorrectly responding to item i is defined as $1 - P_i(\theta_k)$, or $Q_i(\theta_k)$. By taking the square root of the error variance $\sigma_{e|\xi_k}^2$, the CSEM at each raw score level was obtained. (Note: The conditional standard error of measurement is an estimated index of reliability of a student's raw score at a specific score level.)

The classical test theory SEM of the raw scores for the test was the square root of the simple mean of the CSEMs over all N students, or:

$$s_{e,\xi} = \sqrt{\frac{1}{N} \sum \sigma_{e|\xi}^2}.$$

For TAKS, RPTE, EOC, and TAAS exit level tests, CSEMs were estimated for scale scores by first calculating the standard errors for each student proficiency, θ_k , corresponding to each raw score level, k . Proficiency estimate SEMs are inversely related to the root test information function at a given level of student proficiency (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). The test information function is an additive composite that quantifies the psychometric information of each item at every point along the student proficiency distribution. As indicated above, each raw score level has with it only one corresponding proficiency estimate, θ_k . The test information function at a given level of proficiency is calculated as

$$TI(\theta_k) = \sum_{i=1}^n P_i(\theta_k) Q_i(\theta_k),$$

where $P_i(\theta_k)$ is the probability of correctly responding to item i at proficiency k and $Q_i(\theta_k)$ is the probability of incorrectly responding to item i at proficiency k . (Note that the test information function and the raw score error variance at a given level of proficiency, θ_k , are analogous for the Rasch model). The CSEM at a given level of proficiency, θ_k , is simply the root inverse of the test information function at θ_k and is calculated as

$$SE_{\theta_k} = \frac{1}{\sqrt{TI(\theta_k)}}.$$

Finally, the SEM of the proficiency estimates for the total test was calculated as the mean CSEM across all N students, or:

$$SE_{\theta} = \frac{1}{N} \sum SE_{\theta_k}.$$

Because TAKS, RPTE and EOC results are not reported in terms of Rasch proficiency estimates but, instead, are reported in terms of scale scores, proficiency estimate CSEMs had to be converted to a scale score metric. Scale scores reported for TAKS, RPTE and EOC are linear transformations of the underlying proficiency estimates. As such, scale score CSEMs are simply a multiple of the proficiency estimate CSEMs (Kolen, Hanson, & Brennan, 1992). This conversion was made based on the same linear transformation used to convert proficiency estimates to scale scores and is calculated as

$$SE_{SS_k} = (SE_{\theta_k} \times T_1),$$

where SE_{SS_k} is the conditional standard error of measurement of the scale score at proficiency k , SE_{θ_k} is the conditional standard error of measurement at the proficiency level k , and T_1 is the multiplicative scale score transformation constant (see Chapter 13: Sampling, Tables 14 and 15).

Appendix D provides conditional standard errors of measurement for all TAKS, RPTE, SDAA II, and EOC tests. CSEMs are provided for the primary administration of each test only.

Use of the Standard Error of Measurement

The standard error of measurement is helpful for quantifying the margin of uncertainty that occurs on every test. It is particularly useful for estimating a student's true score, which is assumed to fall within one standard error of measurement of the observed score 68% of the time (when errors are normally distributed). Unless the test is perfectly reliable, a student's observed score and true score will differ. A standard error of measurement band placed around an observed score will result in a range of values that will most likely contain the student's true score. For example, suppose a student achieves a scale score of 2025 on a test with a SEM of 50. Placing a one-SEM band around this student's score would result in a scale score range of 1975 to 2075. Furthermore, if it is assumed that the errors are normally distributed, it is likely that across repeated testing occasions, this student's true score would fall in this band 68% of the time. Put differently, if this student took the test 100 times, he or she would be expected to achieve a scale score between 1975 and 2075 about 68 times.

As stated above, the problem with using the standard error of measurement to quantify the margin of error around any individual student's scale score is that it assumes that errors are the same at every scale score level. SEMs are weighted averages of the error associated with each scale score level. By using CSEMs, which are specific to each scale score level, a more precise error band can be placed around a student's observed score. For example, suppose the CSEM of 2025 is smaller than the SEM, say, 42 as compared to 50. Placing a one-CSEM band around this student's score would result in a scale score range of 1983 to 2067. The smaller CSEM at scale score 2025 in this example demonstrates that a scale score estimate of 2025 on this test has less range of error than the average error of the test.

Appendix E provides the reliabilities and SEMs for all subject areas and objectives and for major demographic groups.

Classification Accuracy

Every test administration will result in some error in classifying students' results. Several elements of test construction and cut score determination procedures can reduce these errors. It is important to understand the expected degree of misclassification prior to approval of the final cut scores. To this end, Pearson Educational Measurement (PEM) conducted an analysis of the accuracy in student classifications into performance categories based on test results from the TAKS, RPTE, SDAA II, and EOC tests.

Common procedures for estimating classification accuracy are based on classical test theory conceptualizations of error distributions. However, the TAKS, RPTE, and EOC scale scores are reported and equated using the Rasch model, which does not use classical test theory model assumptions about the shape of the error distribution. (Note that although SDAA II results are reported in terms of raw score levels, raw score cut points are equated using the Rasch model.)

Other recommended procedures that use Item Response Theory, of which the Rasch model is an example, assume either that scaled student proficiency scores will not be reported or that the final score distribution will be normalized, neither of which applies to TAKS, RPTE, SDAA II, and EOC. The procedures used for these tests are similar to those recommended by Rudner (2001, 2005), with modification for use in these special cases.

Under the Rasch model, for a given true proficiency score, θ , the observed proficiency score, $\hat{\theta}$, is expected to be normally distributed with a mean of θ and a standard deviation of $SE(\theta)$. Using this information for a particular level, k , the expected proportion of all students that have a true proficiency score between c and d and an observed proficiency score between a and b is:

$$PropLevel_k = \sum_{\theta=c}^d \left(\phi \left(\frac{b-\theta}{SE(\theta)} \right) - \phi \left(\frac{a-\theta}{SE(\theta)} \right) \right) \phi \left(\frac{\theta-\mu}{\sigma} \right),$$

where ϕ are the cumulative normal distribution functions at the observed score boundaries, and ϕ is the normal density associated with the true score (Rudner, 2005).

This formula was modified for the current case in the following ways:

1. ϕ was replaced with the observed frequency distribution. This is necessary because the Rasch model preserves the shape of the distribution, which is not necessarily normally distributed.
2. The lower bound for lowest performance category (Did Not Meet Standard for TAKS and EOC, beginning for RPTE, and Achievement Level I for SDAA II) and the upper bound for highest performance category (Commended Performance for TAKS and EOC, advanced high for RPTE, and Achievement Level III for SDAA II) were replaced with extreme, but unobserved, true proficiency/raw scores in order to capture the theoretical distribution in the tails.
3. In computing the theoretical cumulative distribution, the lower bounds for the Met Standard performance level for TAKS and EOC, the intermediate and advanced performance levels for RPTE, and Achievement Level II for SDAA II were used as the upper bounds for the adjacent lower levels, even though under the Rasch model there are no observed true proficiency scores between discrete and adjacent raw score points. This was necessary because a small proportion of the theoretical distribution exists between the observed raw scores, given that the theoretical distribution assumes a continuous function between discrete and adjacent raw score points.
4. Actual boundaries were used for person levels, as these are the current observations.

To compute classification accuracy, the proportions were computed for all cells of an “ n performance category by n performance category” classification table. The sum of the diagonal entries represents the accuracy of classification for the test. Classification accuracy rates for each TAKS, RPTE, SDAA II, and EOC grade and subject are provided in Appendix E.

Figures 12 and 13 are examples of classification accuracy values for the 2007 TAKS Exit Level Social Studies test. In each table, the rows represent the theoretical true (expected) proportions of students in each performance level, while the columns represent the observed proportions. The diagonal entries represent the agreement between expected and observed classifications. In Figure 12 there was 86.6% agreement between expected and observed classifications for students who were in the two higher levels of performance.

Figure 12. Classification Accuracy for 2007 TAKS Exit Level Social Studies

Classification	<i>Did Not Meet Standard</i>	<i>Met Standard</i>	<i>Commended Performance</i>	Expected
<i>Did Not Meet Standard</i>	2.8	1.5	0.0	4.3
<i>Met Standard</i>	0.4	62.3	7.2	69.9
<i>Commended Performance</i>	0.0	1.6	24.3	25.9
Observed	3.2	65.4	31.5	100.0

Since TAKS uses Met Standard for Adequate Yearly Progress (AYP) and exit level decision purposes, it is useful to consider decision classification accuracy on a dichotomous classification of Did Not Meet Standard versus Met Standard and above. To compute classification accuracy in this case, the cells associated with Met Standard and Commended Performance are collapsed and compared against Did Not Meet Standard. In Figure 13, it can be seen that there was 95.4% agreement in classifications for students at the Met Standard/Commended Performance levels.

Figure 13. Collapsed Classification Accuracy for 2007 TAKS Exit Level Social Studies

Classification	<i>Did Not Meet Standard</i>	<i>Met Standard/ Commended Performance</i>	Expected
<i>Did Not Meet Standard</i>	2.8	1.5	4.3
<i>Met Standard/ Commended Performance</i>	0.4	95.4	95.7
Observed	3.2	96.9	100.0

Alternate Forms Reliability Estimates

When calculating alternate forms reliability, the goal is to examine how a different set of items introduces error into the estimate. When estimating alternate forms reliability, the process involves giving a group of students alternate forms of a test on more than one occasion. To accurately estimate this reliability, testing conditions should remain the same across testing occasions. Since no representative group of students takes more than one form of the test under similar conditions during any TAKS, SDAA II, RPTE, EOC, or TAAS exit level administration, no information regarding alternate or parallel forms reliability estimates is currently available. Some students take retests; however, the retests are taken after additional

instruction is provided. The added instruction makes the testing conditions different over the occasions and makes the estimate of alternate forms reliability inaccurate.

Gathering Reliability Evidence for TAKS–Alt

As part of the process of developing the TAKS–Alt, evidence that the assessment allows for reliable observation and rating of student performance in the Texas Essential Knowledge and Skills (TEKS) was collected. Unlike other statewide assessments in Texas, TAKS–Alt is not a traditional paper-and-pencil or multiple-choice test. Instead, the assessment involves teachers observing students as they complete instructional activities that link to the grade-level TEKS curriculum. Building reliability evidence for this form of assessment requires a different approach than that used for TAKS.

To gather reliability evidence for TAKS–Alt, an inter-rater reliability approach was used. Inter-rater reliability information can be collected by having two observers rate the same student during the same instructional activity. This permits the determination of the extent of agreement between the two observers in terms of how the student was rated. This type of information will be used to provide evidence that different raters observing the same activity provide the same score for a student when using the TAKS–Alt scoring rubric.

During the TAKS–Alt field test, a sample of teachers were asked to have a second observer provide a rating of student performance on a specified essence statement and its accompanying instructional activity. If a teacher was notified that they would be participating in the inter-rater reliability study, he/she was required to serve as the first rater for the study. In addition, the teacher had to select the student he/she would observe for the study and a qualified person to be the second rater. Teachers selected an appropriate person to serve as the second rater using the TEA guidelines listed below.

TAKS–Alt testing raters should be professionals or under the supervision of professionals who hold valid education credentials such as Texas teacher certificates or permits. Those selected may include the following:

- teachers (including general, special education, and teachers for the visually and auditory impaired)
- paraprofessionals
- assessment Specialists
- speech Therapists
- occupational Therapists
- physical Therapists

Once the teacher had selected a second rater for the study, he/she notified the campus coordinator with the names of both the second rater and the student being observed. The second rater completed all the TAKS–Alt training modules before taking part in the study. After completing the observation, the second rater was asked to submit his/her rating through

an internet survey system that is separate from the TAKS–Alt system. The first and second raters were given directions not to view or discuss each other’s ratings.

Teachers within a district were randomly assigned to have the second rater observe either a mathematics or reading/ELA instructional activity. Writing, science, and social studies will be included in future inter-rater reliability studies. The second observation only had to take place for one of the four essence statements completed within a subject test for the field test. Teachers within a district were randomly assigned to one of the two state-required essence statements for the subject area in which they are conducting the study (reading or mathematics). Inter-rater reliability evidence was collected in this manner for all the state-required essence statements in reading and mathematics.

Data collection for the field test inter-rater reliability study was completed in spring 2007. Analyses included the calculation of the correlation between the two sets of ratings, the weighted kappa statistic, and agreement rates between the first and second rater. Ratings were moderately to highly related to each other with significant correlations ranging from .414 to .858. Weighted kappa values ranged from .349 to .767. Guidelines for interpreting the kappa statistic indicate that values less than or equal to .4 have fair strength of agreement, values greater than .4 have moderate strength of agreement, and values greater than .6 have good agreement. Six percent of the calculated kappa values were of fair strength, 44% were of moderate strength, and 50% were of good strength. The majority of Kappa values reported were of moderate or good strength of agreement. The agreement rates indicate that first and second raters had quite high levels of agreement, with perfect agreement rates ranging from 65% to 89%. When perfect agreement rates are combined with adjacent agreement rates, 91% to 100% of raters had the same or adjacent ratings. It is expected that results from future inter-rater reliability studies will be higher due to increased teacher experience with the assessment. For further information about these analyses refer to the “TAKS–Alt” report in the 2007 Texas Education Agency Technical Report Series which can be found at <http://www.tea.state.tx.us/student.assessment/resources/techdig07/index.html>.

The field test inter-rater reliability study is an initial step in building reliability evidence. A second study (covering all subject areas) will be conducted during the first operational administration of TAKS–Alt. After these two studies have been completed, it should be sufficient to replicate the study every two or three years.