Florida
Standards Assessments

# Florida Standards Assessments

# 2014–2015

# Volume 4
# Evidence of Reliability and Validity

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

## ACKNOWLEDGEMENTS

This technical report was produced on behalf of the Florida Department of Education. Requests for additional information concerning this technical report or the associated appendices should be directed to Dr. Salih Binici at the Florida Department of Education (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from American Institutes for Research (AIR): Dr. Harold Doran, Dr. Elizabeth Ayers-Wright, Dr. Dipendra Subedi, Dr. MinJeong Shin, Dr. AhYoung Shin, Danielle Peterson, and Patrick Kozak. The major contributors from the Florida Department of Education are as follows: Dr. Salih Binici, Dr. Molly Hand, Dr. Qian Liu, Vince Verges, Victoria Ash, Susie Lee, Mengyao Cui, Steve Ash, Renn Edenfield, and Chris Harvey.

## TABLE OF CONTENTS

## LIST OF APPENDICES



## LIST OF TABLES

## LIST OF FIGURES

# 1. Introduction and Overview of Reliability and Validity Evidence

The State of Florida implemented a new assessment program for operational use during the 2014–2015 school year. This new program, named the Florida Standards Assessments (FSA), replaced the Florida Comprehensive Assessment Tests (FCAT) 2.0 in English Language Arts and Mathematics. Students in grades 3 and 4 were administered fixed, operational ELA Reading and Mathematics forms on paper. Students in grades 5 through 10 were administered fixed, operational Reading forms online, and students in grades 5 through 8 were administered fixed, operational Mathematics forms online. End-of-Course (EOC) assessments were administered to students taking Algebra 1, Algebra 2, and Geometry. In addition, students in grades 4 through 10 responded to a text-based Writing prompt, with grades 4 through 7 administered on paper and grades 8 through 10 administered online. Writing and Reading scores were combined to form an overall English Language Arts (ELA) score.

In the grades with online testing, paper forms, in lieu of online forms, were administered to students whose Individual Educational Plans (IEP) or Section 504 plans indicated such a need. Grades 3 and 4 Mathematics and Reading were universally administered on paper, so there were no accommodated forms. Table 1 displays the complete list of test forms for the operational administration.

*Table 1: Test Administration*

| Subject | Administration | Grade/Course |
|---|---|---|
| ELA Reading | Paper | 3–4 |
| ELA Reading | Online | 5–10 |
| | Paper (Accommodated) | |
| ELA Writing | Paper | 4–7 |
| ELA Writing | Online | 8–10 |
| | Paper (Accommodated) | |
| Mathematics | Paper | 3–4 |
| Mathematics | Online | 5–8 |
| | Paper (Accommodated) | |
| EOC | Online | Algebra 1, Algebra 2, Geometry |
| | Paper (Accommodated) | |

With the implementation of these new tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the FSA scores. This volume provides empirical evidence about the reliability and validity of the 2014–2015 FSA, given its intended uses.

The purpose of this volume is to provide empirical evidence to support the following:

- Reliability: Multiple reliability estimates for each test are reported in this volume, including stratified-coefficient *alpha*, Feldt-Raju, and the marginal reliability. The reliability estimates are presented by grade and subject as well as by demographic subgroups. This section also includes conditional standard errors of measurement by grade and subject, as well as standard deviation of theta and mean standard errors of measurement of theta.
- Content validity: Evidence is provided to show that test forms were constructed to measure the Florida Standards with a sufficient number of items targeting each area of the blueprint.
- Internal structure validity: Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among reporting categories per grade. Confirmatory factor analysis has also been performed using the second-order factor model. Additionally, local item independence, an assumption of unidimensional IRT, was tested using the $Q_3$ statistic.
- Comparability of paper-pencil to online tests: By examining the blueprint match between forms and test characteristic curves (TCCs) for both forms, we evaluate comparability of test scores across forms.
- Test fairness: Fairness is statistically analyzed using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

## 1.1 RELIABILITY

Reliability refers to consistency in test scores. Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}.$$

There are various approaches for estimating the reliability of scores. The conventional approaches used are characterized as follows:

- The *test-retest* method measures stability over time. With this method, the same test is administered twice to the same group at two different points in time. If test scores from the two administrations are highly correlated, then the test scores are deemed to have a high level of stability. For example, if the result is highly stable, those who scored high on the first administration tend to obtain a high score on the second administration. The critical factor, however, is the time interval. The time interval should not be too long, which could allow for changes in the examinees' true scores. Likewise, it should not be too short, in which case memory and practice may confound the results. The test-retest method is most effective for measuring constructs that are stable over time, such as intelligence or personality traits.

- The *parallel-forms* method is used for measuring equivalence. With this design, two parallel forms of the test are administered to the same group. This method requires two similar forms of a test. However, it is very difficult to create two strictly parallel forms. When this method is applied, the effects of memory or practice can be eliminated or reduced, since the tests are not purely identical as with the test-retest method. The reliability coefficient from this method indicates the degree to which the two tests are measuring the same construct. While there are a wide variety of possible items to administer to measure any particular construct, it is only feasible to administer a sample of items on any given test. If there is a high correlation between the scores of the two tests, then inferences regarding high reliability of scores can be substantiated. This method is commonly used to estimate the reliability of achievement or aptitude tests.
- The *split-half* method utilizes one test divided into two halves within a single test administration. It is crucial to make the two half-tests as parallel as possible, as the correlation between the two half-tests is used to estimate reliability of the whole test. In general, this method produces a coefficient that underestimates the reliability for the full test. To correct the estimate, the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) can be applied. While this method is convenient, varying splits of the items may yield different reliability estimates.
- The *internal consistency* method can be employed when it is not possible to conduct repeated testing administrations. Whereas other methods often compute the correlation between two separate tests, this method considers each item within a test to be a one-item test. There are several other statistical methods based on this idea: coefficient *alpha* (Cronbach, 1951), Kuder-Richardson Formula 20 (Kuder & Richardson, 1937), Kuder-Richardson Formula 21 (Kuder & Richardson, 1937), stratified coefficient *alpha* (Qualls, 1995), and Feldt-Raju coefficient (Feldt & Qualls, 1996; Feldt & Brennan, 1989).
- *Inter-rater reliability* is the extent to which two or more individuals (coders or raters) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system.

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEM)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory assumes that an observed score (X) of each individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of $X$ can be shown to be the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, we can arrive at:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_x^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

As the fraction of error variance to observed score variance tends to zero, the reliability then tends to 1. The Classical Test Theory (CTT) SEM, which assumes a homoscedastic error, is derived from the classical notion expressed above as $\sigma_X\sqrt{1-\rho_{XX'}}$, where $\sigma_X$ is the standard

deviation of the scaled score and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived.

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \rho_{XX'},$$

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}),$$

$$\sigma_E = \sigma_X\sqrt{(1 - \rho_{XX'})}.$$

In general, the standard error of measurement is relatively constant across samples as the group dependent term, $\sigma_X$, can be shown to cancel out:

$$\sigma_E = \sigma_X\sqrt{(1 - \rho_{XX'})} = \sigma_X\sqrt{(1 - (1 - \frac{\sigma_E^2}{\sigma_X^2}))} = \sigma_X\sqrt{\frac{\sigma_E^2}{\sigma_X^2}} = \sigma_X \cdot \frac{\sigma_E}{\sigma_X} = \sigma_E.$$

This shows that the standard error of measurement is generally stable and invariant to different groups of test-takers.

In IRT, the standard errors of measurement vary over the ability continuum. These heterogeneous errors are a function of a test information function that provides different information about examinees depending on their estimated abilities. Often, the test information function (TIF) is maximized over an important performance cut, such as the proficient cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the "lack" of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut, and have less information at the tails of the score distribution. See Section 3.3 for the derivation of heterogeneous errors in IRT.

## 1.2 VALIDITY

*Validity* refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment." Both of these definitions emphasize evidence and theory to support inferences and interpretations of test scores. *The Standards* (AERA, APA, & NCME, 2014) suggests five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct. In order for test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (see Volume 2 for details). Test scores can be used to support an intended validity claim when they contain minimal construct irrelevant variance. For example, a mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and vocabulary has a high level of construct irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores. Statistical analyses, such as factor analysis or multi-dimensional scaling of relevance, are also used to evaluate content relevance. Results from factor analysis for the FSA are presented in section 5.2. Evidence based on test content is a crucial component of validity, because construct underrepresentation or irrelevancy could result in unfair advantages or disadvantages to one or more group of examinees.

Technology-enhanced items should be examined to ensure that no construct irrelevant variance is introduced. If some aspect of the technology impedes, or advantages, a student in his or her responses to items, this could affect item responses and inferences regarding abilities on the measured construct. Florida makes use of the technology enhanced items developed by AIR, and the items are delivered by the same engine as is used for delivery of the Smarter Balanced assessment. Hence, the FSA makes use of items that have the same technology-enhanced functionality as those found on these other assessments. A cognitive lab study was completed for the Smarter Balanced assessment, providing evidence in support of the item types used for the consortium and also in Florida. The complete study is provided as a compendium to the FSA technical reports in Volume 7, showing support for the item types used on the FSA tests.

The second source of validity evidence is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014).This evidence is collected by surveying examinees about their performance strategies or responses to particular items. Because items are developed to measure particular constructs and intellectual processes, evidence that examinees have engaged in relevant performance strategies to correctly answer the items supports the validity of the test scores.

The third source of evidence for validity is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Differential item functioning, which determines whether particular items may function differently for subgroups of examinees, is one method for analyzing the internal structure of tests (see Volume 1, Section 5.2). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Sections 3 and 5 for details).

A fourth source of evidence for validity is the relationship of test scores to external variables. *The Standards* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant

evidence, a multitrait-multimethod matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restriction may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

A study linking state tests to the NAEP test (Phillips, 2016) found that the Florida grades 4 and 8 level 4 performance standards, in both Mathematics and ELA, mapped to the NAEP proficiency levels. This is a rigorous standard that only Florida met as reported by Phillips (2016).

Fifth, the intended and unintended consequences of test use should be included in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is in fact due to an unintended, confounding aspect of the test, this would interfere with the test's validity. As described in Volume 1 and additionally in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This then allows for one to evaluate if sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores, and subsequently, evidence that the scores can be used to support these inferences.

# 2. PURPOSE OF FLORIDA'S STATE ASSESSMENT

The Florida Standards Assessments (FSA) are standards-based, summative tests that measure students' achievement of Florida's education standards. Assessment supports instruction and student learning, and the results help Florida's educational leadership and stakeholders determine whether the goals of the education system are being met. Assessments help Florida determine whether it has equipped its students with the knowledge and skills they need to be ready for careers and college-level coursework. The tests are constructed to meet rigorous technical criteria (Standards for Educational and Psychological Testing [AERA, APA, & NCME, 2014]) and to ensure that all students have access to the test content via principles of universal design and appropriate accommodations.

The FSA yields test scores that are useful for understanding to what degree individual students have mastered the Florida Standards and, eventually, whether students are improving in their performance over time. Additionally, scores can be aggregated to evaluate the performance of subgroups, and both individual and aggregated scores will be compared over time in program evaluation methods.

The FSA results serve as the primary indicator for the state's accountability system, and the policy and legislative purpose of the FSA is described more thoroughly in Volume 1. The test is a standards-based assessment designed to measure student achievement toward the state content standards. FSA scores are indications of what students know and are able to do relative to the expectations by grade and subject area. While there are student-level stakes associated with the assessment, particularly for Grade 3 ELA (scores inform district promotion decisions) and Grade 10 ELA and Algebra 1 (assessment graduation requirements), the assessment is never the sole determinant in making these decisions.

Test items were selected prior to the test administration to ensure that the test construction aligned to the approved blueprint. The content and psychometric verification log was kept to track the compliance of the test structure to the FSA requirements.

In the FSA administered in 2015, student-level scores included T scores, percentile ranks, and raw scores at the reporting category level. After performance cuts were approved by the State Board of Education on January 6, 2016, scaled scores were retrofitted for spring 2015 tests and reported back to districts. These scale scores and achievement levels will be reported in spring 2016 and beyond. Volume 1 Section 8.1 of the FSA Annual Technical Report describes how each of these scores is computed.

The raw scores for reporting categories were provided for each student to indicate student strengths and weaknesses in different content areas of the test relative to the other areas and to the district and state. These scores serve as useful feedback for teachers to tailor their instruction, provided that they are viewed with the usual caution that accompanies use of reporting category scores. Thus, we must examine the reliability coefficients for these test scores and the validity of the test scores to support practical use across the state.

# 3. RELIABILITY

## 3.1 INTERNAL CONSISTENCY

As the FSA was administered in a single administration, it is necessary to examine the internal consistency of the test to support the reliability of the test scores. For the FSA ELA, Mathematics, and EOC assessments, the reliability coefficients were computed using Cronbach *alpha*, stratified *alpha*, and Feldt-Raju coefficient. In addition to Cronbach *alpha*, stratified *alpha* and Feldt-Raju coefficients were computed treating multiple-choice and non-multiple-choice items as two separate strata.

The FSA ELA, Mathematics, and EOC Assessments included mixed item types: multiple choice, short response, and extended response. Although there are various techniques for estimating the reliability of test scores with multiple item types or parts (Feldt & Brennan, 1989; Lee & Frisbie, 1999; Qualls, 1995), studies (Qualls, 1995; Yoon & Young, 2000) indicate that the use of Cronbach *alpha* underestimates the reliability of test scores for a test with mixed item types.

The Cronbach *alpha* is defined as

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

where $\sigma_i^2$ is the variance of scores on each item, $\sigma_x^2$ is the variance of the total test scores, and $n$ is the number of items.

The stratified Cronbach *alpha* coefficient is computed as

$$\text{stratified } \alpha\ \rho_{XX'} = 1 - \frac{\sum_{i=1}^{k}\sigma_i^2(1-\alpha_i)}{\sigma_x^2},$$

where $\alpha_i$ is the reliability of the $i$th strata, $\sigma_i^2$ is the variance between items in the $i$th strata, and $\sigma_x^2$ is the variance of the total test scores. The stratified Cronbach *alpha* coefficient takes into account the weights proportional to the number of items and mean scores for each stratum. Qualls (1995) incorporated Raju's (1977) and Feldt's (Feldt & Brennan, 1989) techniques for calculating reliability, which is called Feldt-Raju coefficient.

The Feldt-Raju coefficient is defined as

$$\text{Feldt-Raju } \rho_{XX'} = \frac{\sigma_x^2 - \sum_{i=1}^{k}\sigma_i^2}{(1-\sum_{i=1}^{k}\hat{\lambda}_i^2)\sigma_x^2},$$

where $\sigma_x^2$ is the total score variance, (i.e., the variance of the whole test); $\sigma_x^2$ indicates the score *variance* for a part-test (or item type) $i$; and $\hat{\lambda}_i$ is the sum of the variance of item type $i$ and the covariance between item type $i$ and other item types. This is defined as

$$\hat{\lambda}_i = \frac{(\sigma_{i1} + \sigma_{i2} + \sigma_i^2 + \sigma_{i2} + \cdots + \sigma_{ik})}{\sigma_x^2}.$$

Table 2 through Table 5 display item types and their descriptions, as well as the number of items belonging to each item type. These tables were used to classify strata of item types. Because there were not considerable amounts of each of the individual item types, we organized the items into two categories for our analyses: multiple-choice and non-multiple-choice.

*Table 2: Reading Item Types and Descriptions*

| Response Type | Description |
|---|---|
| Multiple-Choice (MC) | Student selects one correct answer from a number of options. |
| Multi-Select (MS) | Student selects all correct answers from a number of options. |
| Editing Task (ET) | Student identifies an incorrect word or phrase and replaces it with the correct word or phrase. |
| Editing Task Choice (ETC) | Student identifies an incorrect word or phrase and chooses the replacement from a number of options. |
| Hot Text (HT) | Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. |
| GRID (GI) | Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer. |
| Evidence Based Selected Response (EBSR) | Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A. |
| Natural Language (NL) | Student uses the keyboard to enter a response into a text field. |

*Table 3: Mathematics Item Types and Descriptions*

| Response Type | Description |
|---|---|
| Multiple-Choice (MC) | Student selects one correct answer from a number of options. |
| Multi-Select (MS) | Student selects all correct answers from a number of options. |
| Short Answer (SA ) | Student writes a numeric response to answer the question. |
| GRID (GI) | Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer. |
| Hot Text (HT) | Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. |
| Equation (EQ) | Student uses a toolbar with a variety of mathematical symbols to create a response. |
| Word Builder (WB) | Student enters a numeric value and bubbles in the corresponding number or symbol. |
| Natural Language (NL) | Student uses the keyboard to enter a response into a text field. |
| Matching (MI) | Student checks a box to indicate if information from a column header matches information from a row. |
| Table (TI) | Student types numeric values into a given table. |

*Table 4: Reading Operational Item Types by Grade*

| Item type * | Grade | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| MC | 37 | 38 | 35 | 28 | 32 | 30 | 35 | 40 |
| MS | 1 | | 4 | 3 | 2 | 4 | 2 | 1 |
| ET | | | | | | 3 | 3 | |
| ETC | 8 | 8 | 8 | 8 | 8 | 6 | 5 | 8 |
| HT | 2 | 2 | 2 | 7 | 5 | 4 | 6 | 4 |
| GI | 1 | | | | | 1 | | |
| EBSR | 1 | 2 | | 5 | 5 | 4 | 2 | 1 |
| NL | | | 1 | 1 | | | 1 | |

* Descriptions for each item type are presented in Table 2

*Table 5: Mathematics Operational Item Types by Grade*

| Item type * | Grade | | | | | | Algebra 1** | Algebra 2** | Geometry** |
|---|---|---|---|---|---|---|---|---|---|
| | **3** | **4** | **5** | **6** | **7** | **8** | | | |
| MC4 | 50 | 51 | 45 | 39 | 36 | 35 | 30; 29; 28; 27 | 21; 20 | 29; 27 |
| MS5 | 1 | | | | | 2 | 1; 2; 1; 1 | 2; 1 | |
| MS6 | | | | | | | 0; 0; 1; 2 | 2; 2 | |
| GI | | | 3 | 8 | 7 | 10 | 7; 7; 6; 7 | 5; 6 | 8; 9 |
| SA | 3 | 3 | | | | | | | |
| EQ | | | 6 | 9 | 13 | 9 | 20; 20; 22; 21 | 28; 29 | 21; 22 |

* Descriptions for each item type are presented in Table 3
** Algebra 1 has four core forms, and Algebra 2 and Geometry have two core forms

Table 6 through Table 8 present the Cronbach *alpha*, stratified *alpha*, and Feldt-Raju coefficients for ELA, Mathematics, and EOC by grade/course and test form. The Cronbach *alpha* ranged from 0.86 to 0.92 for ELA, 0.82 to 0.93 for Mathematics, and 0.84 to 0.94 for EOC. The stratified *alpha* coefficients ranged from 0.85 to 0.92 for ELA, 0.82 to 0.94 for Mathematics, and 0.84 to 0.93 for EOC. The Feldt-Raju coefficients were between 0.84 and 0.91 for ELA, 0.87 and 0.92 for Mathematics, and 0.87 and 0.90 for EOC. The reliability coefficients by each demographic subgroup are presented in Appendix A. Reliability coefficients for each reporting category are also presented in Appendix A.

*Table 6: Reliability Coefficients (ELA)*

| Grade | Form | Cronbach *Alpha* | Stratified *Alpha* | Feldt-Raju |
|---|---|---|---|---|
| 3 | Paper | 0.89 | 0.89 | 0.84 |
| 4 | Paper | 0.89 | 0.89 | 0.86 |
| 5 | Online | 0.89 | 0.90 | 0.88 |
| | Accommodated | 0.86 | 0.85 | 0.84 |
| 6 | Online | 0.92 | 0.92 | 0.91 |
| | Accommodated | 0.90 | 0.89 | 0.87 |
| 7 | Online | 0.90 | 0.91 | 0.89 |
| | Accommodated | 0.90 | 0.88 | 0.85 |
| 8 | Online | 0.92 | 0.92 | 0.90 |
| | Accommodated | 0.91 | 0.89 | 0.87 |
| 9 | Online | 0.91 | 0.92 | 0.89 |
| | Accommodated | 0.90 | 0.88 | 0.87 |
| 10 | Online | 0.90 | 0.90 | 0.88 |
| | Accommodated | 0.89 | 0.87 | 0.85 |

*Table 7: Reliability Coefficients (Mathematics)*

| Grade | Form | Cronbach *Alpha* | Stratified *Alpha* | Feldt-Raju |
|---|---|---|---|---|
| 3 | Paper | 0.93 | 0.93 | 0.94 |
| 4 | Paper | 0.93 | 0.94 | 0.93 |
| 5 | Online | 0.93 | 0.93 | 0.92 |
| | Accommodated | 0.90 | 0.90* | |
| 6 | Online | 0.92 | 0.92 | 0.93 |
| | Accommodated | 0.88 | 0.88* | |
| 7 | Online | 0.93 | 0.93 | 0.91 |
| | Accommodated | 0.90 | 0.90* | |
| 8 | Online | 0.87 | 0.87 | 0.87 |
| | Accommodated | 0.82 | 0.82 | 0.87 |

***** These values are based on the total test. Grades 5, 6, and 7 Mathematics accommodated forms did not have enough non-MC items to compute stratified *alpha*.

*Table 8: Reliability Coefficients (EOC)*

| Course | Form | Cronbach *Alpha* | Stratified *Alpha* | Feldt-Raju |
|---|---|---|---|---|
| Algebra 1 | Online – Core 1 | 0.91 | 0.91 | 0.90 |
| | Online – Core 2 | 0.91 | 0.90 | 0.89 |
| | Online – Core 3 | 0.91 | 0.91 | 0.89 |
| | Online – Core 4 | 0.91 | 0.91 | 0.89 |
| | Accommodated | 0.84 | 0.84 | 0.87 |
| Algebra 2 | Online – Core 1 | 0.92 | 0.92 | 0.89 |
| | Online – Core 2 | 0.92 | 0.92 | 0.89 |
| | Accommodated | 0.86 | 0.86 | 0.87 |
| Geometry | Online – Core 1 | 0.94 | 0.93 | 0.90 |
| | Online – Core 2 | 0.94 | 0.93 | 0.91 |
| | Accommodated | 0.89 | 0.89 | 0.92 |

## 3.2 MARGINAL RELIABILITY

Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale, for all students. The marginal reliability coefficients are nearly identical or close to coefficient *alpha*. For our analysis, the marginal reliability coefficients were produced in IRTPRO.

Within the IRT framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. Thus, test information is a value that is the inverse of the measurement error of the test. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale that is at the middle of the test distribution and greater on scaled values further away from the middle.

The marginal reliability is defined as:

$$\bar{\rho} = 1 - \frac{\int \sigma_e^2(\hat{\theta}) f(\hat{\theta}) d\hat{\theta}}{\sigma_x^2}$$

where $\sigma_e^2(\hat{\theta})$ is the conditional standard error (squared) and $f(\hat{\theta})$ is the assumed population density. Table 9 presents the marginal reliability coefficients for all students. The reliability coefficients ranged from 0.88 to 0.93. Grade 8 Mathematics and Algebra 2 had reliability coefficients of 0.88 and 0.86, respectively. The marginal reliability coefficients for all other subjects and grades were higher than 0.9.

*Table 9: Marginal Reliability Coefficients*

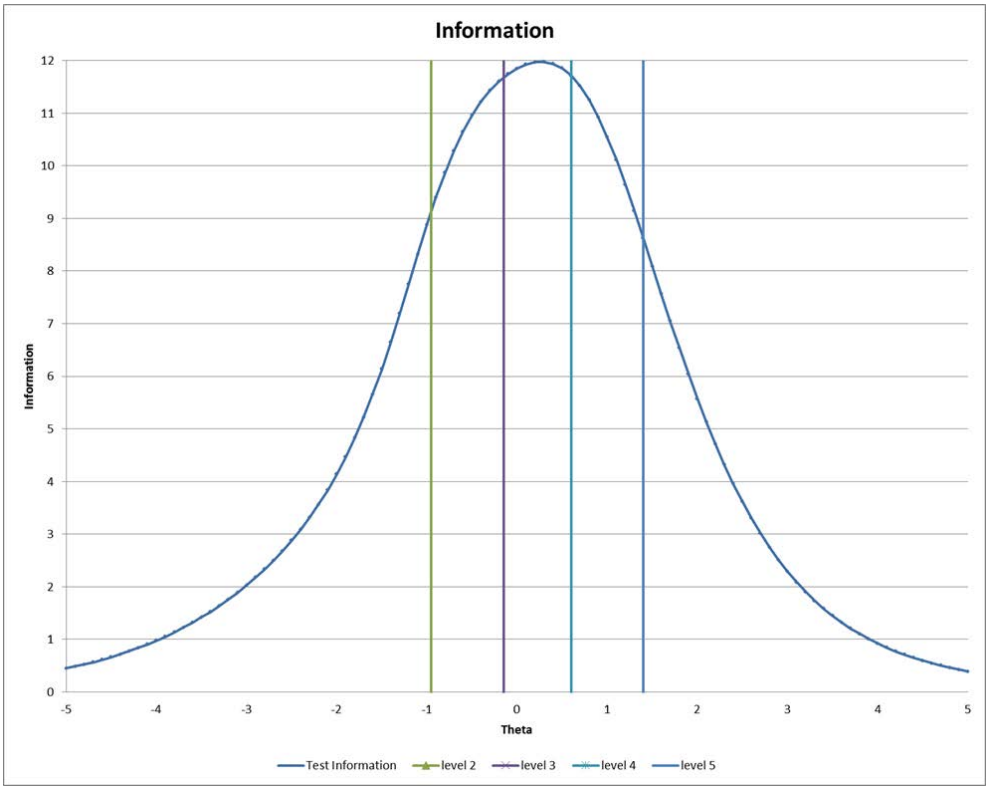
| Subject | Grade | Marginal Reliability for Response Pattern Scores | Subject | Grade/Course | Marginal Reliability for Response Pattern Scores |
|---|---|---|---|---|---|
| ELA | 3 | 0.90 | Mathematics | 3 | 0.92 |
| | 4 | 0.91 | | 4 | 0.93 |
| | 5 | 0.91 | | 5 | 0.93 |
| | 6 | 0.92 | | 6 | 0.92 |
| | 7 | 0.92 | | 7 | 0.92 |
| | 8 | 0.92 | | 8 | 0.88 |
| | 9 | 0.93 | EOC | Algebra 1 | 0.93 |
| | 10 | 0.92 | | Algebra 2 | 0.86 |
| | | | | Geometry | 0.93 |

## 3.3 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

Within the IRT framework, measurement error varies across the range of ability as a result of the test information function (TIF). The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, then less information is being provided by the assessment at the specific ability level.

Figure 1 displays a sample TIF from the 2015 FSA. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. Where the curve is lower at the tails indicates that the test provides less information about examinees at the tails relative to the center. The vertical lines are samples of the performance cuts.

*Figure 1: Sample Test Information Function*




Computing these TIFs is useful to evaluate where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the FSA is calculated as:

$$TIF(\theta_s) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left( \frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} - \left( \frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left( \frac{Q_i}{P_i} \left[ \frac{P_i - c_i}{1 - c_i} \right]^2 \right),$$

where $N_{GPCM}$ is the number of items that are scored using GPCM items, $N_{3PL}$ is the number of items scored using 3PL or 2PL model, i indicates item i $(i \in \{1,2,\ldots,N\})$, $m_i$ is the maximum possible score of the item, s indicates student s, and $\theta_s$ is the ability of student s.

The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se(\theta_s) = \frac{1}{\sqrt{TIF(\theta_s)}}.$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are

presented in Figure 2, Figure 3, and Figure 4, respectively, instead of the TIFs for ELA, Mathematics, and EOC. These plots are based on the T scores reported in 2015. Scaled scores will not be reported until spring 2016. However, the T score is a monotonic function of the original ability estimate, and the scale scores will also be monotonic transformations of the original ability estimate. Hence, the plots below are the same plots that would be observed if they were based on any of those score types.

*Figure 2: Conditional Standard Errors of Measurement (ELA)*

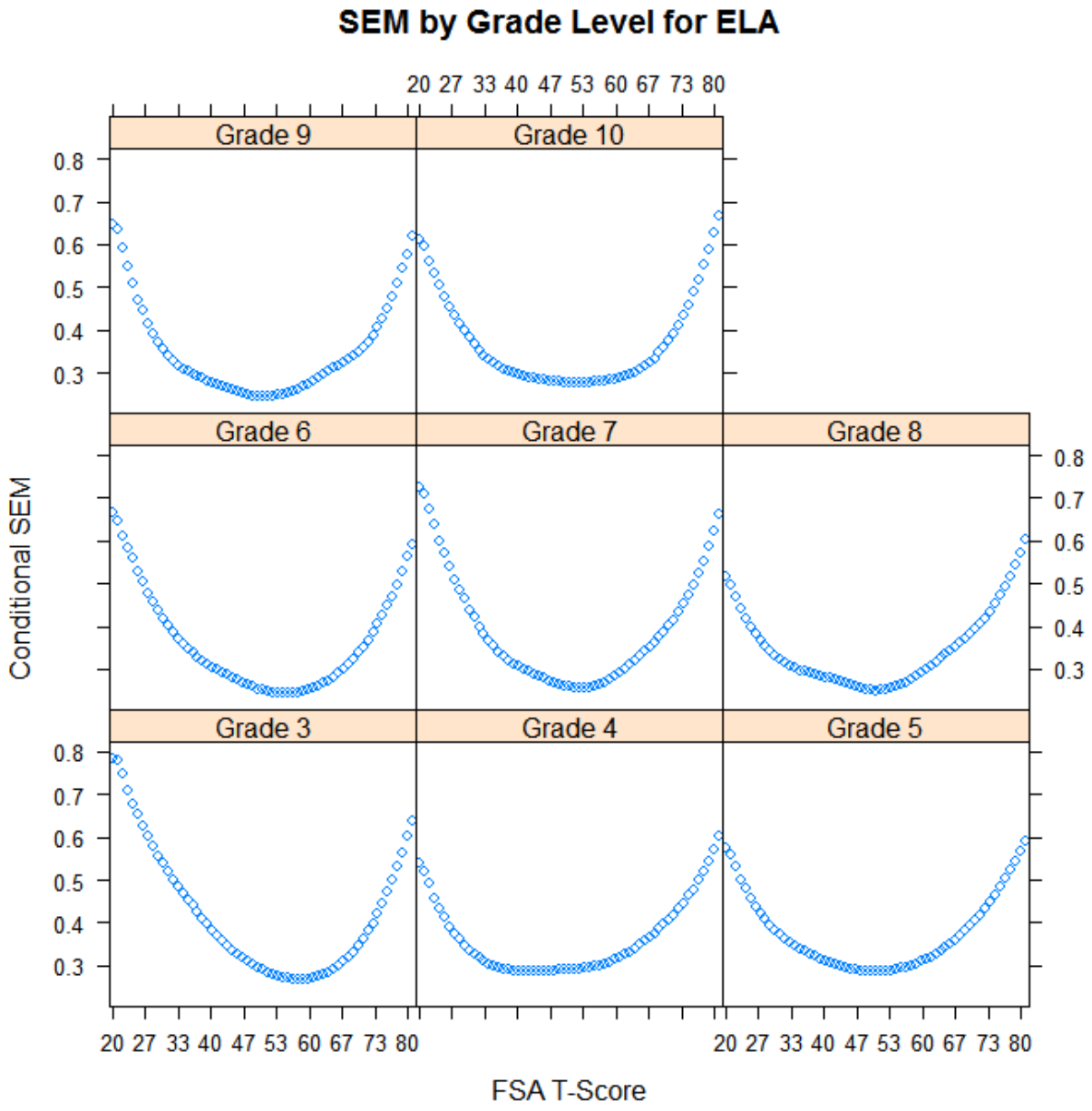


*Figure 3: Conditional Standard Errors of Measurement (Mathematics)*

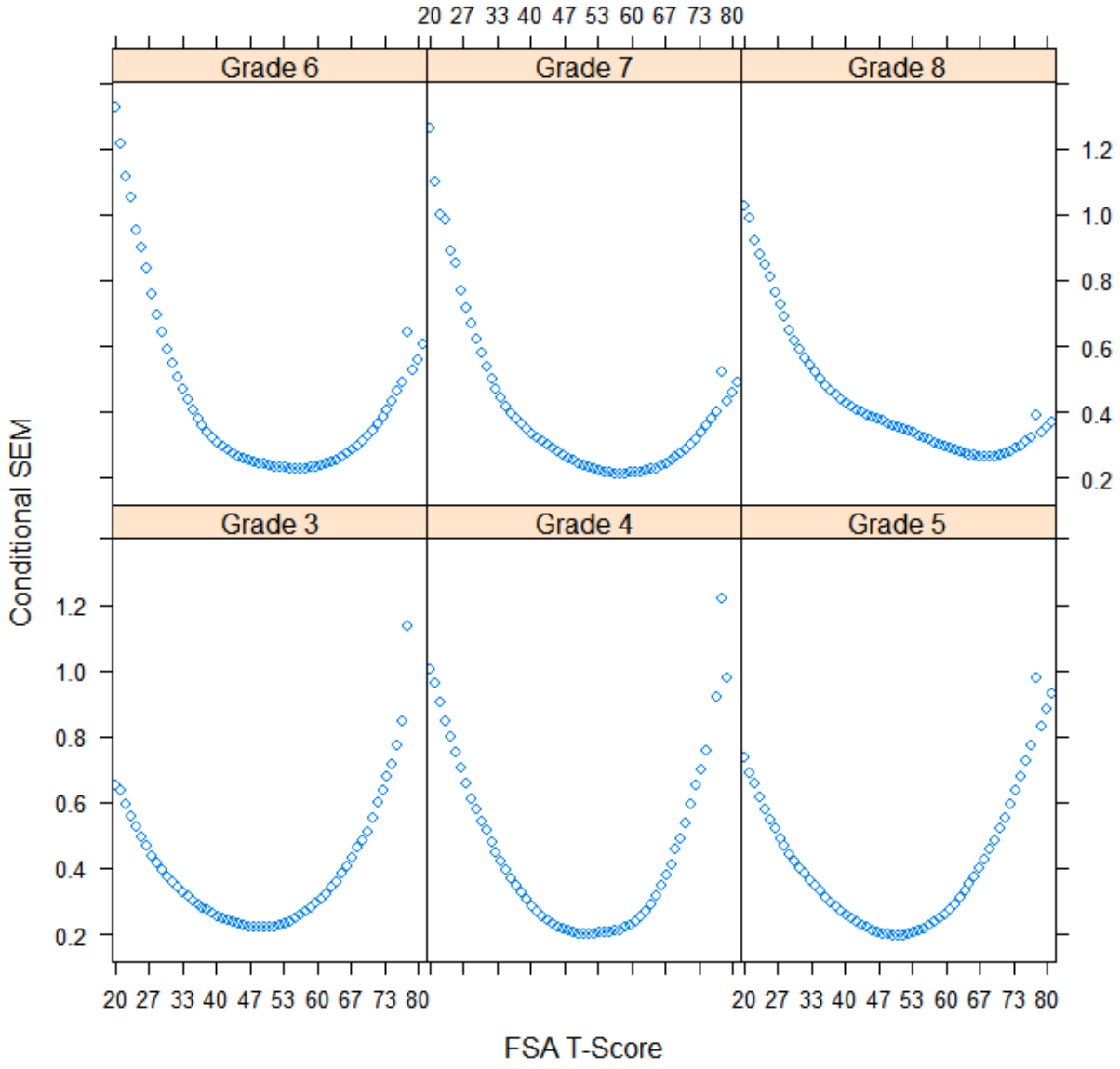
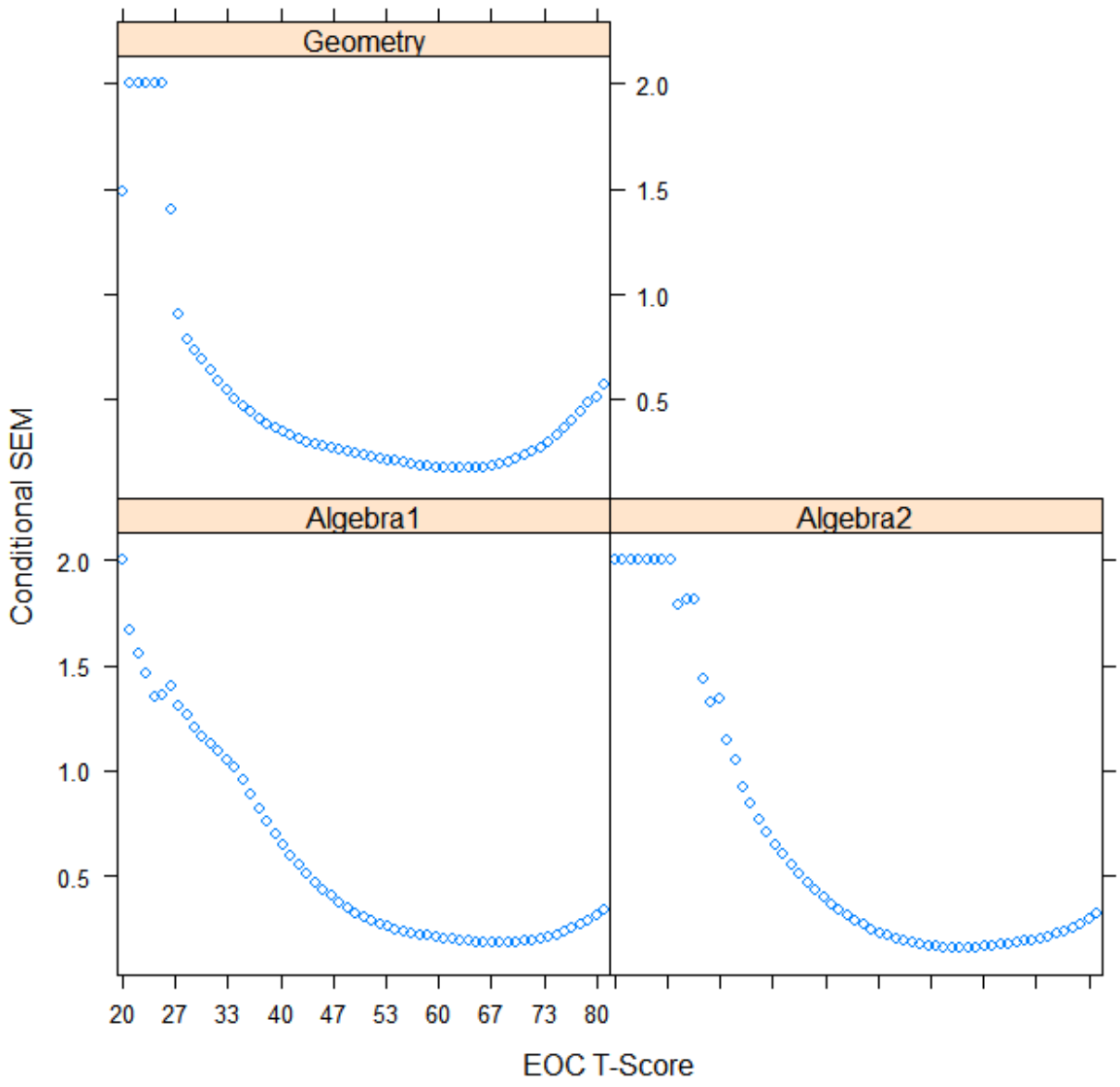


*Figure 4: Conditional Standard Errors of Measurement (EOC)*




For most tests, the standard error curves follow the typical expected trends with more test information regarding scores observed near the middle of the score scale. However, there are two general exceptions. In grade 8 Mathematics and for both Algebra EOC tests, the test is maximized at a higher point along the ability scale. This suggests the items comprising these tests are somewhat challenging relative to the tested population for this initial test administration. Because the testing of Algebra 1 is relatively new and the statewide testing of Algebra 2 is entirely new to these populations, this atypical curve is not unexpected. As students continue to learn these required skills, it is probable that this SEM curve will shift to reflect the expected, normally distributed SEM curve over time.

In future years, vertical lines representing the performance category cut scores will be added to each graphic. Appendix B includes scale score by scale score conditional standard errors of measurement and corresponding achievement levels for each scale score.

In classical test theory, the SEM is defined as $s_x\sqrt{1-r_{xx}}$, where $s_x$ is the standard deviation of the raw score, and $r_{xx}$ is the reliability coefficient. Under classical test theory, measurement error is assumed to be the same at all levels of achievement, and one reliability coefficient can be estimated to acknowledge that error. Standard error of measurement indicates the standard deviation of a single student's repeated test scores, if he or she were to take the same test repeatedly (with no new learning or no memory of questions taking place between test administrations). Reliability coefficients and SEM for each reporting category are also presented in Appendix A.

## 3.4 Reliability of Achievement Classification

Reliability of achievement classification is available in Volume 7.

## 3.5 Precision at Cut Scores

Table 10 through Table 12 present mean conditional standard error of measurement at each achievement level by grade and subject. These tables also include achievement level cut scores and associated conditional standard error of measurement.

*Table 10: Mean Conditional Standard Error of Measurement at each FSA Achievement Level (ELA)*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| 3 | 1 | 9.35 | | |
| 3 | 2 | 6.36 | 285 | 6 |
| 3 | 3 | 5.52 | 300 | 5 |
| 3 | 4 | 5.50 | 315 | 5 |
| 3 | 5 | 7.16 | 330 | 6 |
| 4 | 1 | 6.30 | | |
| 4 | 2 | 5.85 | 297 | 6 |
| 4 | 3 | 5.96 | 311 | 6 |
| 4 | 4 | 6.45 | 325 | 6 |
| 4 | 5 | 8.02 | 340 | 7 |
| 5 | 1 | 7.39 | | |
| 5 | 2 | 6.20 | 304 | 6 |
| 5 | 3 | 6.17 | 321 | 6 |
| 5 | 4 | 6.72 | 336 | 6 |
| 5 | 5 | 8.32 | 352 | 7 |
| 6 | 1 | 8.11 | | |
| 6 | 2 | 5.83 | 309 | 5 |
| 6 | 3 | 5.36 | 326 | 5 |
| 6 | 4 | 5.60 | 339 | 5 |
| 6 | 5 | 7.23 | 356 | 6 |
| 7 | 1 | 7.96 | | |
| 7 | 2 | 5.84 | 318 | 6 |
| 7 | 3 | 5.57 | 333 | 5 |
| 7 | 4 | 6.18 | 346 | 5 |
| 7 | 5 | 8.21 | 360 | 6 |

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| 8 | 1 | 6.67 | | |
| 8 | 2 | 5.61 | 322 | 5 |
| 8 | 3 | 5.52 | 337 | 5 |
| 8 | 4 | 6.31 | 352 | 5 |
| 8 | 5 | 7.94 | 366 | 7 |
| 9 | 1 | 7.02 | | |
| 9 | 2 | 5.45 | 328 | 5 |
| 9 | 3 | 5.40 | 343 | 5 |
| 9 | 4 | 5.99 | 355 | 5 |
| 9 | 5 | 7.47 | 370 | 6 |
| 10 | 1 | 7.37 | | |
| 10 | 2 | 5.97 | 334 | 6 |
| 10 | 3 | 5.92 | 350 | 5 |
| 10 | 4 | 6.15 | 362 | 6 |
| 10 | 5 | 7.74 | 378 | 6 |

*Table 11: Mean Conditional Standard Error of Measurement at each FSA Achievement Level (Mathematics)*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| 3 | 1 | 6.26 | | |
| 3 | 2 | 4.52 | 285 | 5 |
| 3 | 3 | 4.58 | 297 | 4 |
| 3 | 4 | 5.86 | 311 | 5 |
| 3 | 5 | 11.15 | 327 | 6 |
| 4 | 1 | 8.55 | | |
| 4 | 2 | 4.61 | 299 | 5 |
| 4 | 3 | 4.24 | 310 | 4 |
| 4 | 4 | 4.75 | 325 | 4 |
| 4 | 5 | 11.09 | 340 | 6 |
| 5 | 1 | 7.36 | | |
| 5 | 2 | 4.62 | 306 | 5 |
| 5 | 3 | 4.46 | 320 | 4 |
| 5 | 4 | 5.65 | 334 | 5 |
| 5 | 5 | 11.09 | 350 | 7 |
| 6 | 1 | 10.01 | | |
| 6 | 2 | 5.50 | 310 | 4 |
| 6 | 3 | 5.04 | 325 | 4 |
| 6 | 4 | 5.16 | 339 | 4 |
| 6 | 5 | 7.40 | 356 | 6 |
| 7 | 1 | 9.57 | | |
| 7 | 2 | 5.44 | 316 | 5 |
| 7 | 3 | 4.53 | 330 | 4 |
| 7 | 4 | 4.46 | 346 | 4 |
| 7 | 5 | 5.92 | 360 | 5 |
| 8 | 1 | 10.10 | | |
| 8 | 2 | 7.29 | 322 | 6 |
| 8 | 3 | 6.36 | 337 | 5 |

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| 8 | 4 | 5.56 | 353 | 6 |
| 8 | 5 | 5.64 | 365 | 6 |

*Table 12: Mean Conditional Standard Error of Measurement at each Achievement Level (EOC)*

| Grade | Achievement Level | Mean CSEM | Cut Score (Scale Score) | CSEM at Cut Score |
|---|---|---|---|---|
| Algebra_1 | 1 | 26.75 | | |
| Algebra_1 | 2 | 9.67 | 487 | 7 |
| Algebra_1 | 3 | 6.73 | 497 | 6 |
| Algebra_1 | 4 | 5.16 | 518 | 5 |
| Algebra_1 | 5 | 4.86 | 532 | 5 |
| Algebra_2 | 1 | 29.91 | | |
| Algebra_2 | 2 | 6.55 | 497 | 6 |
| Algebra_2 | 3 | 4.59 | 511 | 4 |
| Algebra_2 | 4 | 4.04 | 529 | 4 |
| Algebra_2 | 5 | 4.50 | 537 | 4 |
| Geometry | 1 | 13.77 | | |
| Geometry | 2 | 6.58 | 486 | 6 |
| Geometry | 3 | 5.32 | 499 | 5 |
| Geometry | 4 | 4.45 | 521 | 4 |
| Geometry | 5 | 5.47 | 533 | 4 |

## 3.6 WRITING PROMPTS INTER-RATER RELIABILITY

Writing prompts were handscored by two human raters in grades 4 through 7, and grade 10. For the online tests, prompts were scored by one human rater, and American Institutes for Research's (AIR) scoring engine was used to provide the second score.

The basic method to compute inter-rater reliability is percent agreement. As seen in Table 13, the percentage of exact agreement (when two raters gave the same score), the percentage of adjacent ratings (when the difference between two raters was 1), and the percentage of non-adjacent ratings (when the difference was larger than 1) were all computed. In this example, the exact agreement was 2/4, 50%, and the adjacent and non-adjacent percentages were 25% each.

*Table 13: Percent Agreement Example*

| Response | Rater 1 | Rater 2 | Agreement |
|---|---|---|---|
| 1 | 2 | 3 | 1 |
| 2 | 1 | 1 | 0 |
| 3 | 2 | 2 | 0 |
| 4 | 2 | 0 | 2 |

Likewise, inter-rater reliability monitors how often scorers are in exact agreement with each other and ensures that an acceptable agreement rate is maintained. The calculations for inter-rater reliability in this report are as follows:

- **Percent Exact**: total number of responses by scorer in which scores are equally divided by the number of responses that were scored twice.
- **Percent Adjacent**: total number of responses by scorer in which scores are one score point apart divided by the number of responses that were scored twice.
- **Percent Non-Adjacent**: total number of responses by scorer where scores are more than one score point apart divided by the number of responses that were scored twice, when applicable.

Table 14 displays rater-agreement percentages. The percentage of exact agreement between two raters ranged from 62 to 82. The percentage of non-adjacent rating was between 17 and 36. The non-adjacent percentages fell between 0 and 3.

Grades 8 and 9 had separate percent-agreement values for paper and online administrations. Students who were administered paper prompts in grade 10 had the same scoring rule as online students. Thus, they were reported together on the handscoring reports. The total number of processed responses does not necessarily correspond to the number of students participating in the Writing portion. These numbers could potentially be higher, as some students are scored more than once when rescoring for some responses, as requested.

*Table 14: Inter-Rater Reliability*

| Grade | Item ID | Dimension | Paper administration | | | | Online administration | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | % Exact | % Adjacent | % Not Adjacent | Total Number of Processed Responses | % Exact | % Adjacent | % Not Adjacent | Total Number of Processed Responses |
| 4 | 18474 | Purpose, Focus, & Organization | 70 | 28 | 1 | 401,334 | | | | |
| | | Evidence & Elaboration | 67 | 32 | 2 | 401,334 | | | | |
| | | Conventions | 74 | 25 | 1 | 401,334 | | | | |
| 5 | 18449 | Purpose, Focus, & Organization | 63 | 35 | 2 | 399,412 | | | | |
| | | Evidence & Elaboration | 62 | 35 | 3 | 399,412 | | | | |
| | | Conventions | 81 | 18 | 1 | 399,412 | | | | |
| 6 | 18448 | Purpose, Focus, & Organization | 69 | 30 | 1 | 395,266 | | | | |
| | | Evidence & Elaboration | 69 | 29 | 1 | 395,266 | | | | |
| | | Conventions | 82 | 17 | 1 | 395,266 | | | | |

| Grade | Item ID | Dimension | Paper administration | | | | Online administration | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | % Exact | % Adjacent | % Not Adjacent | Total Number of Processed Responses | % Exact | % Adjacent | % Not Adjacent | Total Number of Processed Responses |
| 7 | 19146 | Purpose, Focus, & Organization | 63 | 35 | 2 | 395,566 | | | | |
| | | Evidence & Elaboration | 66 | 33 | 1 | 395,566 | | | | |
| | | Conventions | 74 | 25 | 1 | 395,566 | | | | |
| 8 | 18463 | Purpose, Focus, & Organization | 67 | 33 | 1 | 2,410 | 69 | 31 | 0 | 245,371 |
| | | Evidence & Elaboration | 69 | 30 | 1 | 2,410 | 72 | 28 | 0 | 245,371 |
| | | Conventions | 72 | 27 | 1 | 2,410 | 81 | 19 | 0 | 245,371 |
| 9 | 19131 | Purpose, Focus, & Organization | 64 | 34 | 2 | 2,656 | 71 | 29 | 0 | 250,455 |
| | | Evidence & Elaboration | 69 | 30 | 1 | 2,656 | 75 | 25 | 0 | 250,455 |
| | | Conventions | 68 | 31 | 1 | 2,656 | 78 | 22 | 0 | 250,455 |
| 10* | 18475 | Purpose, Focus, & Organization | | | | | 62 | 36 | 2 | 399,052 |
| | | Evidence & Elaboration | | | | | 66 | 33 | 1 | 399,052 |
| | | Conventions | | | | | 74 | 25 | 0 | 399,052 |

*Students who were administered the Grade 10 Writing prompt on paper had the same scoring rule as those who were administered the prompt in the online environment, so these data are reported together on the handscoring reports.

In addition to inter-rater reliability, validity coefficients were also calculated. Validity coefficients indicate how often scorers are in exact agreement with previously scored selected responses that are inserted into the scoring queue, and they ensure that an acceptable agreement rate is maintained. The calculations are as follows:

- **Percent Exact**: total number of responses by scorer where scores are equal divided by the total number of responses that were scored.
- **Percent Adjacent**: total number of responses by scorer where scores are one point apart divided by the total number of responses that were scored.
- **Percent Non-Adjacent**: total number of responses by scorer where scores are more than one score point apart divided by the total number of responses that were scored.

Table 15 presents final validity coefficients, which were between 75 and 91.

*Table 15: Validity Coefficients*

| Grade | Purpose, Focus, & Organization | Evidence & Elaboration | Conventions |
|---|---|---|---|
| 4 | 78 | 75 | 83 |
| 5 | 76 | 76 | 83 |
| 6 | 85 | 85 | 91 |
| 7 | 81 | 82 | 82 |
| 8 | 83 | 83 | 84 |
| 9 | 80 | 82 | 86 |
| 10 | 76 | 80 | 84 |

Cohen's kappa (Cohen, 1968) is an index of inter-rater agreement after accounting for the agreement that could be expected due to chance. This statistic can be computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where $P_o$ is the proportion of observed agreement, and $P_c$ indicates the proportion of agreement by chance. Cohen's kappa treats all disagreement values with equal weights. Weighted kappa coefficients (Cohen, 1968), however, allow unequal weights, which can be used as a measure of validity. Weighted kappa coefficients were calculated using the formula below:

$$K_w = \frac{P'_o - P'_c}{1 - P'_c},$$

where

$$P'_o = \frac{\sum w_{ij} p_{oij}}{w_{max}},$$

$$P'_c = \frac{\sum w_{ij} p_{cij}}{w_{max}},$$

where $p_{oij}$ is the proportion of the judgments observed in the ijth cell, $p_{cij}$ is the proportion in the ijth cell expected by chance, and $w_{ij}$ is the disagreement weight.

Weighted kappa coefficients for grades 4 through 10 operational writing prompts by dimension are presented in Table 16. They ranged from 0.49 to 0.71.

*Table 16: Weighted Kappa Coefficients*

| Grade | Scorer | Purpose, Focus, & Organization | Evidence & Elaboration | Conventions |
|---|---|---|---|---|
| 4 | Two Human Raters | 0.63 | 0.60 | 0.53 |
| 5 | Two Human Raters | 0.56 | 0.54 | 0.49 |

| Grade | Scorer | Purpose, Focus, & Organization | Evidence & Elaboration | Conventions |
|---|---|---|---|---|
| 6 | Two Human Raters | 0.62 | 0.61 | 0.54 |
| 7 | Two Human Raters | 0.61 | 0.63 | 0.51 |
| 8 | Machine and Human | 0.61 | 0.63 | 0.63 |
| | Two Human Raters | 0.68 | 0.68 | 0.54 |
| 9 | Machine and Human | 0.64 | 0.65 | 0.65 |
| | Two Human Raters | 0.64 | 0.71 | 0.60 |
| 10 | Two Human Raters | 0.65 | 0.67 | 0.54 |

Grades 8, 9, and 10 Writing prompts were administered online. Grade 10 was scored by two scorers, while students in grades 8 and 9 received one human score and one machine score through AIR's artificial intelligence (AI) scoring engine. To train AIR's AI scoring engine, a subset of papers was scientifically selected and scored by two human raters. Score discrepancies were resolved before being sent from DRC to AIR. The subset was split into a training set with 1,500 papers, and a validation set with 610 papers. The scoring engine used probit regression with a final resolved human score as the outcome. The training was done by trait or dimension, which varied on Latent Semantic Analysis (LSA) dimensions and independent variables such as punctuation and verb diversity. Coefficients from the LSA dimensions and independent variables from the training set were used to calculate predicted scores for the validation set of papers. Scores from the validation set were compared to human scores to ensure that the computer-to-human agreement rate was at least as reasonable as the human-to-human agreement rate. If the scoring engine results were not comparable, LSA dimensions and/or independent variables were updated and models were rerun. The total number of LSA dimensions used was 50 for both grades 8 and 9. The numbers of independent variables were 69 and 70 for grades 8 and 9, respectively. Table 17 shows that the scoring engine produced comparable results with human scores (see Volume 7 for details).

*Table 17: Percent Agreement in Handscoring and Scoring Engine*

| Grade | Dimension | Handscoring from DRC | | | AIR Scoring Engine | | |
|---|---|---|---|---|---|---|---|
| | | % Exact | % Adjacent | % Not Adjacent | % Exact | % Adjacent | % Not Adjacent |
| 8 | Purpose, Focus, & Organization | 67 | 32 | 1 | 68.69 | 30.66 | 0.66 |
| | Evidence & Elaboration | 76 | 24 | 1 | 71.97 | 27.87 | 0.16 |
| | Conventions | 69 | 30 | 1 | 80.82 | 18.69 | 0.49 |

| Grade | Dimension | Handscoring from DRC | | | AIR Scoring Engine | | |
|---|---|---|---|---|---|---|---|
| | | % Exact | % Adjacent | % Not Adjacent | % Exact | % Adjacent | % Not Adjacent |
| 9 | Purpose, Focus, & Organization | 65 | 34 | 1 | 75.57 | 24.1 | 0.33 |
| | Evidence & Elaboration | 73 | 26 | 1 | 72.13 | 27.54 | 0.33 |
| | Conventions | 72 | 28 | 0 | 76.72 | 22.79 | 0.49 |

Table 18 compares the correlations between the scoring engine and the human rater with the correlations between two human raters. The quadratic weighted kappa coefficients are presented in Table 16, which also shows the agreement between the scoring engine and human raters and the agreement between two human scorers. The results demonstrate that scores from the scoring engine were comparable to scores given by two human raters.

*Table 18: Correlations between Scores from Scoring Engine and from Human Raters*

| Grade | Dimension | Machine and Human | | Human and Human | |
|---|---|---|---|---|---|
| | | Pearson Correlations | Spearman Correlations | Pearson Correlations | Spearman Correlations |
| 8 | Purpose, Focus, & Organization | 0.61 | 0.58 | 0.68 | 0.68 |
| | Evidence & Elaboration | 0.64 | 0.63 | 0.68 | 0.68 |
| | Conventions | 0.64 | 0.63 | 0.54 | 0.51 |
| 9 | Purpose, Focus, & Organization | 0.64 | 0.61 | 0.64 | 0.63 |
| | Evidence & Elaboration | 0.66 | 0.66 | 0.71 | 0.72 |
| | Conventions | 0.66 | 0.66 | 0.59 | 0.56 |

# 4. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the FSA were representative of the content standards of the larger knowledge domain. We describe the content standards for FSA and discuss the test development process, mapping FSA tests to the standards. A complete description of the test development process can be found in Volume 2, Test Development.

## 4.1 CONTENT STANDARDS

The FSA was aligned to the Florida Standards, which were approved by the Florida State Board of Education on February 18, 2014, to be the educational standards for all public schools in the state. The Florida Standards are intended to implement higher standards, with the goal of challenging and motivating Florida's students to acquire stronger critical thinking, problem solving, and communications skills. The Language Arts Florida Standards (LAFS) and the Mathematics Florida Standards (MAFS) are available for review at www.flstandards.org.

Table 19, Table 20, and Table 21 present the reporting categories by grade and test, as well as the number of items measuring each category.

*Table 19: Number of Items for Each ELA Reporting Category*

| Reporting Category | Grade | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Key Ideas and Details | 10 | 11 | 12 | 11 | 9 | 13 | 12 | 12 |
| Craft and Structure | 16 | 18 | 16 | 18 | 18 | 16 | 17 | 17 |
| Integration of Knowledge and Ideas | 13 | 11 | 14 | 15 | 12 | 12 | 14 | 15 |
| Language and Editing Task | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| Text-Based Writing | | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

*Table 20: Number of Items for Each Mathematics Reporting Category*

| Grade | Reporting Category | Number of Items |
|---|---|---|
| 3 | Operations, Algebraic Thinking, and Numbers in Base Ten | 26 |
| | Numbers and Operations – Fractions | 8 |
| | Measurement, Data, and Geometry | 19 |
| 4 | Operations and Algebraic Thinking | 11 |
| | Numbers and Operations in Base Ten | 11 |
| | Numbers and Operations – Fractions | 14 |
| | Measurement, Data, and Geometry | 18 |
| 5 | Operations, Algebraic Thinking, and Fractions | 21 |
| | Numbers and Operations in Base Ten | 15 |
| | Measurement, Data, and Geometry | 17 |
| 6 | Ratio and Proportional Relationships | 8 |
| | Expressions and Equations | 16 |
| | Geometry | 8 |
| | Statistics and Probability | 10 |
| | The Number System | 12 |
| 7 | Ratio and Proportional Relationships | 14 |
| | Expressions and Equations | 12 |
| | Geometry | 13 |
| | Statistics and Probability | 9 |
| | The Number System | 7 |
| 8 | Expressions and Equations | 16 |
| | Functions | 13 |
| | Geometry | 13 |
| | Statistics & Probability and the Number System | 9 |

*Table 21: Number of Items for Each EOC Reporting Category*

| Course | Reporting Category | Core Form | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Algebra 1 | Algebra and Modeling | 22 | 22 | 22 | 22 |
| | Functions and Modeling | 21 | 21 | 21 | 21 |
| | Statistics and the Number System | 8 | 8 | 8 | 8 |
| Algebra 2 | Algebra and Modeling | 21 | 21 | | |
| | Functions and Modeling | 18 | 17 | | |
| | Statistics, Probability, and the Number System | 15 | 15 | | |
| Geometry | Congruence, Similarity, Right Triangles and Trigonometry | 25 | 25 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations | 17 | 17 | | |
| | Modeling with Geometry | 8 | 8 | | |

## 4.2 TEST SPECIFICATIONS

Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards that they were intended to measure. For more detail, please see Volume 2, Section 2. The Florida Standards Assessments (FSA) were composed of test items that included traditional multiple-choice items, items that required students to type or write a response, and technology-enhanced items (TEI). Technology-enhanced items are computer-delivered items that require students to interact with test content to select, construct, and support their answers. The blueprints specified the percentage of operational items that were to be administered. The blueprints also included the minimum and maximum number of items for each of the reporting categories, and constraints on selecting items for the depth of knowledge (DOK) levels in Reading. The minimum and maximum number of items by grade and subject and other details on the blueprint are presented in appendices of Volume 2.

## 4.3 TEST DEVELOPMENT

For the 2015 Florida Standards Assessments administration, American Institutes for Research in collaboration with the Florida Department of Education and its Test Development Center (TDC), constructed test forms for ELA grades 3 through 10 and grade 10 retake, Mathematics grades 3 through 8, and End-of-Course Assessments (Algebra 1, Algebra 2, Geometry).

Test construction began during the summer of 2014, when all parties gathered in Washington, DC, to identify items from the Utah Student Assessment of Growth and Excellence (SAGE) item bank that aligned to the FSA standards and blueprints designed for the FSA. Curricular, psychometric, and policy experts constructed test forms carefully, evaluating the fit of each item's statistical characteristics and the alignment of the item to the Florida's standards. The content guidelines, which describe standards coverage and item type coverage, are outlined in detail in Appendices A and B of Volume 2, Test Development.

The Florida Standards Assessments item pool grows each year by field testing new items. Any item used on an assessment was field tested before it was used as an operational item. In spring 2015, field test items were embedded on online forms. Future FSA items were not being field tested on paper, so there were no field test items in grades 3 and 4. The following tests and grades included field test items:

- Grades 5 through 10 in ELA;
- Grades 5 through 8 in Mathematics; and
- End-of-Course Assessments (Algebra 1, Algebra 2, Geometry).

With FSA, field testing was conducted during the spring as a part of the regular administration. The field test items utilized the same positions as anchor items. In order to keep the test length consistent, placeholder items were placed into the field test positions on some of the forms. The number of forms constructed for a given grade and subject was at most 40, including field test and anchor forms.

After operational forms were developed, the AIR and TDC content specialists worked together to assign newly developed items to field test forms for field testing. The teams addressed the following factors when embedding field test items into operational test forms for the spring administration:

- Ensured field test items did not cue or clue answers to other field test items on the form.
- Ensured field test items that cued or clued answers to operational items were not field tested.
- Included a mix of items covering multiple reporting categories and standards on each form.
- Selected items in the field test sets that reflect a range of difficulty levels and cognitive levels.
- Minimized abrupt transitions from one subject strand or mental construct to another.
- Selected items that were needed for appropriate standard coverage in the item bank.
- Selected items that were needed for appropriate format variety in the item bank.
- Maintained awareness of the distribution of keys and the number of adjacent items having the same key.

## 4.4 Alignment of FSA Item Banks to the Content Standards and Benchmarks

A third party, independent alignment study has not yet been completed; those results will be summarized here upon completion and are expected to be completed in May 2016. A full report on alignment will be provided in the 2016 technical reports.

While a comprehensive alignment study has not yet been completed, a small alignment review of the 2015 test items was completed by Alpine Testing Solutions in a subset of the tested FSA grades. That report can be found online at https://feaweb.org/_data/files/2015_DOE/FSA_Validity_Study/FSA-Final-Report_08312015.pdf. Additionally, a summary of the major

findings from Alpine were presented to the Florida State Senate, and in their final presentation to the Education Pre-K to 12 Committee, Alpine reported that their independent review found that all but two of the items reviewed aligned to a specific Florida Standard. The Alpine presentation to the Senate Committee is found in Volume 7.

# 5. EVIDENCE ON INTERNAL STRUCTURE

In this section, we explore the internal structure of the assessment using the scores provided at the reporting category level. The relationship of the subscores is just one indicator of the test dimensionality.

In ELA grades 4 through 10, there are five reporting categories per grade: Key Ideas and Details, Craft and Structure, Integration of Knowledge and Ideas, Language and Editing Task, and Text-Based Writing. Reading grade 3 has the same reporting categories, with the exception of Text-Based Writing. In Mathematics and EOC tests, reporting categories differ in each grade or course (see Table 19, Table 20, and Table 21 for reporting category information).

Raw scores based on each reporting category were provided to students. Evidence is needed to verify that the raw score for each reporting category provides both different and useful information for student achievement.

It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional IRT model difficult, although we could then easily justify reporting these separate scores. On the contrary, if the reporting categories were perfectly correlated, we could justify a unidimensional model, but we could not justify the reporting of separate scores.

One pathway to explore the internal structure of the test is via a second-order factor model, assuming a general mathematics construct (first factor) with reporting categories (second factor), and that the items load onto the reporting category they intend to measure. If the first-order factors are highly correlated and the model fits for the second-order model, this provides evidence of unidimensionality as well as reporting subscores.

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following section.

## 5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 22 through Table 24 present the observed correlation matrix of the reporting category raw scores for each subject area. In ELA, the correlations among the reporting categories range from 0.45 to 0.75. The Language and Editing Task items and Text-Based Writing items exhibited slightly lower correlations with the other reporting categories ranging from 0.45 to 0.61. For Mathematics, the correlations were between 0.54 and 0.83. For EOC, the correlations between the three subscales fell between 0.69 and 0.83. Observed correlations from the accommodated forms are presented in Table 25 through Table 27. The correlations varied between 0.35 and 0.74 for ELA, 0.41 and 0.74 for Mathematics, and 0.51 and 0.74 for EOC.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level,

given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously, which the Department cautions each year when scores are released.

Table 28 through Table 33 display disattenuated correlations. In ELA, the Writing dimension had the lowest correlations among the five reporting categories. For the Writing dimension, the average value was 0.79 and the minimum was 0.66, whereas the overall average disattenuated correlation for ELA was 0.90.

*Table 22: Correlation Matrix among Reporting Categories (ELA)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 3 | Key Ideas and Details (Cat1) | 10 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.71 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 13 | 0.65 | 0.68 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.52 | 0.58 | 0.49 | 1.00 | |
| 4 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.66 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 11 | 0.64 | 0.68 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.48 | 0.53 | 0.49 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.51 | 0.54 | 0.53 | 0.45 | 1.00 |
| 5 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.65 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 0.70 | 0.67 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.54 | 0.56 | 0.55 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.49 | 0.49 | 0.50 | 0.45 | 1.00 |
| 6 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.74 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 0.73 | 0.75 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.58 | 0.60 | 0.58 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.55 | 0.57 | 0.56 | 0.47 | 1.00 |
| 7 | Key Ideas and Details (Cat1) | 9 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.73 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.63 | 0.70 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.51 | 0.55 | 0.47 | 1.00 | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 7 (cont'd.) | Text-Based Writing (Cat5) | 3 | 0.56 | 0.61 | 0.53 | 0.51 | 1.00 |
| 8 | Key Ideas and Details (Cat1) | 13 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.73 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.74 | 0.72 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.58 | 0.59 | 0.58 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.58 | 0.58 | 0.58 | 0.53 | 1.00 |
| 9 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 0.74 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 0.69 | 0.71 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.54 | 0.56 | 0.54 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.60 | 0.61 | 0.59 | 0.55 | 1.00 |
| 10 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 0.69 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 0.70 | 0.72 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.48 | 0.50 | 0.50 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.57 | 0.57 | 0.59 | 0.48 | 1.00 |

*Table 23: Correlation Matrix among Reporting Categories (Mathematics)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 3 | Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1) | 26 | 1.00 | | | | |
| | Numbers and Operations – Fractions (Cat2) | 8 | 0.72 | 1.00 | | | |
| | Measurement, Data, and Geometry (Cat3) | 19 | 0.83 | 0.72 | 1.00 | | |
| 4 | Operations and Algebraic Thinking (Cat1) | 11 | 1.00 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 11 | 0.75 | 1.00 | | | |
| | Numbers and Operations – Fractions (Cat3) | 14 | 0.76 | 0.74 | 1.00 | | |
| | Measurement, Data, and Geometry (Cat4) | 18 | 0.75 | 0.73 | 0.78 | 1.00 | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 5 | Operations, Algebraic Thinking, and Fractions (Cat1) | 21 | 1.00 | 0.82 | 0.77 | | |
| | Numbers and Operations in Base Ten (Cat2) | 15 | 0.82 | 1.00 | 0.76 | | |
| | Measurement, Data, and Geometry (Cat3) | 17 | 0.77 | 0.76 | 1.00 | | |
| 6 | Ratio and Proportional Relationships (Cat1) | 8 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 16 | 0.73 | 1.00 | | | |
| | Geometry (Cat3) | 8 | 0.61 | 0.65 | 1.00 | | |
| | Statistics and Probability (Cat4) | 10 | 0.63 | 0.66 | 0.58 | 1.00 | |
| | The Number System (Cat5) | 12 | 0.72 | 0.76 | 0.64 | 0.64 | 1.00 |
| 7 | Ratio and Proportional Relationships (Cat1) | 14 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 12 | 0.75 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.75 | 0.73 | 1.00 | | |
| | Statistics and Probability (Cat4) | 9 | 0.68 | 0.65 | 0.64 | 1.00 | |
| | The Number System (Cat5) | 7 | 0.64 | 0.63 | 0.61 | 0.59 | 1.00 |
| 8 | Expressions and Equations (Cat1) | 16 | 1.00 | | | | |
| | Functions (Cat2) | 13 | 0.63 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.63 | 0.57 | 1.00 | | |
| | Statistics & Probability and the Number System (Cat4) | 9 | 0.58 | 0.54 | 0.59 | 1.00 | |

*Table 24: Correlation Matrix among Reporting Categories (EOC)*

| Course/Form | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 1/Core 1 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.83 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.75 | 0.76 | 1.00 |
| Algebra 1/Core 2 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.80 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.71 | 0.69 | 1.00 |

| Course/Form | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 1/Core 3 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.81 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.70 | 0.74 | 1.00 |
| Algebra 1/Core 4 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.79 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.71 | 0.69 | 1.00 |
| Algebra 2/Core 1 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 18 | 0.76 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 0.79 | 0.74 | 1.00 |
| Algebra 2/Core 2 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 17 | 0.75 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 0.79 | 0.73 | 1.00 |
| Geometry/Core 1 | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.78 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.69 | 0.72 | 1.00 |
| Geometry/Core 2 | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.78 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.70 | 0.71 | 1.00 |

*Table 25: Correlation Matrix among Reporting Categories (ELA Accommodated Forms)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 5 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.57 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 0.62 | 0.61 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.47 | 0.49 | 0.47 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.43 | 0.44 | 0.42 | 0.42 | 1.00 |
| 6 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.70 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 0.70 | 0.68 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.58 | 0.57 | 0.54 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.53 | 0.51 | 0.51 | 0.44 | 1.00 |
| 7 | Key Ideas and Details (Cat1) | 9 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.71 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.59 | 0.62 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.54 | 0.53 | 0.42 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.57 | 0.58 | 0.43 | 0.46 | 1.00 |
| 8 | Key Ideas and Details (Cat1) | 13 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.74 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.69 | 0.70 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.51 | 0.53 | 0.46 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.52 | 0.52 | 0.49 | 0.45 | 1.00 |
| 9 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 0.72 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 0.63 | 0.64 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.48 | 0.54 | 0.46 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.51 | 0.52 | 0.45 | 0.47 | 1.00 |
| 10 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 10 (cont'd.) | Craft and Structure (Cat2) | 17 | 0.70 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 0.69 | 0.66 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.44 | 0.48 | 0.47 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.50 | 0.51 | 0.50 | 0.35 | 1.00 |

*Table 26: Correlation Matrix among Reporting Categories (Mathematics Accommodated Forms)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 5 | Operations, Algebraic Thinking, and Fractions (Cat1) | 21 | 1.00 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 15 | 0.74 | 1.00 | | | |
| | Measurement, Data, and Geometry (Cat3) | 17 | 0.70 | 0.66 | 1.00 | | |
| 6 | Ratio and Proportional Relationships (Cat1) | 8 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 16 | 0.61 | 1.00 | | | |
| | Geometry (Cat3) | 8 | 0.43 | 0.50 | 1.00 | | |
| | Statistics and Probability (Cat4) | 10 | 0.53 | 0.55 | 0.44 | 1.00 | |
| | The Number System (Cat5) | 12 | 0.60 | 0.66 | 0.48 | 0.54 | 1.00 |
| 7 | Ratio and Proportional Relationships (Cat1) | 14 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 12 | 0.63 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.65 | 0.61 | 1.00 | | |
| | Statistics and Probability (Cat4) | 9 | 0.65 | 0.57 | 0.57 | 1.00 | |
| | The Number System (Cat5) | 7 | 0.59 | 0.61 | 0.54 | 0.56 | 1.00 |
| 8 | Expressions and Equations (Cat1) | 16 | 1.00 | | | | |
| | Functions (Cat2) | 13 | 0.51 | 1.00 | | | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 8 (cont'd.) | Geometry (Cat3) | 13 | 0.53 | 0.46 | 1.00 | | |
| | Statistics & Probability and the Number System (Cat4) | 9 | 0.46 | 0.41 | 0.53 | 1.00 | |

*Table 27: Correlation Matrix among Reporting Categories (EOC Accommodated Forms)*

| Course | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 1 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.67 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.51 | 0.52 | 1.00 |
| Algebra 2 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 18 | 0.66 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 0.68 | 0.62 | 1.00 |
| Geometry | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.74 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.65 | 0.64 | 1.00 |

*Table 28: Disattenuated Correlation Matrix among Reporting Categories (ELA)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 3 | Key Ideas and Details (Cat1) | 10 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.99 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 13 | 0.98 | 0.98 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.79 | 0.83 | 0.77 | 1.00 | |
| 4 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.97 | 1.00 | | | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 4 (cont'd.) | Integration of Knowledge and Ideas (Cat3) | 11 | 0.98 | 0.98 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.81 | 0.85 | 0.81 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.71 | 0.71 | 0.73 | 0.68 | 1.00 |
| 5 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.97 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 1.00* | 0.98 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.86 | 0.88 | 0.86 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.67 | 0.67 | 0.68 | 0.66 | 1.00 |
| 6 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 1.00* | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 1.00 | 1.00 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.90 | 0.91 | 0.88 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.72 | 0.71 | 0.71 | 0.67 | 1.00 |
| 7 | Key Ideas and Details (Cat1) | 9 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 1.00* | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 1.00* | 1.00 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.86 | 0.83 | 0.81 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.77 | 0.75 | 0.75 | 0.75 | 1.00 |
| 8 | Key Ideas and Details (Cat1) | 13 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 1.00 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.99 | 1.00 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.89 | 0.94 | 0.91 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.72 | 0.75 | 0.73 | 0.77 | 1.00 |
| 9 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 0.99 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 0.98 | 0.96 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.86 | 0.87 | 0.88 | 1.00 | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 9 (cont'd.) | Text-Based Writing (Cat5) | 3 | 0.77 | 0.75 | 0.76 | 0.81 | 1.00 |
| 10 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 1.00* | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 1.00* | 1.00 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.88 | 0.87 | 0.87 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.77 | 0.74 | 0.76 | 0.78 | 1.00 |

*Table 29: Disattenuated Correlation Matrix among Reporting Categories (Mathematics)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 3 | Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1) | 26 | 1.00 | | | | |
| | Numbers and Operations – Fractions (Cat2) | 8 | 0.94 | 1.00 | | | |
| | Measurement, Data, and Geometry (Cat3) | 19 | 0.99 | 0.95 | 1.00 | | |
| 4 | Operations and Algebraic Thinking (Cat1) | 11 | 1.00 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 11 | 0.98 | 1.00 | | | |
| | Numbers and Operations – Fractions (Cat3) | 14 | 0.96 | 0.95 | 1.00 | | |
| | Measurement, Data, and Geometry (Cat4) | 18 | 0.94 | 0.93 | 0.97 | 1.00 | |
| 5 | Operations, Algebraic Thinking, and Fractions (Cat1) | 21 | 1.00 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 15 | 0.98 | 1.00 | | | |
| | Measurement, Data, and Geometry (Cat3) | 17 | 0.94 | 0.95 | 1.00 | | |
| 6 | Ratio and Proportional Relationships (Cat1) | 8 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 16 | 0.97 | 1.00 | | | |
| | Geometry (Cat3) | 8 | 0.93 | 0.94 | 1.00 | | |
| | Statistics and Probability (Cat4) | 10 | 0.96 | 0.94 | 0.96 | 1.00 | |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 6 (cont'd.) | The Number System (Cat5) | 12 | 0.98 | 0.98 | 0.96 | 0.95 | 1.00 |
| 7 | Ratio and Proportional Relationships (Cat1) | 14 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 12 | 0.96 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.96 | 0.96 | 1.00 | | |
| | Statistics and Probability (Cat4) | 9 | 0.92 | 0.92 | 0.91 | 1.00 | |
| | The Number System (Cat5) | 7 | 0.92 | 0.94 | 0.92 | 0.94 | 1.00 |
| 8 | Expressions and Equations (Cat1) | 16 | 1.00 | | | | |
| | Functions (Cat2) | 13 | 0.96 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.90 | 0.91 | 1.00 | | |
| | Statistics & Probability and the Number System (Cat4) | 9 | 0.89 | 0.90 | 0.93 | 1.00 | |

*Table 30: Disattenuated Correlation Matrix among Reporting Categories (EOC)*

| Course/Form | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 1/Core 1 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 1.00* | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 1.00* | 1.10 | 1 |
| Algebra 1/Core 2 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.98 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 1.00* | 1.00* | 1 |
| Algebra 1/Core 3 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.99 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 1.00* | 1.00* | 1 |
| Algebra 1/Core 4 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.98 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.98 | 1.00 | 1 |
| Algebra 2/Core 1 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 18 | 0.96 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 0.95 | 0.97 | 1 |

| Course/Form | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 2/Core 2 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 17 | 0.96 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 0.94 | 0.94 | 1 |
| Geometry/Core 1 | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.91 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.87 | 0.92 | 1 |
| Geometry/Core 2 | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.90 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.87 | 0.91 | 1 |

*Table 31: Disattenuated Correlation Matrix among Reporting Categories (ELA Accommodated Forms)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 5 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 0.95 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 1.00* | 0.97 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.83 | 0.82 | 0.80 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.63 | 0.61 | 0.59 | 0.63 | 1.00 |
| 6 | Key Ideas and Details (Cat1) | 11 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.99 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 1.00* | 0.99 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.92 | 0.91 | 0.89 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.68 | 0.66 | 0.67 | 0.64 | 1.00 |

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 7 | Key Ideas and Details (Cat1) | 9 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 18 | 0.97 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.97 | 0.96 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.82 | 0.75 | 0.72 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.76 | 0.71 | 0.64 | 0.64 | 1.00 |
| 8 | Key Ideas and Details (Cat1) | 13 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 16 | 1.00* | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 12 | 0.98 | 1.00* | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.71 | 0.78 | 0.68 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.65 | 0.69 | 0.66 | 0.60 | 1.00 |
| 9 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 0.98 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 14 | 1.00* | 0.99 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.79 | 0.83 | 0.85 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.67 | 0.65 | 0.68 | 0.71 | 1.00 |
| 10 | Key Ideas and Details (Cat1) | 12 | 1.00 | | | | |
| | Craft and Structure (Cat2) | 17 | 1.00 | 1.00 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 15 | 1.00* | 0.99 | 1.00 | | |
| | Language and Editing Task (Cat4) | 8 | 0.78 | 0.86 | 0.87 | 1.00 | |
| | Text-Based Writing (Cat5) | 3 | 0.65 | 0.67 | 0.67 | 0.57 | 1.00 |

*Table 32: Disattenuated Correlation Matrix among Reporting Categories (Mathematics Accommodated Forms)*

| Grade | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|---|
| 5 | Operations, Algebraic Thinking, and Fractions (Cat1) | 21 | 1.00 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 15 | 0.96 | 1.00 | | | |
| | Measurement, Data, and Geometry (Cat3) | 17 | 0.92 | 0.92 | 1.00 | | |
| 6 | Ratio and Proportional Relationships (Cat1) | 8 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 16 | 0.92 | 1.00 | | | |
| | Geometry (Cat3) | 8 | 0.82 | 0.90 | 1.00 | | |
| | Statistics and Probability (Cat4) | 10 | 0.90 | 0.90 | 0.91 | 1.00 | |
| | The Number System (Cat5) | 12 | 0.94 | 0.98 | 0.90 | 0.92 | 1.00 |
| 7 | Ratio and Proportional Relationships (Cat1) | 14 | 1.00 | | | | |
| | Expressions and Equations (Cat2) | 12 | 0.93 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.98 | 0.96 | 1.00 | | |
| | Statistics and Probability (Cat4) | 9 | 0.93 | 0.85 | 0.87 | 1.00 | |
| | The Number System (Cat5) | 7 | 0.89 | 0.96 | 0.87 | 0.85 | 1.00 |
| 8 | Expressions and Equations (Cat1) | 16 | 1.00 | | | | |
| | Functions (Cat2) | 13 | 0.94 | 1.00 | | | |
| | Geometry (Cat3) | 13 | 0.89 | 0.87 | 1.00 | | |
| | Statistics & Probability and the Number System (Cat4) | 9 | 0.79 | 0.81 | 0.92 | 1.00 | |

*Table 33: Disattenuated Correlation Matrix among Reporting Categories (EOC Accommodated Forms)*

| Course | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 1 | Algebra and Modeling (Cat1) | 22 | 1.00 | | |
| | Functions and Modeling (Cat2) | 21 | 0.97 | 1.00 | |
| | Statistics and the Number System (Cat3) | 8 | 0.94 | 1.01 | 1.00 |

| Course | Reporting Category | Number of Items | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|---|
| Algebra 2 | Algebra and Modeling (Cat1) | 21 | 1.00 | | |
| | Functions and Modeling (Cat2) | 18 | 0.96 | 1.00 | |
| | Statistics, Probability, and the Number System (Cat3) | 15 | 1.01 | 1.00 | 1.00 |
| Geometry | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 25 | 1.00 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 17 | 0.96 | 1.00 | |
| | Modeling with Geometry (Cat3) | 8 | 0.92 | 0.97 | 1.00 |

## 5.2 CONFIRMATORY FACTOR ANALYSIS

The FSA had test items designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores on the various reporting categories were also provided as indices of strand-specific performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

The results in this section are intended to provide evidence that the methods for reporting FSA strand scores align with the underlying structure of the test and also provide evidence for appropriateness of the selected IRT models. This section is based on a second-order confirmatory factor analysis, in which the first order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto factors they are intended to measure. The underlying structure of the ELA and Mathematics tests was generally common across all grades, which is useful for comparing the results of our analyses across the grades.

While the test consisted of items targeting different standards, all items within a grade and subject were calibrated concurrently using the various IRT models described in this technical report. This implies the pivotal IRT assumption of local independence (Lord, 1980). Formally stated, this assumption posits that the probability of the outcome on item *i* depends only on the student's ability and the characteristics of the item. Beyond that, the score of item *i* is independent of the outcome of all other items. From this assumption, the joint density (i.e., the likelihood) is viewed as the product of the individual densities. Thus, maximum likelihood estimation of person and item parameters in traditional Item Response Theory is derived on the basis of this theory.

The measurement model and the score reporting method assume a single underlying factor, with separate factors representing each of the reporting categories. Consequently, it is important to collect validity evidence on the internal structure of the assessment to determine the rationality of conducting concurrent calibrations, as well as using these scoring and reporting methods.

### 5.2.1 Factor Analytic Methods

A series of confirmatory factor analyses (CFA) were conducted using the statistical program MPlus [version 7.31] (Muthén & Muthén, 2012) for each grade and subject assessment. The "lavaan" package (Rosseel, 2012) in R was also used for cross-validation and to supplement MPlus. These psychometric tools are used for collecting validity evidence on the internal structure of assessments. In our analysis, these two software produced comparable results, but the MPlus outputs are mainly reported in this document as it is more commonly used for conducting CFA. The CFA was performed with the R package lavaan whenever MPlus failed to converge or confronted any estimation issues. Weighted least squares means and variance adjusted (WLSMV) and weighted least squares (WLS) estimators, implemented in both psychometric tools, were used for the factor analysis. WLSMV, also referred to as the robust WLS, was the primary estimation method employed because it is less sensitive to the size of the sample and the model and is also shown to perform well with categorical variables (Muthén, du Toit, & Spisic, 1997). WLS estimator was only used when there were estimation problems under WLSMV.

As previously stated, the method of reporting scores used for the state of Florida implies separate factors for each reporting category, connected by a single underlying factor. This model is subsequently referred to as the implied model. In factor analytic terms, this suggests that test items load onto separate first-order factors, with the first-order factors connected to a single underlying second-order factor. The use of the CFA in this section establishes some validity evidence for the degree to which the implied model is reasonable.

A chi-square difference test is often applied to assess model fit. However, it is sensitive to sample size, almost always rejecting the null hypothesis when the sample size is large. Therefore, instead of conducting a chi-square difference test, other goodness-of-fit indices were used to evaluate the implied model for FSA.

If the internal structure of the test was strictly unidimensional, then the overall person ability measure, theta $(\theta)$, would be the single common factor, and the correlation matrix among test items would suggest no discernable pattern among factors. As such, there would be no empirical or logical basis to report scores for the separate performance categories. In factor analytic terms, a test structure that is strictly unidimensional implies a single-order factor model, in which all test items load onto a single underlying factor. The development below expands the first-order model to a generalized second-order parameterization to show the relationship between the models.

The factor analysis models are based on the matrix $\boldsymbol{S}$ of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix $\boldsymbol{W}$ of asymptotic covariances among these sample correlations (Jöreskog, 1994) is employed as a weight matrix in a weighted least squares estimation approach (Browne, 1984; Muthén, 1984) to minimize the fit function:

$$F_{WLS} = \text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}})' \boldsymbol{W^{-1}} \text{vech}(\boldsymbol{S} - \widehat{\boldsymbol{\Sigma}})$$

In the equation above, $\widehat{\boldsymbol{\Sigma}}$ is the implied correlation matrix, given the estimated factor model, and the function vech vectorizes a symmetric matrix. That is, vech stacks each column of the matrix to form a vector. Note that the WLSMV approach (Muthén, du Toit, & Spisic, 1997) employs a weight matrix of asymptotic variances (i.e., the diagonal of the weight matrix) instead of the full asymptotic covariances.

We posit a first-order factor analysis where all test items load onto a single common factor, as the base model. The first-order model can be mathematically represented as:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta},$$

Where $\boldsymbol{\Lambda}$ is the matrix of item factor loadings (with $\boldsymbol{\Lambda}'$ representing its transpose), and $\boldsymbol{\Theta}$ is the uniqueness, or measurement error. The matrix $\boldsymbol{\Phi}$ is the correlation among the separate factors. For the base model, items are thought only to load onto a single underlying factor. Hence $\boldsymbol{\Lambda}$ is a *p x 1* vector, where *p* is the number of test items and $\boldsymbol{\Phi}$ is a scalar equal to 1. Therefore, it is possible to drop the matrix $\boldsymbol{\Phi}$ from the general notation. However, this notation is retained to more easily facilitate comparisons to the implied model, such that it can subsequently be viewed as a special case of the second-order factor analysis.

For the implied model, we posit a second-order factor analysis in which test items are coerced to load onto the reporting categories they are designed to target, and all reporting categories share a common underlying factor. The second-order factor analysis can be mathematically represented as:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}(\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}' + \boldsymbol{\Theta},$$

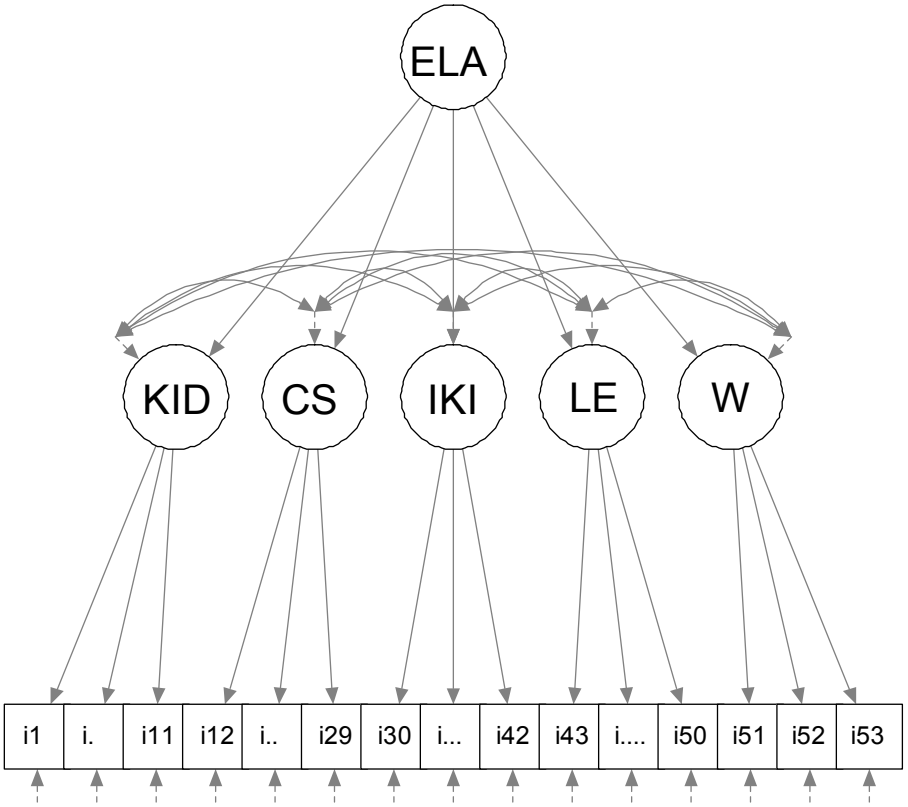
Where $\hat{\boldsymbol{\Sigma}}$ is the implied correlation matrix among test items, $\boldsymbol{\Lambda}$ is the *p x k* matrix of first-order factor loadings relating item scores to first-order factors, $\boldsymbol{\Gamma}$ is the *k x 1* matrix of second-order factor loadings relating the first-order factors to the second-order factor with *k* denoting the number of factors, $\boldsymbol{\Phi}$ is the correlation matrix of the second-order factors, and $\boldsymbol{\Psi}$ is the matrix of first-order factor residuals. All other notation is the same as the first-order model. Note that the second-order model expands the first-order model such that $\boldsymbol{\Phi} \rightarrow \boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi}$. As such, the first-order model is said to be nested within the second-order model.

There is a separate factor for each of 4–5 reporting categories for ELA, 3–5 categories for Mathematics, and 3 categories for EOC (see Table 19, Table 20, and Table 21 for reporting category information). Therefore, the number of rows in $\boldsymbol{\Gamma}$ (*k)* differs between subjects, but the general structure of the factor analysis is consistent across ELA and Mathematics.

The second-order factor model can also be represented graphically and a sample of the generalized approaches is provided on the following page. The general structure of the second-order factor analysis for ELA is illustrated in Figure 5. This figure is generally representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across the grades.

The purpose of conducting confirmatory factor analysis for the FSA was to provide evidence that each individual assessment in the FSA implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

*Figure 5: Second-Order Factor Model (ELA)*




## 5.2.2 Results

Several goodness-of-fit statistics from each of the analyses are presented in the tables below. Table 34 presents the summary results obtained from confirmatory factor analysis. Three goodness-of-fit indices were used to evaluate model fit of the item parameters to the manner in which students actually responded to the items. The root mean square error of approximation (RMSEA) is referred to as a badness-of-fit index so that a value closer to 0 implies better fit and a value of 0 implies best fit. In general, RMSEA below 0.05 is considered as good fit and RMSEA over 0.1 suggests poor fit (Browne & Cudeck, 1993). The Tucker-Lewis index (TLI) and the comparative fit index (CFI) are incremental goodness-of-fit indices. These indices compare the implied model to the baseline model where no observed variables are correlated (i.e., there are no factors). Values

greater than 0.9 are recognized as acceptable, and values over 0.95 are considered as good fit (Hu & Bentler, 1999).

Based on the fit indices, the model showed good fit across all content domains. For all tests, RMSEA was below 0.05, and for most tests, CFI and TLI were equal or greater than 0.95.

*Table 34: Goodness-of-Fit Second-Order CFA*

| **Mathematics** | | | | | |
|---|---|---|---|---|---|
| **Grade** | *df* | *RMSEA* | *CFI* | *TLI* | *Convergence* |
| 3 | 1322 | 0.02 | 0.98 | 0.98 | Yes |
| 4 | 1373 | 0.03 | 0.96 | 0.96 | Yes |
| 5 | 1322 | 0.03 | 0.96 | 0.96 | Yes |
| 6 | 1372 | 0.02 | 0.97 | 0.97 | Yes |
| 7 | 1425 | 0.03 | 0.97 | 0.97 | Yes |
| 8 | 1220 | 0.02 | 0.95 | 0.95 | Yes |
| **ELA** | | | | | |
| **Grade** | *df* | *RMSEA* | *CFI* | *TLI* | *Convergence* |
| 3 | 1030 | 0.03 | 0.96 | 0.96 | Yes |
| 4 | 1219 | 0.02 | 0.98 | 0.98 | Yes |
| 5 | 1320 | 0.02 | 0.99 | 0.99 | Yes |
| 6 | 1425 | 0.02 | 0.99 | 0.99 | Yes |
| 7* | 1170 | 0.01 | 0.997 | 0.996 | Yes |
| 8*** | 1264 | 0.02 | 0.99 | 0.99 | Yes |
| 9 | 1372 | 0.02 | 0.99 | 0.99 | Yes |
| 10** | 1425 | 0.01 | 0.99 | 0.99 | Yes |
| **EOC** | | | | | |
| **Subject/Form** | *df* | *RMSEA* | *CFI* | *TLI* | *Convergence* |
| Alg 1/Core 1* | 1221 | 0.02 | 0.99 | 0.99 | Yes |
| Alg 1/Core 2* | 1221 | 0.02 | 0.99 | 0.99 | Yes |
| Alg 1/Core 3** | 1221 | 0.02 | 0.84 | 0.83 | Yes |
| Alg 1/Core 4* | 1221 | 0.02 | 0.99 | 0.99 | Yes |
| Alg 2/Core 1* | 1374 | 0.03 | 0.99 | 0.99 | Yes |
| Alg 2/Core 2** | 1322 | 0.02 | 0.94 | 0.93 | Yes |
| Geo/Core 1*** | 1172 | 0.03 | 0.99 | 0.99 | Yes |
| Geo/Core 2 | 1173 | 0.03 | 0.97 | 0.97 | Yes |

* estimated from R lavaan using WLSMV.
** estimated from MPlus or R lavaan using WLS.
*** estimated based on a first-order model with correlated factors.

The second-order factor model converged for all tests except for grade 8 ELA and Geometry Core 1. Thus, a different model had to be identified for these two tests. For grade 8 ELA and Geometry Core 1, we allowed the residuals of the first-order factors to correlate to each other and removed

the second-order factor from the second-order model. The conceptual definition of the first-order model with correlated factors is similar to the second-order factor model. Both of these models assume that the first-order factors are highly correlated with one another and that there is a common variance among factors. The second-order model is, however, more parsimonious in terms of the number of parameters and more comprehensible when the theoretical framework implies a hierarchical structure. Hence, the implied model for these two tests reflected the same test construct as other FSA tests.

As indicated in Section 3.1, FSA items are operationally calibrated by IRTPRO software; however, factor analyses presented here were conducted with MPlus software and "lavaan" package in R. There are some noted differences between these software packages in terms of their model parameter estimation algorithms and item-specific measurement models. First, IRTPRO employs full information maximum likelihood and chooses model parameter estimates so that the likelihood of data can be maximized, whereas Mplus and lavaan package in R use WLS or WLSMV based on limited information maximum likelihood and choose model parameter estimates so that the likelihood of the observed covariations among items can be maximized. Secondly, IRTPRO allows one to model pseudo-guessing via the 3PL model, whereas Mplus and lavaan package in R do not include the same flexibility. However, CFA results presented here still indicated acceptable fit indices even though pseudo-guessing was constrained to zero or not taken into account.

In Table 35, Table 36, and Table 37, we provide the estimated correlations between the reporting categories from the second-order factor model for Mathematics, ELA and EOC respectively. In all cases, these correlations are very high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

*Table 35: Correlations among Mathematics Factors*

| Grade | Reporting Category | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|
| 3 | Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1) | 1 | | | | |
| | Numbers and Operations – Fractions (Cat2) | 0.93 | 1 | | | |
| | Measurement, Data, and Geometry (Cat3) | 0.98 | 0.95 | 1 | | |
| 4 | Operations and Algebraic Thinking (Cat1) | 1 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 0.95 | 1 | | | |
| | Numbers and Operations – Fractions (Cat3) | 0.90 | 0.89 | 1 | | |
| | Measurement, Data, and Geometry (Cat4) | 0.94 | 0.94 | 0.88 | 1 | |
| 5 | Operations, Algebraic Thinking, and Fractions (Cat1) | 1 | | | | |
| | Numbers and Operations in Base Ten (Cat2) | 0.96 | 1 | | | |

| Grade | Reporting Category | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|
| 5 (cont'd.) | Measurement, Data, and Geometry (Cat3) | 0.92 | 0.94 | 1 | | |
| 6 | Ratio and Proportional Relationships (Cat1) | 1 | | | | |
| | Expressions and Equations (Cat2) | 0.94 | 1 | | | |
| | Geometry (Cat3) | 0.94 | 0.93 | 1 | | |
| | Statistics and Probability (Cat4) | 0.92 | 0.91 | 0.92 | 1 | |
| | The Number System (Cat5) | 0.96 | 0.95 | 0.96 | 0.94 | 1 |
| 7 | Ratio and Proportional Relationships (Cat1) | 1 | | | | |
| | Expressions and Equations (Cat2) | 0.96 | 1 | | | |
| | Geometry (Cat3) | 0.96 | 0.95 | 1 | | |
| | Statistics and Probability (Cat4) | 0.93 | 0.92 | 0.93 | 1 | |
| | The Number System (Cat5) | 0.93 | 0.92 | 0.93 | 0.90 | 1 |
| 8 | Expressions and Equations (Cat1) | 1 | | | | |
| | Functions (Cat2) | 0.94 | 1.00 | | | |
| | Geometry (Cat3) | 0.91 | 0.90 | 1.00 | | |
| | Statistics & Probability and the Number System (Cat4) | 0.92 | 0.91 | 0.88 | 1 | |

*Table 36: Correlations among ELA Factors*

| Grade | Reporting Category | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|
| 3 | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.98 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.96 | 0.97 | 1 | | |
| | Language and Editing Task (Cat4) | 0.81 | 0.82 | 0.81 | 1 | |
| 4 | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.97 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.96 | 0.98 | 1 | | |
| | Language and Editing Task (Cat4) | 0.85 | 0.86 | 0.85 | 1 | |
| | Text-Based Writing (Cat5) | 0.73 | 0.74 | 0.74 | 0.65 | 1 |
| 5 | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.98 | 1 | | | |

| Grade | Reporting Category | Cat1 | Cat2 | Cat3 | Cat4 | Cat5 |
|---|---|---|---|---|---|---|
| 5 (cont'd.) | Integration of Knowledge and Ideas (Cat3) | 0.98 | 0.97 | 1 | | |
| | Language and Editing Task (Cat4) | 0.88 | 0.87 | 0.87 | 1 | |
| | Text-Based Writing (Cat5) | 0.70 | 0.69 | 0.70 | 0.62 | 1 |
| 6 | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.99 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.99 | 0.98 | 1 | | |
| | Language and Editing Task (Cat4) | 0.90 | 0.89 | 0.89 | 1 | |
| | Text-Based Writing (Cat5) | 0.72 | 0.71 | 0.71 | 0.64 | 1 |
| 7* | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.98 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.996 | 0.98 | 1 | | |
| | Language and Editing Task (Cat4) | 0.86 | 0.84 | 0.85 | 1 | |
| | Text-Based Writing (Cat5) | 0.75 | 0.74 | 0.75 | 0.64 | 1 |
| 8*** | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.99 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.96 | 0.97 | 1 | | |
| | Language and Editing Task (Cat4) | 0.87 | 0.92 | 0.88 | 1 | |
| | Text-Based Writing (Cat5) | 0.70 | 0.74 | 0.71 | 0.74 | 1 |
| 9 | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.97 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.97 | 0.96 | 1 | | |
| | Language and Editing Task (Cat4) | 0.88 | 0.87 | 0.88 | 1 | |
| | Text-Based Writing (Cat5) | 0.77 | 0.76 | 0.76 | 0.69 | 1 |
| 10** | Key Ideas and Details (Cat1) | 1 | | | | |
| | Craft and Structure (Cat2) | 0.999 | 1 | | | |
| | Integration of Knowledge and Ideas (Cat3) | 0.998 | 0.999 | 1 | | |
| | Language and Editing Task (Cat4) | 0.90 | 0.90 | 0.90 | 1 | |
| | Text-Based Writing (Cat5) | 0.78 | 0.78 | 0.78 | 0.70 | 1 |

* estimated from R lavaan using WLSMV.
** estimated from MPlus using WLS.
*** estimated based on a first-order model with correlated factors.

*Table 37: Correlations among EOC Factors*

| Course/Form | Reporting Category | Cat1 | Cat2 | Cat3 |
|---|---|---|---|---|
| Algebra 1/Core 1* | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.97 | 1 | |
| | Statistics and the Number System (Cat3) | 0.98 | 0.97 | 1 |
| Algebra 1/Core 2* | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.96 | 1 | |
| | Statistics and the Number System (Cat3) | 0.97 | 0.99 | 1 |
| Algebra 1/Core 3** | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.98 | 1 | |
| | Statistics and the Number System (Cat3) | 0.97 | 0.98 | 1 |
| Algebra 1/Core 4* | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.96 | 1 | |
| | Statistics and the Number System (Cat3) | 0.97 | 0.96 | 1 |
| Algebra 2/Core 1* | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.98 | 1 | |
| | Statistics, Probability, and the Number System (Cat3) | 0.94 | 0.93 | 1 |
| Algebra 2/Core 2** | Algebra and Modeling (Cat1) | 1 | | |
| | Functions and Modeling (Cat2) | 0.98 | 1 | |
| | Statistics, Probability, and the Number System (Cat3) | 0.98 | 0.96 | 1 |
| Geometry/Core1*** | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 1 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 0.94 | 1 | |
| | Modeling with Geometry (Cat3) | 0.98 | 0.98 | 1 |
| Geometry/Core 2 | Congruence, Similarity, Right Triangles and Trigonometry (Cat1) | 1 | | |
| | Circles, Geometric Measurement and Geometric Properties with Equations (Cat2) | 0.93 | 1 | |
| | Modeling with Geometry (Cat3) | 0.97 | 0.96 | 1 |

* estimated from R lavaan using WLSMV.
** estimated from R lavaan using WLS.
*** estimated based on a first-order model with correlated factors.

### 5.2.3 Discussion

In all scenarios, the empirical results suggest the implied model fits the data well. That is, these results indicate that reporting an overall score in addition to separate scores for the individual reporting categories is reasonable, as the intercorrelations among items suggest that there are detectable distinctions among reporting categories.

Clearly, the correlations among the separate factors are high, which is reasonable. This again provides support for the measurement model, given that the calibration of all items is performed concurrently. If the correlations among factors were very low, this could possibly suggest that a different IRT model would be needed (e.g., multidimensional IRT) or that the IRT calibration should be performed separately for items measuring different factors. The high correlations among the factors suggest these alternative methods are unnecessary and that our current approach is in fact preferable.

Overall, these results provide empirical evidence and justification for the use of our scoring and reporting methods. Additionally, the results provide justification for the current IRT model employed.

## 5.3 LOCAL INDEPENDENCE

The validity of the application of Item Response Theory (IRT) depends greatly on meeting the underlying assumptions of the models. One such assumption is local independence, which means that for a given proficiency estimate, the (marginal) likelihood is maximized, assuming the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

$$\mathrm{L}(\theta) = \int \prod_{j=1}^{K} \Pr\left(x_j|\theta\right) f(\theta)\mathrm{d}\theta$$

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that "local independence follows automatically from unidimensionality" (as cited in Bejar, 1980, p. 5). From a dimensionality perspective, there may be nuisance factors that are influencing relationships among certain items, after accounting for the intended construct of interest. These nuisance factors can be influenced by a number of testing features, such as speededness, fatigue, item chaining, and item or response formats (Yen, 1993).

Yen's $Q_3$ statistic (Yen, 1984) was used to measure local independence, which was derived from the correlation between the performances of two items. Simply, the $Q_3$ statistic is the correlation among IRT residuals and is computed using the following equations:

$$d_{ij} = u_{ij} - T_j\left(\hat{\theta}_i\right).$$

where $u_{ij}$ is the item score of the *i*th examinee for item $j$, $T_j\left(\hat{\theta}_i\right)$ is the estimated true score for item *j* of examinee *i*, which is defined as

$$T_j(\hat{\theta}_i) = \sum_{k=1}^{m} y_{jk} P_{jk}(\hat{\theta}_i)$$

where $y_{jk}$ is the weight for response category *k*, *m* is the number of response categories, and $P_{jk}(\hat{\theta}_i)$ is the probability of response category *k* to item *j* by examinee *i* with the ability estimate $\hat{\theta}_i$.

The pairwise index of local dependence $Q_3$ between item *j* and item *j'* is

$$Q_{3jj'} = r(d_j, d_{j'}),$$

where *r* refers to the Pearson product-moment correlation.

When there are *n* items, $n(n\text{-}1)/2$, $Q_3$ statistics will be produced. The $Q_3$ values are expected to be small. Table 38, Table 39, and Table 40 present summaries of the distributions of $Q_3$ statistics—minimum, 5th percentile, median, 95th percentile, and maximum values from each grade and subject. The results show that at least 90% of the items, between the 5th and 95th percentiles, for all grades and subjects were smaller than a critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997).

*Table 38: ELA $Q_3$ Statistic*

| Grade | Average Zero-Order Correlation | Q3 Distribution | | | | | Within Passage $Q_3$ | |
|---|---|---|---|---|---|---|---|---|
| | | Minimum | 5th Percentile | Median | 95th Percentile | Maximum* | Minimum | Maximum |
| 3 | 0.022 | -0.091 | -0.048 | -0.018 | 0.008 | 0.408 | -0.058 | 0.067 |
| 4 | 0.023 | -0.132 | -0.073 | -0.014 | 0.018 | 0.710 | -0.028 | 0.119 |
| 5 | 0.036 | -0.283 | -0.089 | -0.016 | 0.061 | 0.736 | -0.094 | 0.220 |
| 6 | 0.043 | -0.295 | -0.116 | -0.009 | 0.083 | 0.782 | -0.195 | 0.192 |
| 7 | 0.037 | -0.354 | -0.096 | -0.013 | 0.054 | 0.788 | -0.080 | 0.246 |
| 8 | 0.038 | -0.475 | -0.097 | -0.012 | 0.065 | 0.839 | -0.133 | 0.175 |
| 9 | 0.040 | -0.418 | -0.104 | -0.013 | 0.069 | 0.772 | -0.231 | 0.441 |
| 10 | 0.033 | -0.278 | -0.090 | -0.011 | 0.053 | 0.777 | -0.117 | 0.112 |

* Maximum $Q_3$ values of grades 4 through 10 are from elaboration and organization dimensions of the Writing prompt.

*Table 39: Mathematics $Q_3$ Statistic*

| Grade | Average Zero-Order Correlation | $Q_3$ Distribution | | | | |
|---|---|---|---|---|---|---|
| | | Minimum | 5th Percentile | Median | 95th Percentile | Maximum |
| 3 | 0.023 | -0.100 | -0.049 | -0.018 | 0.015 | 0.257 |
| 4 | 0.022 | -0.092 | -0.048 | -0.017 | 0.014 | 0.626 |
| 5 | 0.046 | -0.303 | -0.126 | -0.014 | 0.085 | 0.429 |
| 6 | 0.046 | -0.303 | -0.126 | -0.014 | 0.085 | 0.429 |
| 7 | 0.030 | -0.156 | -0.074 | -0.017 | 0.048 | 0.212 |
| 8 | 0.041 | -0.224 | -0.107 | -0.017 | 0.072 | 0.258 |

*Table 40: EOC $Q_3$ Statistic*

| Course | Average Zero-Order Correlation | $Q_3$ Distribution | | | | |
|---|---|---|---|---|---|---|
| | | Minimum | 5th Percentile | Median | 95th Percentile | Maximum |
| Algebra 1 | 0.042 | -0.410 | -0.107 | -0.016 | 0.078 | 0.485 |
| Algebra 2 | 0.062 | -0.449 | -0.145 | -0.016 | 0.122 | 0.440 |
| Geometry | 0.049 | -0.426 | -0.119 | -0.018 | 0.101 | 0.357 |

# 6. EVIDENCE OF COMPARABILITY

As the FSA was administered in multiple modes (both online and paper-and-pencil), it is important to provide evidence of comparability between the versions. If the content between forms varies, then one cannot justify score comparability.

Student scores should not depend on the mode of administration or the type of test form. FSA had online assessments for grades 5 through 10 ELA, grades 5 through 8 Mathematics, and EOC. To improve the accessibility of the statewide assessment, alternate assessments were provided to students whose Individual Educational Plans (IEP) or Section 504 Plans indicated such a need. Thus, the comparability of scores obtained via alternate means of administration must be established and evaluated. For grades 3 and 4 ELA and Mathematics, there were no accommodated forms, as the tests were universally administered on paper for these grades. For Algebra 1, Algebra 2, and Geometry, only the Core 1 form was administered as an accommodated version.

## 6.1 MATCH-WITH-TEST BLUEPRINTS FOR BOTH PAPER-AND-PENCIL AND ONLINE TESTS

For the 2014–2015 FSA, the paper-and-pencil versions of the tests were developed according to the same test specifications used for the online tests. These paper tests matched the same blueprints designed for the online tests. In this section, evidence of matching blueprints for both online and paper tests is provided. The procedures used to establish comparable forms are provided in Volume 2, Test Development, of the 2015 FSA Technical Report.

## 6.2 COMPARABILITY OF FSA TEST SCORESOVER TIME

This section is not applicable this year.

## 6.3 COMPARABILITY OF ONLINE AND PAPER-AND-PENCIL TEST SCORES

Test forms for paper-and-pencil administration were offered as a special accommodation for students who qualified, according to their Individual Educational Plans (IEP) or Section 504 Plans. These forms aligned to the same test specifications as the online forms and used the same item parameters for scoring for items that were common in both forms. However, without an online system, technology-enhanced items could not be administered with paper-pencil testing. Thus, some items were replaced with comparable items formatted for paper. This was the only difference between the two versions.

After replacing technology-enhanced items with multiple-choice items, accommodated forms were somewhat different from online forms. This pattern can be easily found in TCCs for Mathematics, in which a relatively large number of items were replaced in the accommodated forms. However, this is not concerning since all of the items are on the same IRT scale. In Mathematics, TCCs for the accommodated forms are above those of the online forms, slightly shifted upward compared to the TCCs for the online forms. Conversely, there is an overlap of TCCs for the online and accommodated forms in ELA. As seen in Table 41, only a few items were replaced in ELA. TCCs for both ELA and Mathematics are presented in Appendix C.

Common items between the online and accommodated forms were used to link the forms. By linking item parameters across the modes of testing administrations, we placed item parameters from the paper-and-pencil forms to the same scale of the online forms, which in turn supports comparability.

*Table 41: Number of Item Replacements for the Accommodated Forms*

| Mathematics | Number of Items Replaced | ELA | Number of Items Replaced |
|---|---|---|---|
| Grade 5 | 9 | Grade 5 | 3 |
| Grade 6 | 17 | Grade 6 | 3 |
| Grade 7 | 20 | Grade 7 | 2 |
| Grade 8 | 18 | Grade 8 | 8 |
| Algebra 1* | 22 | Grade 9 | 9 |
| Algebra 2* | 21 | Grade 10 | 4 |
| Geometry* | 26 | Grade 10 Retake | 4 |

* EOC accommodated forms were created by replacing items in Core 1 forms.

# 7. FAIRNESS AND ACCESSIBILITY

## 7.1 FAIRNESS IN CONTENT

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population;
2. Precisely defined constructs;
3. Accessible, non-biased items;
4. Amenable to accommodations;
5. Simple, clear, and intuitive instructions and procedures;
6. Maximum readability and comprehensibility; and
7. Maximum legibility.

Test development specialists have received extensive training on the principles of universal design and apply these principles in the development of all test materials. In the review process, adherence to the principles of universal design is verified by Florida educators and stakeholders.

## 7.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is not sufficient to determine the fairness of a test. Rather, it must be accompanied by statistical processes. While a variety of item statistics were reviewed during form building to evaluate the quality of items, one notable statistic that was utilized was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF, according to the DIF classification convention illustrated in Volume 1. Furthermore, items were categorized positively (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American/black, Hispanic, or female), or negatively (i.e., –A, –B, or–C), signifying that the item favored the reference group (e.g., white or male). Items were flagged if their DIF statistics indicated the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal or the reference group. The details surrounding this review of items for bias is further described in Volume 2, Test Development.

DIF analyses were conducted for all items to detect potential item bias from a statistical perspective across major ethnic and gender groups. DIF analyses were performed for the following groups:

- Male/Female
- White/African-American

- White/Hispanic
- Not student with disability (SWD)/SWD
- Not English language learner (ELL)/ELL

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 5.2, of the 2014–2015 FSA Annual Technical Report. The DIF statistics for each test item are presented in the appendices of Volume 1 of the 2014–2015 FSA Annual Technical Report.

**Summary**

This report is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. The overall results can be summarized as follows:

- Reliability: Various measures of reliability are provided at the aggregate and subgroup levels, showing the reliability of all tests is in line with acceptable industry standards.
- Content validity: Evidence is provided to support the assertion that content coverage on each form was consistent with test specifications of the blueprint across testing modes.
- Internal structural validity: Evidence is provided to support the selection of the measurement model, the tenability of local independence, and the reporting of an overall score and subscores at the reporting category levels.

# 8. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

American Psychological Association. (2014). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87*, 513–524.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.

Chen, W. -H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. 16*, 297–334.

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

Feldt, L. S., & Qualls, A. L. (1996). Bias in Coefficient Alpha Arising from Heterogeneity of Test Content. *Applied Measurement in Education, 9*, 277–286.

Hu, L. T. and Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381–389.

Kuder, G. F., & Richardson, N. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151–160.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L. K. and Muthén, B. O. (1998–2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*, 111–120.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, 42, 549–565.

Rosseel Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2),* 1-36. URL http://www.jstatsoft.org/v48/i02/.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th Ed.). New Jersey: Lawrence Erlbaum.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retreived October 2002, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling Performance Assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 20*, 187-213.

Yoon, B., & Young, M. J. (2000). *Estimating the reliability for test scores with mixed item formats: Internal consistency and generalizability.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.