

CHAPTER 14: RELIABILITY

Reliability is the first and foremost technical characteristic of any measurement endeavor. The reliability of the scores resulting from an assessment should be demonstrated before issues such as validity, fairness, and interpretability can be discussed. Because the TAKS, SDAA II, RPTE, and the exit level TAAS tests provide observed scores that serve as a proxy for direct measurement of underlying achievement levels, their scores contain some amount of error, and test reliability quantifies this error. There are many ways to estimate test reliability: Suppose for example, the true achievement level for a particular group of students was known. The correlation between these true scores and the observed scores obtained from TAKS, SDAA II, RPTE, or exit level TAAS would be an estimate of the reliability of the test. See *Introduction to Classical and Modern Test Theory*, Crocker and Algina (1986) for a more thorough discussion of test reliability.

Internal Consistency Estimates

Test reliability indicates the consistency of measurement. TAKS, SDAA II, RPTE, and exit level TAAS test reliabilities are based on internal consistency measures, in particular on the Kuder-Richardson Formula 20 (KR20) for tests involving dichotomously scored (multiple-choice) items and on 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 with reliabilities for TAKS assessments ranging from 0.81 to 0.93, SDAA II assessments ranging from 0.74 to 0.89, and RPTE assessments ranging from 0.93 to 0.94. For the spring 2005 live administrations, SDAA II tests were lengthened to increase reliabilities. However, reliabilities may still be lower on some SDAA II tests (K–2) as a function of test length; these tests are shorter to reduce the burden on this population of students.

Appendix 20 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, mean, standard deviation, number of items) and related statistics such as the standard error of measurement and mean p-value. All values in Appendix 20 were verified by two independent data analysts using two separate programming routines.

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; it is generally expressed symbolically as the following:

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

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 (i.e., 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 the test, σ_E^2 is the observed score variance, and p_i is the proportion of students who got item i correct, (i.e., 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 the test, σ_X^2 is the observed score variance, and σ_i^2 is the variance of item i .

The stratified coefficient alpha is a further extension of coefficient alpha to the situation where 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, and/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 interrater 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. Appendix 21 provides a research report showing the comparison of stratified coefficient alpha to other mixed-model reliability estimates for TAKS.

For SDAA II Writing tests, KR20 estimates of reliability are reported for the multiple choice portion of the tests and interrater correlations (between ratings of the first two raters) are reported for the essay prompts. Unlike TAKS Writing, where these two measures are combined to form the stratified coefficient alpha for the entire test, 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 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 interrater 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 is based on the premise 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 actually achieved on a test) to fluctuate above or below his or her true score (the true ability of the student). The SEM is calculated as

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

where r is the reliability estimate (e.g., 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. However, it is generally accepted (e.g., Peterson, Kolen, and 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 associated with them. For this reason, it is generally useful to report not only a test level SEM estimate, but individual score level estimate as well. Individual score level estimates of error are commonly referred to as conditional standard errors of measurement (CSEM).

Conditional Standard Error of Measurement

The CSEM provides an estimate of reliability, conditional on the proficiency estimate. In other words, it 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 there.

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, and exit level TAAS 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, and exit level TAAS report CSEM in terms of scale score units.

For SDAA II tests, SEM estimates were calculated using the following steps that average the weighted CSEM at each raw score level across the entire test (Lord, 1980). First, the error variance for a given raw score level was defined as the sum of the binomial probabilities of correctly responding to each of n items on a test given the specific proficiency level associated with that raw score level. 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 binomial probabilities of a student with a proficiency level θ_k correctly responding to each item, i , on the total test, n .

Under the assumptions of the Rasch measurement model, each raw score level, k , has associated with it only one corresponding proficiency estimate, θ_k . For every value of θ_k a specific probability of responding correctly, P_i , to each item 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 $\sigma_{e|\xi_k}^2$ at each raw score level, 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 calculated as the weighted root mean square CSEM across all N students

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

For TAKS, RPTE, and exit level TAAS tests, CSEM 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 can be calculated as the weighted CSEM across all N students

$$SE_{\theta} = \frac{1}{N} \sum_{k=1}^N SE_{\theta_k} .$$

Because TAKS and RPTE results are not reported in terms of Rasch proficiency estimates but in terms of scale scores, proficiency CSEM had to be converted to a scale score metric. Scale scores reported for TAKS and RPTE 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 of the proficiency estimate k , and T_i is the multiplicative scale score transformation constant (see Chapter 12, Tables 13 and 14).

Appendix 22 provides conditional standard errors of measurement for all TAKS, RPTE, and SDAA II tests. CSEMs are provided for the primary administration of each test only. All values in Appendix 22 were verified by two independent data analysts using two separate programming routines.

Use of the Standard Error of Measurement

The standard error of measurement is helpful for quantifying the margin of error 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 percent of the time (when errors are normally distributed). The standard error of measurement is used to quantify the reliability of a test into the metric on which scores will be reported. 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 percent 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 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 CSEM, 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. Placing a one-SEM 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 error than the average error of the test.

Appendix 20 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. Several elements of test construction and guidelines around setting cut scores can assist in minimizing these

errors. However, it is important to investigate the expected degree of misclassification prior to approval of the final cutscores. PEM conducted an analysis of the accuracy in student classifications into performance categories based on test results from the TAKS, RPTE, and SDAA II tests.

Common procedures for estimating classification accuracy are based on classical test theory conceptualizations of error distributions. However, the TAKS and RPTE scale scores are reported and equated using the Rasch model, which makes different model assumptions than classical test theory about the shape of the error distribution. (Note that although SDAA II results are reported in terms of raw score levels, raw score cutpoints are equated using the Rasch model). Other recent recommended procedures that use Item Response Theory, of which the Rasch model is an example, are either based on the assumption 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, and SDAA II. The procedures used for these tests are similar to those recommended by Rudner (2001, 2005), with modification for use in this special case.

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, the expected proportion of students with true scores in any particular level is

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

where a and b are Rasch scale points representing the score boundaries for the classification levels, d and c are the Rasch scale points representing score boundaries for persons in the classification levels, ϕ are the cumulative distribution functions of the achievement level boundaries, and φ is the normal density associated with the true score (Rudner, 2004).

This formula was modified for the current case. Modifications include

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 (Below Standard for TAKS, beginning for RPTE, and Achievement Level I for SDAA II) and the upper bound for highest performance category (commended for TAKS, 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, 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 because 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 tables for each TAKS, RPTE, and SDAA II grade and subject are provided in Appendix 23. In the tables below, 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.

Table 16. Classification Accuracy for TAKS Exit Level Social Studies 2005

Classification	<i>Below Standard</i>	<i>Met Standard</i>	<i>Commended</i>	Expected
<i>Below Standard</i>	4.9	2.0	0.0	6.9
<i>Met Standard</i>	0.9	64.8	5.4	71.0
<i>Commended</i>	0.0	3.2	18.9	22.1
Observed	5.7	70.0	24.3	100.0

Since TAKS uses Met the Standard for AYP and exit level decision purposes, it is useful to consider decision classification accuracy on a dichotomous classification of below Met Standard versus Met the Standard and above. To compute classification accuracy in this case, the cells associated with Met the Standard and Commended are collapsed and compared against Below Standard.

Table 17. Collapsed Classification Accuracy for TAKS Exit Level Social Studies 2005

Classification	<i>Below Standard</i>	<i>Met Standard/ Commended</i>	Expected
<i>Below Standard</i>	4.9	2.0	6.9
<i>Met Standard/ Commended</i>	0.9	92.3	93.1
Observed	5.7	94.3	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. No information regarding alternate or parallel forms reliability estimates is currently available, since no representative group of students takes more than one form of the test under similar conditions during any TAKS, SDAA II, RPTE, or exit level TAAS administration. 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.