Chapter 16: Reliability

Overview

Reliability is the most critical technical characteristic of any measurement because scores from a test with weak reliability cannot be interpreted in a valid way. Thus, the reliability of 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 Texas Assessment of Knowledge and Skills (TAKS), TAKS–Modified (TAKS–M), TAKS–Alternate (TAKS–Alt), Texas English Language Proficiency Assessment System (TELPAS) reading, Texas Assessment of Academic Skills (TAAS), and end-of-course (EOC) 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, TAKS–M, TELPAS reading (paper and online), EOC, and TAAS reliability data are based on internal consistency measures. These include 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 .87 to .90; for TAKS–M assessments ranging from .82 to .88; for paper versions of the TELPAS reading assessment ranging from .93 to .94 and online versions of the assessment ranging from .92 to .95. The reliabilities were .92, .91 and .91 for the Algebra I, biology, and geometry EOC assessments respectively.

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:

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

where the reliability, Pxx', 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 observed score 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 the "Determining An Appropriate Index of Reliability" report in the 2007 Texas Education Agency Technical Report Series which can be found at http://www.tea.state.tx.us/student.assessment/resources/techdigest/.

Classical Standard Error of Measurement (SEM)

The 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, TAKS–M, TELPAS reading, EOC, and 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. TAKS, TELPAS reading, EOC, and TAAS report CSEM in terms of scale score units, whereas some EOC (biology and geometry) report CSEM in raw score units.

For TAKS, TAKS–M, TELPAS reading, EOC Algebra I, and TAAS, 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:

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

Because TAKS, TAKS–M, TELPAS reading, and EOC Algebra I 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. However, EOC biology and geometry tests are reported in raw score units and, hence, the Rasch proficiency estimates will be reported in raw score units rather than scale score units. Scale scores reported for TAKS, TAKS–M, TELPAS reading, 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

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

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.

Appendix D provides conditional standard errors of measurement for all TAKS, TAKS–M, TELPAS reading, and EOC tests. CSEMs are provided for the primary administration of each test only.

Use of the SEM

The SEM 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 an 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 SEM 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, for instance, 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 conducted an analysis of the accuracy in student classifications into performance categories based on test results from the TAKS, TAKS–M, TELPAS reading, and EOC Algebra I tests. Classification accuracy was not calculated for EOC biology or geometry because standards have not been set.

Common procedures for estimating classification accuracy are based on classical test theory conceptualizations of error distributions. However, the TAKS, TAKS–M, TELPAS reading, and EOC Algebra I 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. 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, TAKS–M, TELPAS reading, 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(θ). 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:

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

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:

- 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.
- The lower bound for lowest performance category (Did Not Meet Standard for TAKS, TAKS–M, and EOC Algebra I, and Beginning for TELPAS reading) and the upper bound for highest performance category (Commended Performance for TAKS, TAKS–M, and EOC Algebra I, and Advanced High for TELPAS reading) were replaced with extreme, but unobserved, true proficiency/raw scores in order to capture the theoretical distribution in the tails.

- In computing the theoretical cumulative distribution, the lower bounds for the Met Standard performance level for TAKS, TAKS–M, and EOC Algebra I, and the Intermediate and Advanced performance levels for TELPAS reading, 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.
- 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, TAKS–M, TELPAS reading, and EOC Algebra I grade and subject are provided in Appendix E.

Figures 8 and 9 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 8 there was 86.6% agreement between expected and observed classifications for students who were in the two higher levels of performance.

Figure 8. Classification Accuracy for 2007 TAKS Exit Level Social Studies

Classification	*Did Not Meet Standard*	*Met Standard*	*Commended Performance*	***Expected***
Did Not Meet Standard	2.8	1.5	0.0	4.3
Met Standard	0.4	62.3	7.2	69.9
Commended Performance	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. Figure 9 shows that there was 95.4% agreement in classifications for students at the Met Standard/Commended Performance levels.

Figure 9. Collapsed Classification Accuracy for 2007 TAKS Exit Level Social Studies

Classification	*Did Not Meet Standard*	*Met Standard/ Commended Performance*	***Expected***
Did Not Meet Standard	2.8	1.5	4.3
Met Standard/ Commended Performance	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, TAKS–M, TELPAS reading, TAAS, or EOC 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 TAKS–Alt, evidence that the assessment allows for reliable observation and rating of student performance in the 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.

Information on the inter-rater reliability study that was conducted during the 2007 TAKS–Alt field test can be found in the 2006–2007 Technical Digest. An inter-rater reliability study was not conducted during the 2007–2008 school year due to a number of changes planned for the assessment in the following year. The next inter-rater reliability study will be conducted for the 2008–2009 school year. During the 2008–2009 school year, a sample of teachers will be asked to have a second observer provide a rating of student performance on a specified essence statement and its accompanying assessment task. The inter-rater reliability study will then be replicated for future TAKS–Alt administrations.

Gathering Reliability Evidence for TELPAS

TELPAS consists of two basic types of assessments, multiple-choice and holistically rated assessments. Reliability evidence for these types of assessments is gathered differently.

Multiple-choice TELPAS reading tests are used to assess English language learners (ELLs) in grades 2–12. Reliability evidence for these tests is demonstrated through similar reliability estimates as used for TAKS and other multiple-choice assessments. For the spring 2008 TELPAS reading tests, internal consistency estimates were rather high, with reliabilities for the online version ranging from .92 to .95 and reliabilities for the paper version ranging from .93 to .94. Appendix C presents reliability estimates by objectives for all students as well as for major demographic groups. In addition, conditional standard errors of measurement (see Appendix D), and classification accuracy (see Appendix E) inform about the consistency of scores across different item sets.

The TELPAS holistically rated components assess reading in K–1 and listening, speaking, and writing in K–12. Evidence that the assessment allows for reliable observation and rating of student performance is collected in two general ways. First, information about the consistency with which raters adhere to the strict administration protocol is provided. Second, evidence is shown to support interrater reliability, which informs about the consistency of scores provided by different raters.

TELPAS Audits

Evidence supporting the reliability of the TELPAS holistically rated components comes from evidence demonstrating that raters followed the defined protocol for rating these TELPAS components and from interrater reliability analyses conducted as part of the annual audits. Since the 2004–2005 school year, the Texas Education Agency (TEA) has conducted annual audits of the TELPAS assessment processes. As part of the audit, reports of rater adherence to the assessment protocol and interrater reliability evidence is collected. The rater adherence information is collected through surveys and state audits. Interrater reliability is collected by having a second rater (in addition to the rater of record) provide a rating for a sample of audited students. For the writing audits, the second rater provides a rating based on the same sample of student work used by the

first rater. For the listening and speaking audit, the second rater provides an independent rating based on his or her observation of the student in the classroom. Both approaches result in evidence of interrater reliability. Information about TELPAS audits is also included in Chapter 17: Validity because the same crucial pieces of evidence support both the validity and reliability of TELPAS holistically rated components.

Writing Audits

The TELPAS writing assessments require trained raters to use the writing proficiency level descriptors (PLDs) from the state's English language proficiency standards (ELPS) and student writing from classroom assignments to assign student English language proficiency levels. For four years, the Texas Education Agency (TEA) has required sampled school districts to submit to the state student writing collections they rated for the writing component of the assessment. Audit raters trained by the Pearson Performance Scoring Center re-rated the writing collections to monitor how well district raters were applying the PLDs as holistic scoring rubrics during the live assessment. The writing collections themselves were also examined for adherence to standardized protocols to ensure that they contained the necessary types and number of student writing assignments, as stipulated in the administration manuals.

The first audit, conducted in 2005, was relatively small and had the objective of helping the state establish appropriate audit procedures. The second audit, conducted in 2006, was a larger study in which information from a large, representative number of districts and students was collected. This audit provided results for regional education service centers (ESCs) and large districts referred to as training entities because of their role in directly providing TELPAS training to teacher raters. The 2006 audit results provided reliability and validity evidence supporting the accuracy of teacher ratings for the writing domain. The spring 2007 and spring 2008 TELPAS audits were smaller and served to provide ongoing evidence of validity and reliability at the level of the state rather than at the training entity level.

The results of the last three years of audits are presented in Table 12. The 2008 writing audit provided evidence of rater accuracy at a level similar to that reported in the 2007 audit. The overall perfect agreement rate of 79% was found to be satisfactory based on the Pearson Performance Scoring Center ISO Standards, and the adjacent agreement rate was 98%. In addition, the presence of a high correlation ($r = 0.89$) and a high weighted kappa ($w = 0.81$) value underscored the strong agreement between the state ratings and the teacher ratings.

Table 12. Writing Audit Results from 2006 to 2008

Year	Sample Size	Perfect Agreement Rate	Correlation Between District and State Raters	Kappa
2006	13357	77%	0.87	0.79
2007	542	76%	0.87	0.78
2008	1932	79%	0.89	0.81

Table 12 illustrates consistently high rating accuracy across multiple years and provides evidence of stability in interrater agreement over time. Since TELPAS scores across years are used in reporting student progress in language acquisition, the state's finding of rater accuracy over time supports inferences about annual student progress from TELPAS scores.

The state will continue to audit the writing process periodically to provide ongoing monitoring of rating effectiveness and to give district personnel feedback to support the administration of this assessment.

Listening and Speaking Audits

The process for collecting reliability evidence supporting the listening and speaking components parallels the process used for collecting evidence for the writing component. The first small-scale pilot audit of TELPAS listening and speaking components was conducted in spring 2008. In this small-scale study, audit raters with previous TELPAS rater training were provided a face-to-face refresher training in May 2008 and then sent to school districts to re-rate ELL students who had been assessed in March–April 2008 in their natural classroom settings.

The pilot encompassed two TELPAS grade clusters, grade 2 and grades 6–8, and included 43 internal (from TEA and Pearson) and external (from districts and education service centers) audit raters. Each audit rater was given a list of several students to rate blindly over the course of three days. The objectives of this small pilot were to examine the viability of this method of auditing rating efficacy and the feasibility of conducting this type of audit on a larger scale.

The audit raters received face-to-face training in two main areas: rating skills and audit logistics. The audit raters received refresher training in the proficiency level rubrics in order to observe students, assign ratings, document reasons for their ratings, and determine whether the students fell into the early, middle, or late stage within the proficiency levels assigned.

The pilot enabled audit raters to re-rate approximately 150 students in each of the two grade clusters. Although the pilot was small in scope and used only volunteer audit raters, students were sampled in a way that maximized representation of different regions of the state to the extent possible.

Questionnaire information from the audit raters was collected along with their ratings of students. The questionnaire results indicated that the majority of audit raters (60%) were able to rate 1–2 students per day if the students belonged to different classrooms at the same campus and 3–4 students if the students belonged to the same classroom at the same campus. The audit raters strongly agreed (62%) or agreed (38%) that the training and the materials provided during training prepared them well to holistically rate listening and speaking proficiency. The audit raters also strongly agreed (88%) or agreed (12%) that procedural documents provided at audit rater training were clear and complete.

An important finding of the pilot was that the time of year of the audit posed logistical challenges for the participants. Audit raters indicated that they would not volunteer again in May because end-of-year duties are too time-consuming to allow them to participate in the audit. They also indicated that teachers of the students would need to be instructed ahead of time to plan academic lessons that engaged students in more listening and speaking activities. Because of the limited time spent with the students, audit raters sometimes indicated that they would have preferred to see students in a greater variety of academic listening and speaking interactions to be confident about determining their proficiency levels.

This small-scale study provided the opportunity for TEA to gather feedback from audit raters on logistics and data collection procedures. Input from this pilot will help determine the feasibility and standardized processes needed for future larger-scale audits.

Internal Consistency of TELPAS Composite Ratings

After administering all components of TELPAS, the listening, speaking, reading, and writing domain ratings are combined into an overall TELPAS composite rating used for state and federal accountability purposes. The composite rating weighting formula is shown in Table 13 below.

Table 13. Weights of the Language Domains in TELPAS Composite Ratings

Listening	Speaking	Reading	Writing
5%	5%	75%	15%

The composite English language proficiency rating is used to inform educators, parents, and policymakers of the progress ELLs make in developing academic English language proficiency. The weights were determined based on input from educators and technical experts.

The reading and writing domains receive the most weight in the composite rating because of the belief of Texas educators that ELLs who can read and write proficiently in English but have weaker oral communication skills are more likely to succeed academically than students who have strong oral skills but weaker reading and writing skills. The reading and writing domains are also the domains for which the state has the most reliability evidence. The ability to make consistent inferences about students' English language proficiency is stronger at the composite level, which combines performance across domains, than at the individual domain level.