Florida
Standards Assessments

Florida Standards Assessments

2015–2016

Volume 4
Evidence of Reliability and Validity

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

ACKNOWLEDGMENTS

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, Patrick Kozak, Mayumi Rezwan, and Kathryn Conway. The major contributors from the Florida Department of Education are as follows: Dr. Salih Binici, Dr. Qian Liu, Vince Verges, Victoria Ash, Susie Lee, Mengyao Cui, Steve Ash, Sally Rhodes, 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 (ELA) 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 ELA score. In the spring of 2016 the grade 4 Reading portion of the ELA assessment transitioned to an online delivery.

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 Grade 3 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
ELA Reading	Online	4–10
	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 2015–2016 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 and classification accuracy results by grade and subject.
- 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-and-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 $\sigma_X\sqrt{1-\rho_{XX\prime}}$, where σ_X is the standard deviation of the scaled score and $\rho_{XX\prime}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived.

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

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

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

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

In general, the standard error of measurement is relatively constant across samples as the group dependent term, σ_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 in the classical test theory is assumed to be homoscedastic error irrespective of the standard deviation of a group.

In contrast, the standard errors of measurement in IRT 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 (see Section 4.4). 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 (see Volume 7 of 2014-2015 FSA technical reports).

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 National Assessment of Educational Progress (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 outlined in *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 2016, student-level scores included scale scores and raw scores at the reporting category level. Based on the performance cuts approved by the State Board of Education on January 6, 2016, scale scores and achievement levels were reported in spring 2016. 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 σ_i^2 is the variance of scores on each item, σ_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 α_i is the reliability of the *i*th strata, σ_i^2 is the variance between items in the *i*th strata, and σ_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 σ_x^2 is the total score variance, (i.e., the variance of the whole test); σ_i^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

$$\hat{\lambda}_i = \frac{(\sigma_{i1} + \sigma_{i2} + \sigma_i^2 + \sigma_{i(i+1)} + \cdots + \sigma_{ik})}{\sigma_x^2}.$$

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 large numbers 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	33	28	27	28	22	32	34	31
MS		2	2	4	6	4	3	3
ET								
ETC	8	7	8	8	10	8	8	12
HT	3	8	8	8	8	6	4	4
GI								
EBSR	1	3	4	4	6	2	5	3
NL	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	44	43	9	17	13	20	23; 20; 21	16; 18	17; 21
MS5***	3		11	3	2	4	2; 0; 0	2;1	2;1
MS6***		1		2	2		2; 0; 0	0; 2	3;1
GI			6	5	6	10	6; 9; 8	8; 5	9; 8
SA	2	2							
HT									2;3
TI			1	1	1	1	1;1; 0		1;1
MI			1	2			2;1;1	2; 1	
NL								0;1	1; 0
EQ	5	8	26	25	32	20	22; 27; 28	30; 30	23; 23

* Descriptions for each item type are presented in Table 3
** Algebra 1 has three core forms, and Algebra 2 and Geometry have two core forms
*** MS5 and MS6 refer to the number of multi-select options

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 0.87 to 0.93 for ELA, 0.90 to 0.95 for Mathematics, and 0.83 to 0.94 for EOC. The stratified *alpha* coefficients ranged from 0.87 to 0.93 for ELA, 0.90 to 0.95 for Mathematics, and 0.83 to 0.94 for EOC. The Feldt-Raju coefficients were between 0.85 and 0.91 for ELA, 0.87 and 0.93 for Mathematics, and 0.88 and 0.92 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	0.90	0.90	0.88
4	Online	0.90	0.90	0.88
	Accommodated	0.87	0.87	0.85
5	Online	0.91	0.91	0.89
	Accommodated	0.88	0.88	0.85
6	Online	0.93	0.93	0.91
	Accommodated	0.90	0.90	0.88
7	Online	0.91	0.91	0.90
	Accommodated	0.90	0.90	0.88
8	Online	0.91	0.91	0.90
	Accommodated	0.90	0.90	0.87
9	Online	0.91	0.91	0.90
	Accommodated	0.91	0.91	0.89
10	Online	0.91	0.91	0.88
	Accommodated	0.90	0.90	0.86

Table 7: Reliability Coefficients (Mathematics)

Grade	Form	Cronbach *Alpha*	Stratified *Alpha*	Feldt-Raju
3	Paper	0.93	0.93	0.92
4	Paper	0.93	0.93	0.92
5	Online	0.95	0.95	0.93
	Accommodated	0.94	0.94	0.90
6	Online	0.95	0.95	0.93
	Accommodated	0.92	0.92	0.89
7	Online	0.95	0.95	0.93
	Accommodated	0.93	0.93	0.90
8	Online	0.92	0.92	0.90
	Accommodated	0.90	0.90	0.87

Table 8: Reliability Coefficients (EOC)

Course	Form	Cronbach *Alpha*	Stratified *Alpha*	Feldt-Raju
Algebra 1	Online – Core 1	0.92	0.92	0.92
	Online – Core 2	0.93	0.93	0.92
	Online – Core 3	0.93	0.93	0.92
	Accommodated	0.83	0.83	0.88
Algebra 2	Online – Core 1	0.94	0.94	0.91
	Online – Core 2	0.93	0.93	0.91
	Accommodated	0.89	0.89	0.88
Geometry	Online – Core 1	0.94	0.94	0.92
	Online – Core 2	0.94	0.94	0.91
	Accommodated	0.89	0.89	0.90

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 using operational items.

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. The inverse of the test information function represents the standard error of measurement. The standard error of measurement is equal to the inverse square root of information. 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 function generating the standard error of measurement and $f(\hat{\theta})$ is the assumed population density. Table 9 presents the marginal reliability coefficients for all students. The marginal reliability coefficients for all subjects and grades were higher than 0.9, ranging from 0.90 to 0.94.

Table 9: Marginal Reliability Coefficients

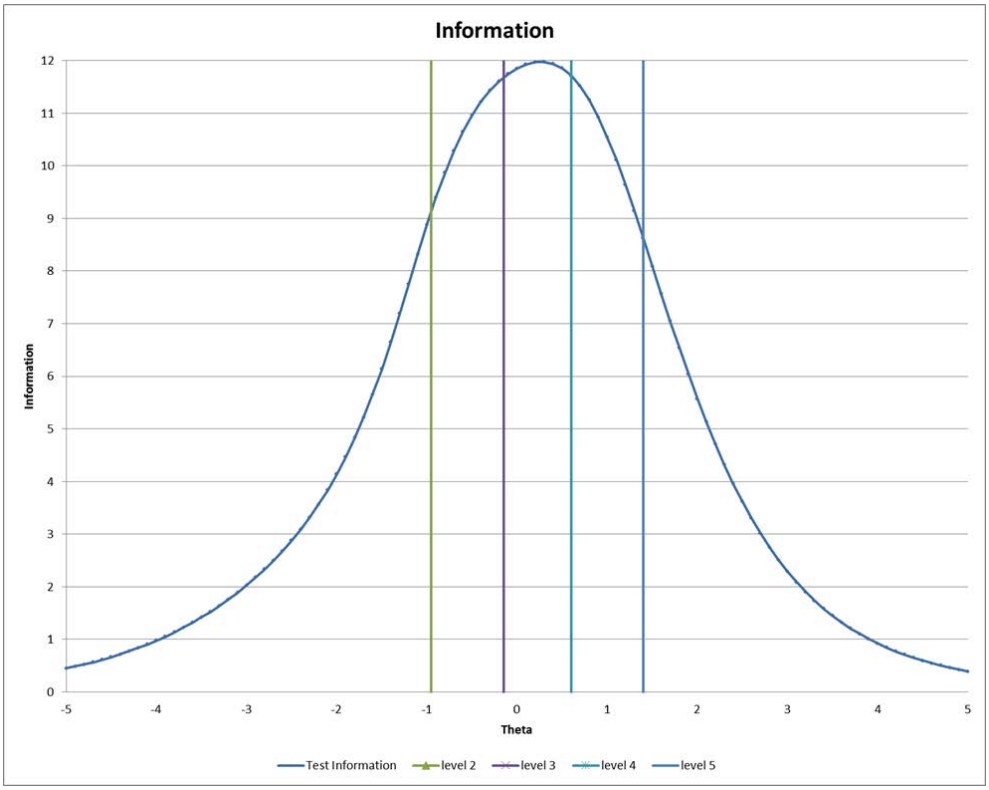
Subject	Grade	Marginal Reliability for Response Pattern Scores	Subject	Grade/Course	Marginal Reliability for Response Pattern Scores
ELA	3	0.91	Mathematics	3	0.92
	4	0.91		4	0.93
	5	0.91		5	0.94
	6	0.93		6	0.94
	7	0.92		7	0.94
	8	0.92		8	0.92
	9	0.92	EOC	Algebra 1	0.92
	10	0.92		Algebra 2	0.90
				Geometry	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 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:

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

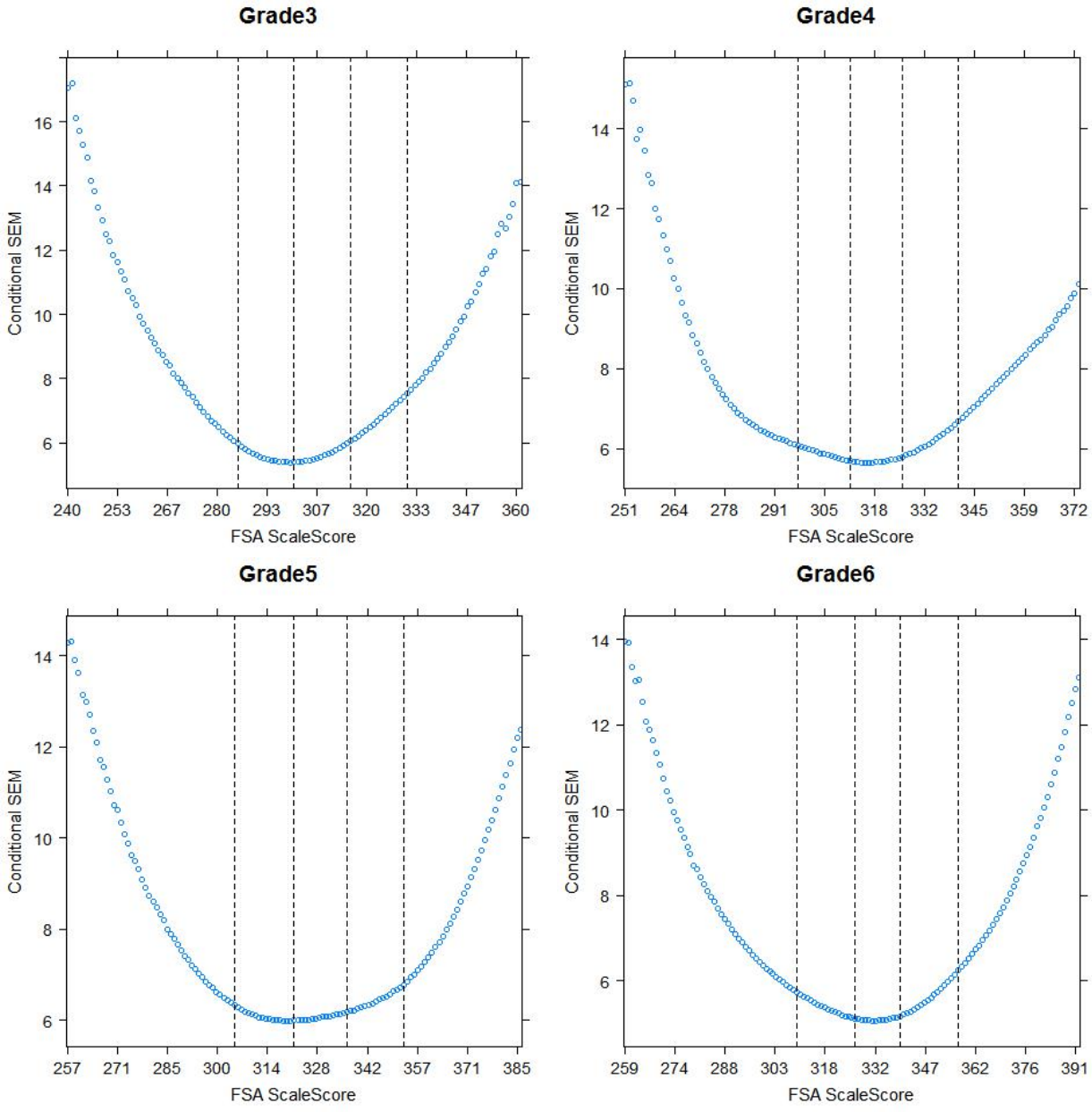
where N_{GPCM} is the number of items that are scored using generalized partial credit model (GPCM) items, N_{3PL} is the number of items scored using 3PL or 2PL model, i indicates item i ($i \in \{1,2,\ldots,N\}$), m_i is the maximum possible score of the item, s indicates student s, and θ_s is the 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 scaled scores reported in 2016. Vertical lines represent the four performance category cut scores.

Figure 2: Conditional Standard Errors of Measurement (ELA)



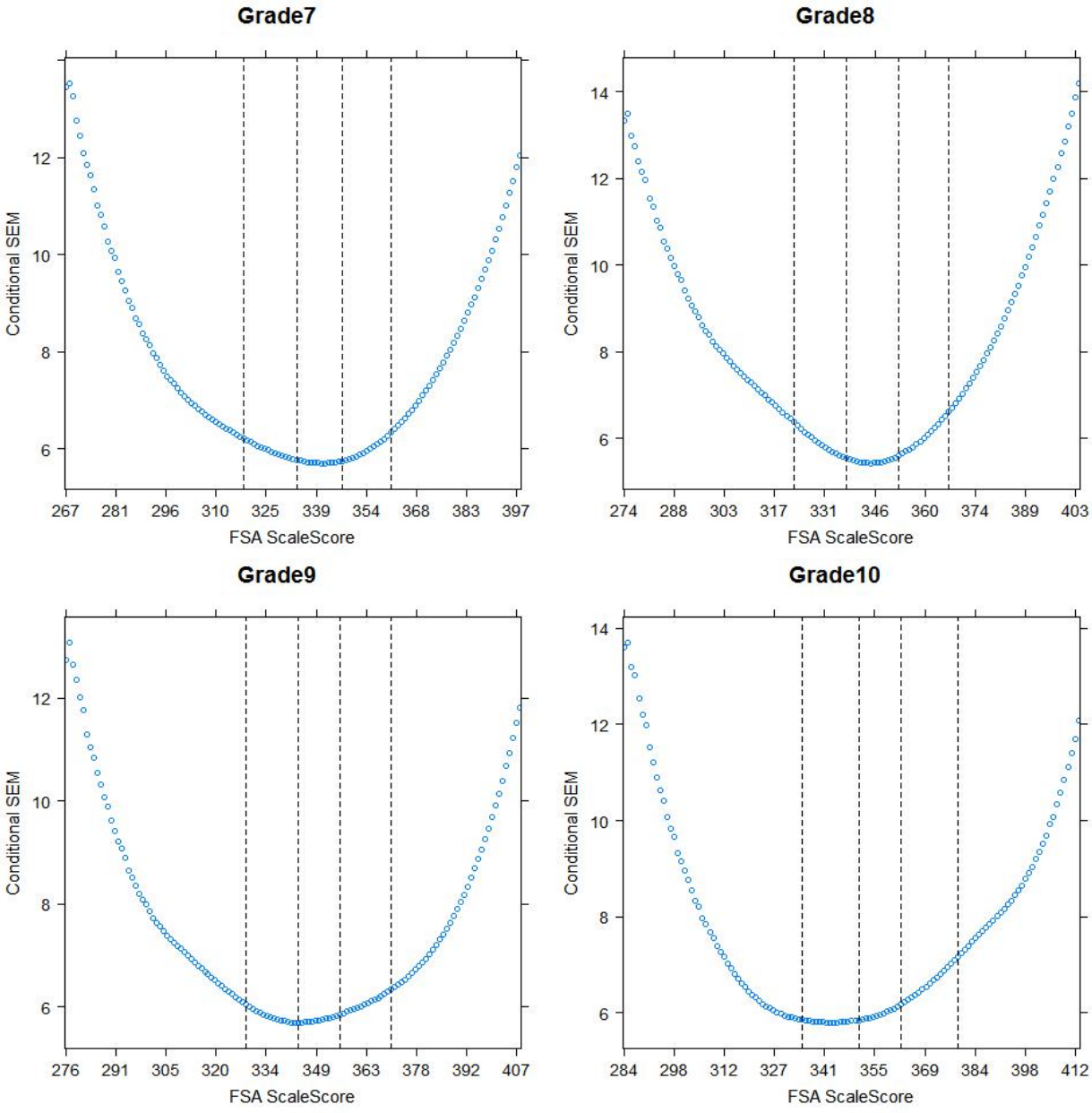
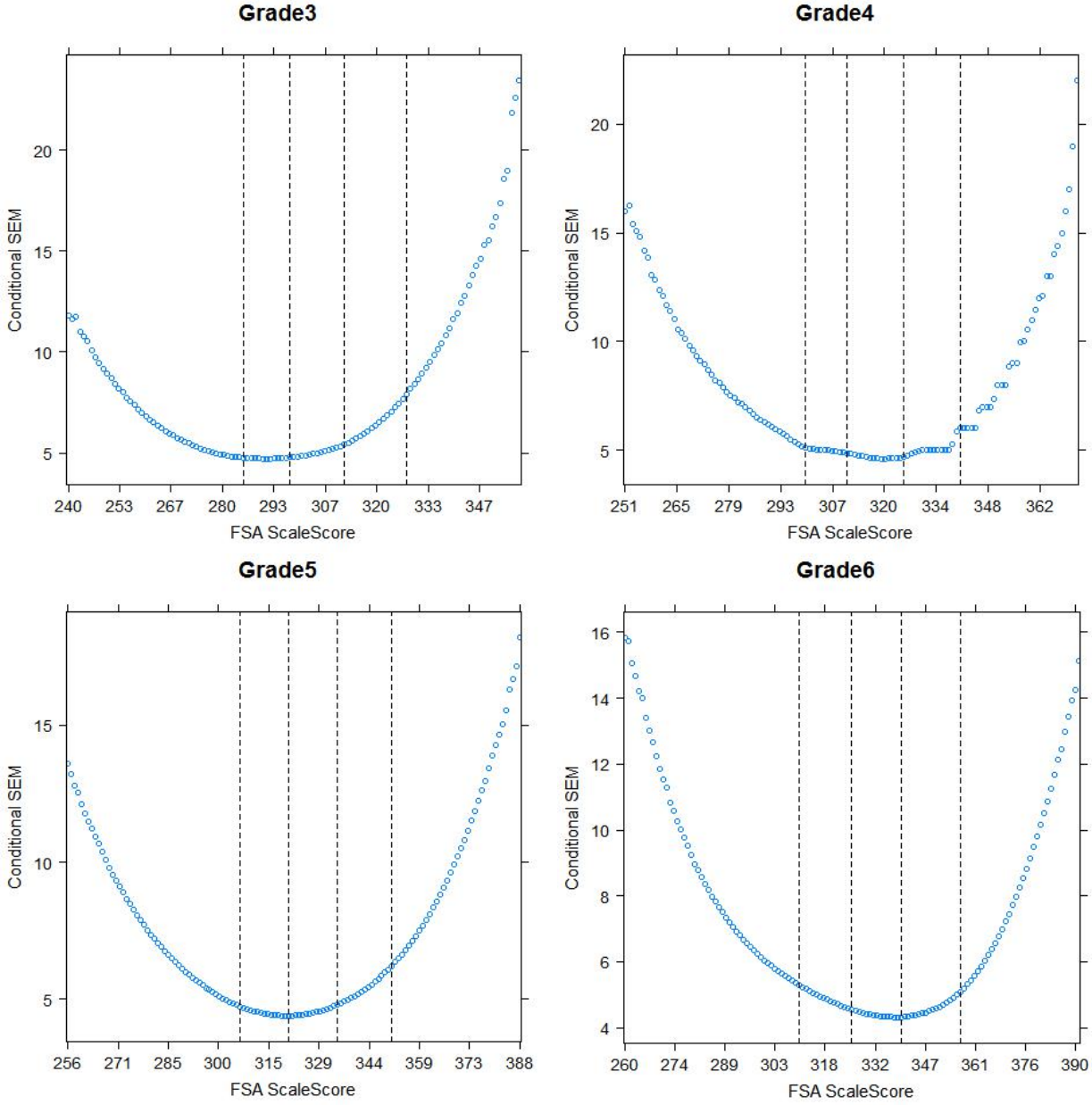

Grade7
Conditional SEM
FSA ScaleScore
267 281 296 310 325 339 354 368 383 397
6 8 10 12
Grade8
Conditional SEM
FSA ScaleScore
274 288 303 317 331 346 360 374 389 403
6 8 10 12 14
Grade9
Conditional SEM
FSA ScaleScore
276 291 305 320 334 349 363 378 392 407
6 8 10 12
Grade10
Conditional SEM
FSA ScaleScore
284 298 312 327 341 355 369 384 398 412
6 8 10 12 14

Figure 3: Conditional Standard Errors of Measurement (Mathematics)



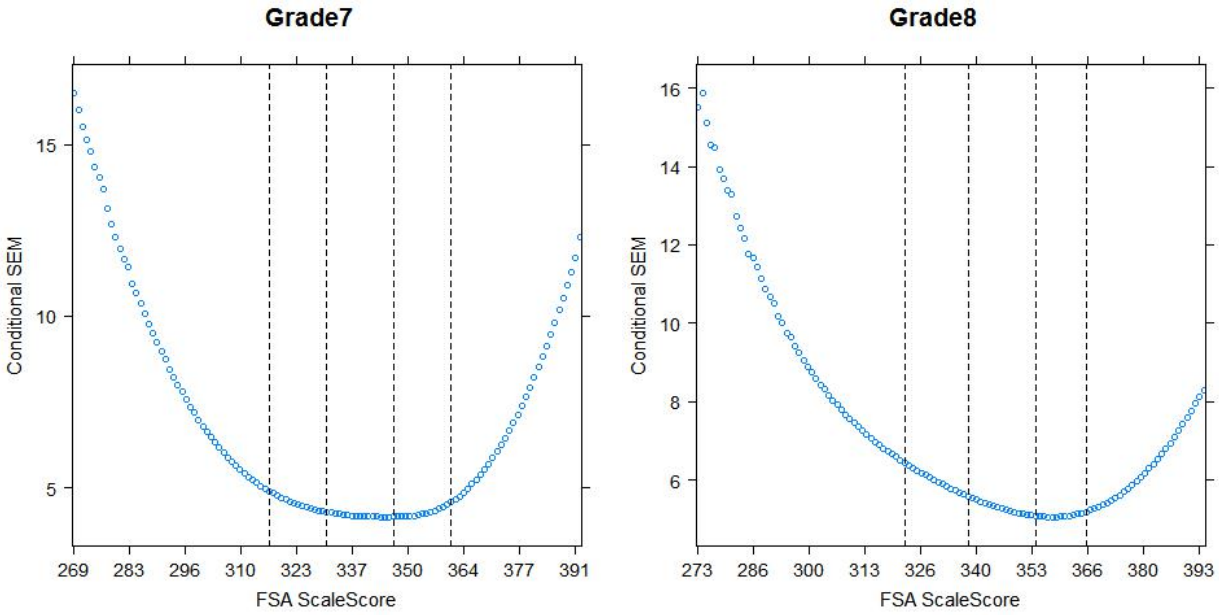
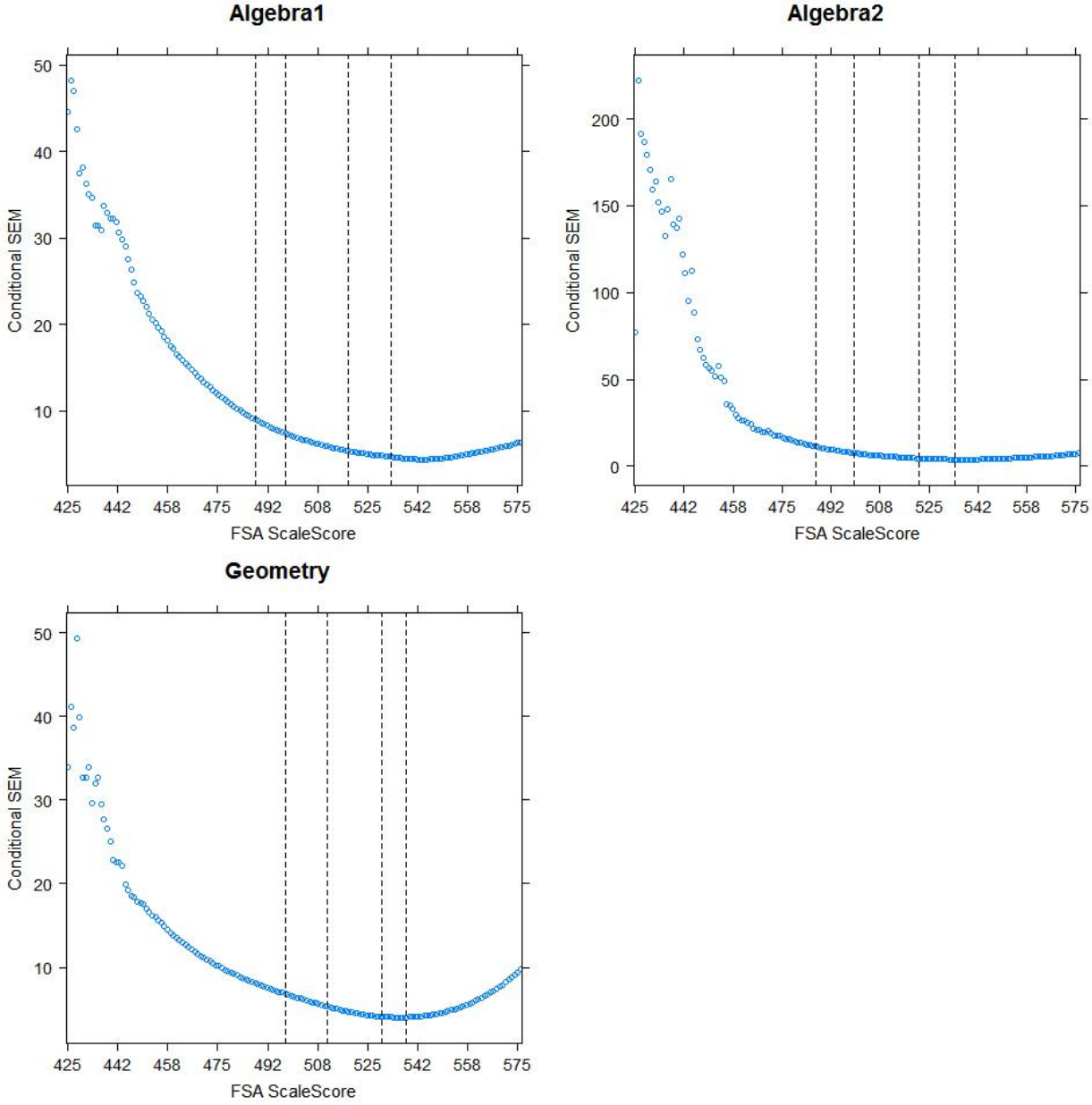

Grade7
Conditional SEM
15
10
5
269
283
296
310
323
337
350
364
377
391
FSA ScaleScore
Grade8
Conditional SEM
16
14
12
10
8
6
273
286
300
313
326
340
353
366
380
393
FSA ScaleScore

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. Because the testing of Algebra 1 and Algebra 2 is still relatively 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.

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

When students complete the FSA, they are placed into one of five achievement levels given their observed scaled score. The cut scores for student classification into the different achievement levels were determined after the FSA standard-setting process.

During test construction, techniques are implemented to minimize misclassification of students, which can occur on any assessment. In particular, standard error of measurement (SEM) curves can be constructed to ensure that smaller SEMs are expected near important cut scores of the test.

Misclassification probabilities are computed for all achievement level standards (i.e., for the cuts between levels 1 and 2, levels 2 and 3, levels 3 and 4, and levels 4 and 5). The achievement level cut between level 2 and level 3 is of primary interest because students are classified as Satisfactory or Below Satisfactory using this cut. Students with observed scores far from the level 3 cut are expected to be classified more accurately as Satisfactory or Below Satisfactory than students with scores near this cut. This report estimates classification reliabilities using two different methods: one based on observed abilities and a second based on estimating a latent posterior distribution for the true scores.

Two approaches for estimating classification probabilities are provided. The first is an observed score approach to computing misclassification probabilities and is designed to explore the following research questions:

1. What is the overall classification accuracy index (CAI) of the total test?
2. What is the classification accuracy rate index for each individual performance cut within the test?

The second approach computes misclassification probabilities using an IRT-based method for students scoring at each score point. This approach is designed to explore the following research questions:

1. What is the probability that the student's true score is below the cut point? And;
2. What is the probability that the student's true score is above the cut point?

Both approaches yield student-specific classification probabilities that can be aggregated to form overall misclassification rates for the test.

3.4.1 Classification Accuracy Estimation Methods

In the first approach, we used students from the spring 2016 FSA population data files with the status of reported scores. However, in the second approach, item level data from the calibration sample were used. Since there were multiple core forms in EOC tests, the classification accuracy analysis was performed for each form, as operational items varied by form. Also, the item level data used in IRT-based approach did not include accommodated tests because the sample was too small to compute classification accuracy.

Table 10 provides the sample size, mean, and standard deviation of the observed theta for the data used in the first method described above. The theta scores are based on the maximum likelihood estimates (MLEs) obtained from AIR's scoring engine. Similarly, Table 11 provides the sample size, mean, and standard deviation of the observed theta for the data used in the second method.

Table 10: Descriptive Statistics from Population Data

ELA Grade	Sample Size	Average Theta	Standard Deviation of Theta	Mathematics Grade/EOC Subject	Sample Size	Average Theta	Standard Deviation of Theta
3	220786	0.05	1.08	3	220898	0.06	1.08
4	209339	-0.07	1.07	4	212258	0.01	1.11
5	200723	-0.03	1.07	5	202798	0.02	1.09
6	197353	0.03	1.03	6	194400	-0.08	1.11
7	194423	-0.06	1.06	7	176304	-0.06	1.15
8	196609	0.01	1.11	8	135324	0.06	1.21
9	201784	-0.02	1.07	Algebra 1	219884	-0.21	1.23
10	196165	-0.01	1.05	Algebra 2	137035	-0.02	1.27
				Geometry	202784	-0.12	1.15

Table 11: Descriptive Statistics from Calibration Data

ELA				Mathematics				EOC			
Grade	N	Average Theta	SD of Theta	Grade	N	Average Theta	SD of Theta	Subject/Core	N	Average Theta	SD of Theta
3	28455	0.10	1.08	3	29729	0.09	1.08	Alg1/Core5	109616	-0.22	1.25
4	27029	-0.05	1.08	4	28833	0.03	1.12	Alg1/Core6	53766	-0.17	1.19
5	25971	-0.03	1.08	5	24740	0.03	1.05	Alg1/Core7	53669	-0.20	1.23
6	25176	0.06	1.06	6	25905	-0.10	1.11	Alg2/Core3	85343	-0.03	1.28
7	24068	-0.04	1.07	7	23363	-0.07	1.13	Alg2/Core4	51136	0.00	1.26
8	22890	0.03	1.08	8	93452	0.01	1.16	Geo/Core3	126215	-0.13	1.15
9	24005	0.03	1.05					Geo/Core4	75040	-0.10	1.15
10	23996	0.03	1.06								

The observed score approach (Rudner, 2001) implemented to assess classification accuracy is based on the probability that the true score, θ, for student i is within performance level $j = 1,2,\cdots,J$. This probability can be estimated from evaluating the following integral

$$p_{ij} = \Pr(\lambda_l \le \theta_i < \lambda_u | \hat{\theta}_i, \hat{\sigma}_i^2) = \int_{\lambda_l}^{\lambda_u} f\left(\theta_i \middle| \hat{\theta}_i, \hat{\sigma}_i^2\right) d\theta_i,$$

where λ_u and λ_l denote the score corresponding to the upper and lower limits of the performance level, respectively, $\hat{\theta}_i$ is the ability estimate of the *i*th student with standard error of measurement of $\hat{\sigma}_i$ and using the asymptotic property of normality of the maximum likelihood estimate, $\hat{\theta}_i$, we take $f(\cdot)$ as asymmetrically normal, so the above probability can be estimated by

$$p_{ij} = \Phi\left(\frac{\lambda_u - \hat{\theta}_i}{\hat{\sigma}_i}\right) - \Phi\left(\frac{\lambda_l - \hat{\theta}_i}{\hat{\sigma}_i}\right),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (CDF).
The expected number of students at level *j* based on students from observed level *k* can be expressed as

$$E_{kj} = \sum_{pl_i \in k} p_{ij},$$

where pl_i is the *i*th student's performance level, the values of E_{kj} are the elements used to populate the matrix $\boldsymbol{E}$, a 5 ×5 matrix of conditionally expected numbers of students to score within each performance level bin based on their true scores. The overall CAI of the test can then be estimated from the diagonal elements of the matrix:

$$\text{CAI} = \frac{tr(\boldsymbol{E})}{N},$$

where $N = \sum_{k=1}^{5} N_k$, N_k is the observed number of students scoring in performance level k. The classification accuracy index for the individual cuts (CAIC) is estimated by forming square partitioned blocks of the matrix $\boldsymbol{E}$ and taking the summation over all elements within the block as follows:

$$\text{CAIC} = \left(\sum_{k=1}^{p}\sum_{j=1}^{p} E_{kj} + \sum_{k=p+1}^{5}\sum_{j=p+1}^{5} E_{kj}\right) \Big/ N,$$

where p is the element of one of the cuts of interest.

The IRT-based approach (Guo, 2006) makes use of student-level item response data from the 2016 FSA test administration. We can estimate a posterior probability distribution for the latent true score and from this estimate the probability that a true score is above the cut as

$$p(\theta > c) = \frac{\int_c^{\infty} p(z|\theta) f(\theta|\mu,\sigma) d\theta}{\int_{-\infty}^{\infty} p(z|\theta) f(\theta|\mu,\sigma)\, d\theta},$$

where c is the cut score required for passing in the same assigned metric, θ is true ability in the true-score metric, z is the item score, μ is the mean, and σ is the standard deviation of the population distribution. The function $p(z|\theta)$ is the probability of the particular pattern of responses given the theta, and $f(\theta)$ is the density of the proficiency θ in the population.

Similarly we can estimate the probability that a true score is below the cut as

$$p(\theta < c) = \frac{\int_{-\infty}^{c} p(z|\theta) f(\theta|\mu, \sigma) d\theta}{\int_{-\infty}^{\infty} p(z|\theta) f(\theta|\mu, \sigma)\, d\theta}.$$

From these misclassification probabilities, we can estimate the overall false positive rate (FPR) and false negative rate (FNR) of the test. The FPR is expressed as the proportion of individuals who scored above the cut based on their observed score, but their true score would otherwise have classified them as below the cut. The FNR is expressed as the proportion of individuals who scored below the cut based on their observed score, but otherwise would have been classified as above the cut based on their true scores. These rates are estimated as follows:

$$\text{FPR} = \sum_{i \in \theta < c} p(\theta > c)/N$$

$$\text{FNR} = \sum_{i \in \theta \geq c} p(\theta < c)/N.$$

In addition to these rates, we computed the accuracy rates for each cut as

$$\text{Accuracy} = 1 - (\text{FPR} + \text{FNR}).$$

3.4.2 Results

Table 12 and Table 13 provide the overall classification accuracy index (CAI) and the classification accuracy index for the individual cuts (CAIC) for the ELA and Mathematics tests, respectively, based on the observed score approach. Here the overall classification accuracy of the test ranges from 0.747 to around 0.810 for Mathematics and EOC, and from 0.733 to 0.767 for ELA.

The overall cut accuracy rates are much higher, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is typically above or close to 0.9.

Table 12: Classification Accuracy Index (ELA)

Grade	Overall Accuracy Index	Cut Accuracy Index			
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4	Cut 4 and Cut 5
3	0.736	0.938	0.920	0.923	0.954
4	0.735	0.933	0.917	0.923	0.961
5	0.740	0.937	0.916	0.925	0.961
6	0.767	0.941	0.929	0.933	0.963
7	0.741	0.931	0.920	0.931	0.957
8	0.746	0.942	0.926	0.928	0.950
9	0.741	0.939	0.920	0.924	0.955
10	0.733	0.936	0.914	0.922	0.958

Table 13: Classification Accuracy Index (Mathematics and EOC)

Grade/Subject	Overall Accuracy Index	Cut Accuracy Index			
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4	Cut 4 and Cut 5
3	0.747	0.951	0.928	0.918	0.948
4	0.774	0.942	0.936	0.938	0.957
5	0.788	0.951	0.938	0.940	0.959
6	0.801	0.943	0.936	0.950	0.972
7	0.810	0.948	0.941	0.951	0.970
8	0.769	0.929	0.923	0.946	0.969
Algebra 1	0.757	0.901	0.910	0.950	0.973
Algebra 2	0.760	0.883	0.913	0.957	0.970
Geometry	0.779	0.915	0.922	0.959	0.976

Table 14 and Table 15 provide the FPR and FNR from the IRT-based approach for both ELA and Mathematics tests. In Mathematics, the FPR and FNR rates for the level 2/3 cut are around 6% to 7%, with the exception of the EOC tests (shown in Table 16), which have slightly larger rates given the challenging nature of those tests. In ELA, the rates are around 8%. Table 14 and Table 15 also provide the overall accuracy rates after accounting for both false positive and false negative rates. For example, the overall accuracy rate of 0.851 for the level 2/3 cut in grade 3 Mathematics suggests 85.1% of the students estimated to have a true score status at level 3 are correctly classified into that category by their observed scores. As expected, the overall accuracy rates are reasonable in all cuts except at the extreme cuts. A high false negative rate at the cut between levels 4 and 5 is also expected due to large standard error.

Table 14: False Classification Rates and Overall Accuracy Rates (ELA)

	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Grade	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
3	0.156	0.036	0.808	0.082	0.079	0.839	0.045	0.178	0.777	0.018	0.368	0.614
4	0.145	0.041	0.814	0.081	0.085	0.834	0.047	0.175	0.778	0.016	0.326	0.658
5	0.159	0.036	0.805	0.083	0.087	0.830	0.043	0.176	0.781	0.015	0.322	0.663
6	0.143	0.033	0.824	0.070	0.070	0.860	0.040	0.129	0.831	0.016	0.266	0.718
7	0.141	0.044	0.815	0.075	0.087	0.838	0.040	0.148	0.812	0.020	0.260	0.720
8	0.152	0.035	0.813	0.083	0.073	0.845	0.045	0.142	0.813	0.022	0.272	0.707
9	0.128	0.038	0.833	0.081	0.083	0.836	0.048	0.151	0.800	0.020	0.278	0.701
10	0.139	0.038	0.823	0.080	0.089	0.831	0.050	0.164	0.787	0.018	0.311	0.671

Table 15: False Classification Rates and Overall Accuracy Rates (Mathematics)

	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Grade	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
3	0.124	0.027	0.849	0.086	0.062	0.851	0.054	0.143	0.803	0.022	0.341	0.636
4	0.124	0.036	0.841	0.076	0.054	0.870	0.042	0.107	0.852	0.019	0.227	0.754
5	0.121	0.030	0.849	0.073	0.058	0.869	0.040	0.108	0.852	0.018	0.228	0.754
6	0.119	0.037	0.845	0.062	0.067	0.871	0.030	0.101	0.868	0.012	0.203	0.785
7	0.110	0.033	0.857	0.061	0.061	0.878	0.029	0.121	0.850	0.011	0.225	0.764
8	0.140	0.049	0.811	0.069	0.095	0.835	0.027	0.162	0.811	0.011	0.234	0.755

Table 16: False Classification Rates and Overall Accuracy Rates (EOC)

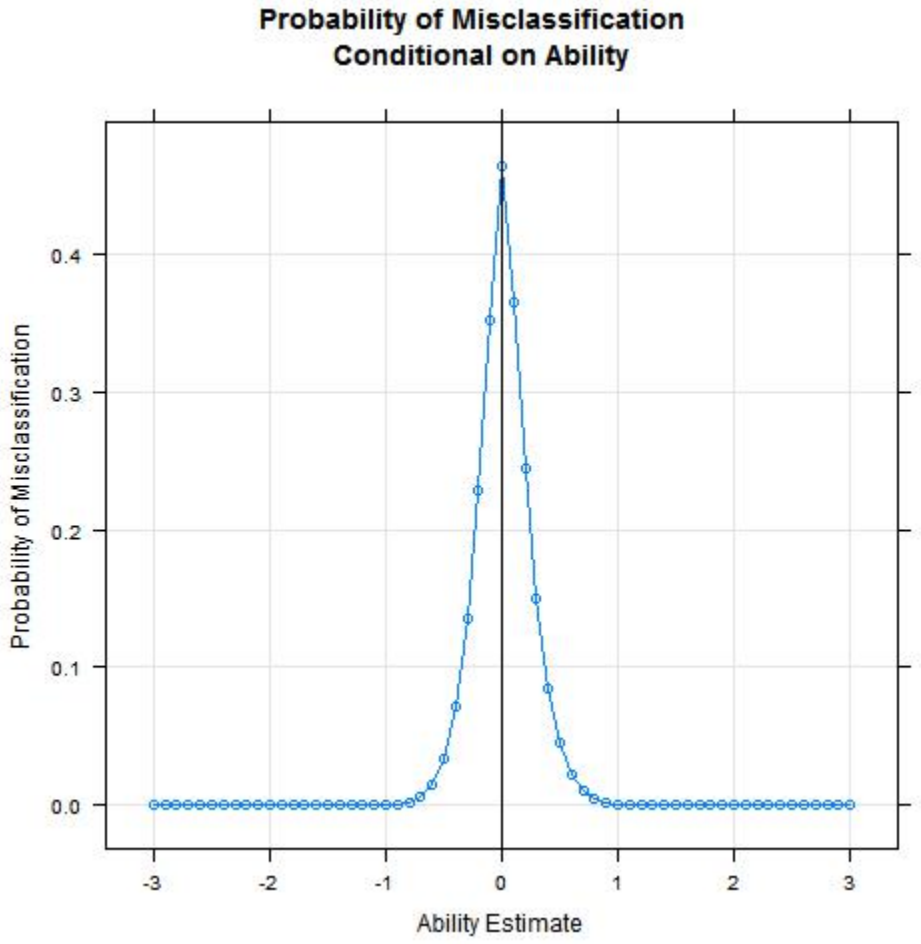
	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Grade	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
Algebra 1 Core5	0.146	0.067	0.787	0.083	0.083	0.834	0.026	0.118	0.856	0.011	0.148	0.841
Algebra 1 Core6	0.148	0.063	0.788	0.083	0.080	0.837	0.027	0.111	0.863	0.012	0.139	0.849
Algebra 1 Core7	0.150	0.066	0.784	0.086	0.082	0.831	0.026	0.113	0.862	0.011	0.135	0.854
Algebra 2 Core3	0.112	0.080	0.808	0.048	0.092	0.860	0.016	0.117	0.866	0.010	0.138	0.853
Algebra 2 Core4	0.115	0.081	0.804	0.050	0.095	0.855	0.016	0.120	0.863	0.010	0.139	0.850
Geometry Core3	0.143	0.056	0.801	0.074	0.077	0.848	0.021	0.109	0.870	0.011	0.137	0.852
Geometry Core4	0.148	0.058	0.793	0.078	0.078	0.844	0.021	0.104	0.875	0.011	0.131	0.858

Figure 5 shows a plot exhibiting the probability of misclassification for grade 3 ELA. The plot displays that students with scores below -0.346 on the theta scale which corresponds to a scale score of 293, and students with scores above 0.373 corresponding to a scale score of 307 are classified accurately at least 90% of the time. Scale scores representing 90% of classification accuracy by each grade and subject are displayed in Appendix C.

Appendix C also includes plots of the misclassification probabilities for the level 2/3 cuts from the IRT-based approach conditional on ability for all grades and subject as well as by subgroups (ELLs and SWDs). The vertical bar within each graph represents the cut score required to achieve level 3 (i.e., Satisfactory). A properly functioning test yields increased misclassification probabilities approaching the cut, as the density of the posterior probability distribution is symmetric, and approximately half of its mass will fall on either side of the proficiency level cut as $\theta \rightarrow c$.

These visual displays are useful heuristics to evaluate the probability of misclassification for all levels of ability. Students far from the level 3 cut have very small misclassification probabilities, and the probabilities approach a peak near 50% as $\theta \rightarrow c$, as expected.

Figure 5: Probability of Misclassification Conditional on Ability



These results demonstrate that classification reliabilities are generally high, with some lower rates affecting tests known to be particularly challenging. The classification accuracy results presented

in this report (Table 12 and Table 13) are generally equivalent to or higher than those reported in the 2013 FCAT 2.0 and EOC technical reports. Based on the Florida Statewide Assessments 2013 Yearbook (Florida Department of Education, 2013), the classification accuracy rates in Mathematics ranged from 0.690 in grade 4 to 0.719 in grade 5 (see page 112 for details). Similarly, the classification accuracy rates in Reading ranged from 0.664 in grade 10 to 0.718 in grade 3 (see page 264 for details). The classification accuracy rates in Algebra 1 vary from 0.716 to 0.737 (see page 413 for details). Additionally, we can compare the FSA classification accuracy rates to those of the State of New York, which is comparable in population size (New York State Education Department, 2014). Although New York administers a different testing program, estimated accuracy rates here range from 77% to 81% in ELA and from 81% to 85% in Mathematics (see page 100 for details). While the overall cut accuracy results for New York are slightly higher than those of the FSA, as there are only three achievement level cuts compared to four FSA cuts, the individual cut accuracy was comparable between New York and Florida. Florida showed from 92% to 93% in ELA and from 92% to 96% in Mathematics for the level 2/3 cut. New York showed from 91% to 93% in ELA and from 93% to 95% in Mathematics for the proficiency cut.

3.5 Precision at Cut Scores

Table 17 through Table 19 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 17: Achievement Levels and Associated Conditional Standard Error of Measurement (ELA)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	8.17		
3	2	5.53	285	6
3	3	5.64	300	5
3	4	6.60	315	6
3	5	8.97	330	8
4	1	7.62		
4	2	5.95	297	7
4	3	5.89	311	6
4	4	6.08	325	6
4	5	7.43	340	8
5	1	8.04		
5	2	6.06	304	7
5	3	6.06	321	7
5	4	6.35	336	7
5	5	7.66	352	8
6	1	7.14		

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
6	2	5.30	309	6
6	3	5.00	326	5
6	4	5.47	339	6
6	5	7.33	356	7
7	1	7.39		
7	2	6.02	318	7
7	3	5.91	333	6
7	4	5.99	346	6
7	5	7.22	360	7
8	1	8.03		
8	2	5.93	322	7
8	3	5.44	337	6
8	4	6.01	352	6
8	5	7.98	366	7
9	1	7.31		
9	2	5.94	328	7
9	3	5.92	343	6
9	4	6.01	355	6
9	5	7.00	370	7
10	1	7.07		
10	2	5.98	334	7
10	3	6.00	350	7
10	4	6.52	362	7
10	5	7.92	378	8

Table 18: Achievement Levels and Associated Conditional Standard Error of Measurement (Mathematics)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	5.97		
3	2	4.95	285	5
3	3	5.01	297	5
3	4	6.28	311	5
3	5	12.06	327	8
4	1	7.49		

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
4	2	5.00	299	5
4	3	4.68	310	5
4	4	4.98	325	5
4	5	9.02	340	6
5	1	6.45		
5	2	4.46	306	5
5	3	4.50	320	4
5	4	5.30	334	5
5	5	8.93	350	6
6	1	7.44		
6	2	4.98	310	5
6	3	4.21	325	5
6	4	4.44	339	4
6	5	7.04	356	5
7	1	7.68		
7	2	4.53	316	5
7	3	4.00	330	4
7	4	4.04	346	4
7	5	5.93	360	5
8	1	8.65		
8	2	6.04	322	6
8	3	5.17	337	6
8	4	5.00	353	5
8	5	5.77	365	5

Table 19: Achievement Levels and Associated Conditional Standard Error of Measurement (EOC)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Algebra_1	1	19.02		
Algebra_1	2	8.23	487	9
Algebra_1	3	6.35	497	7
Algebra_1	4	5.05	518	6
Algebra_1	5	4.60	532	5
Algebra_2	1	29.28		

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Algebra_2	2	6.72	497	8
Algebra_2	3	4.71	511	5
Algebra_2	4	4.00	529	4
Algebra_2	5	4.39	537	4
Geometry	1	14.38		
Geometry	2	7.47	486	8
Geometry	3	5.58	499	7
Geometry	4	4.09	521	5
Geometry	5	4.44	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 20, 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 20: 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 equal 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 21 displays rater-agreement percentages. The percentage of exact agreement between two raters ranged from 63% to 81%. The percentage of non-adjacent rating was between 19% and 36%. The non-adjacent percentages fell between 0% and 2%. 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 21: Inter-Rater Reliability

Grade	Item ID	Dimension	% Exact	% Adjacent	% Not Adjacent	Total Number of Processed Responses
4	23271	Purpose, Focus, & Organization	73	26	0	425,774
		Evidence & Elaboration	72	28	1	425,774
		Conventions	66	33	1	425,774
5	23328	Purpose, Focus, & Organization	63	36	1	406,959
		Evidence & Elaboration	63	35	2	406,959
		Conventions	71	29	0	406,959
6	23333	Purpose, Focus, & Organization	65	34	1	403,374
		Evidence & Elaboration	65	33	1	403,374
		Conventions	70	29	1	403,374
7	23273	Purpose, Focus, & Organization	65	33	2	398,792
		Evidence & Elaboration	65	33	1	398,792
		Conventions	74	25	1	398,792
8	23383	Purpose, Focus, & Organization	64	34	2	402,170
		Evidence & Elaboration	63	35	2	402,170
		Conventions	80	20	0	402,170
9	23385	Purpose, Focus, & Organization	73	26	1	418,151
		Evidence & Elaboration	71	28	1	418,151
		Conventions	81	19	0	418,151
10	23413	Purpose, Focus, & Organization	68	31	1	408,356
		Evidence & Elaboration	68	31	1	408,356
		Conventions	78	22	0	408,356

In addition to inter-rater reliability, validity coefficients, percent exact agreement on validity true scores and human scores, were also calculated. Validity true scores for each dimension were

determined by scoring directors, and TDC content experts approved those scores. 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 22 presents final validity coefficients, which were between 74 and 92.

Table 22: Validity Coefficients

Grade	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
4	82	81	74
5	79	79	82
6	78	80	78
7	80	81	84
8	91	90	92
9	90	89	92
10	87	85	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 *ij*th cell, p_{cij} is the proportion in the *ij*th 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 23. They ranged from 0.6 to 0.94.

Table 23: Weighted Kappa Coefficients

Grade	Scorer	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
4	Two Human Raters	0.71	0.68	0.62
5	Two Human Raters	0.66	0.64	0.62
6	Two Human Raters	0.66	0.64	0.60
7	Two Human Raters	0.72	0.72	0.67
8	Machine and Human	0.92	0.94	0.89
9	Machine and Human	0.91	0.92	0.90
10	Two Human Raters	0.75	0.73	0.70

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.

3.6.1 Automated Scoring Engine

AIR's essay scoring engine, Autoscore, uses a statistical process to evaluate Writing prompts. Autoscore evaluates papers against the same rubric used by human raters, but a statistical process is used to analyze each paper and assign scores for each of the three dimensions. The engine uses the same process for scoring essays every time a new prompt is submitted, regardless of whether the data is obtained from an operational assessment or an independent field test.

Statistical rubrics are effectively proxy measures. Although they can directly measure some aspects of Writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not directly measure argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in Writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Furthermore, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may predict whether a response to a factual item is correct— higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the actual reason that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

AIR's essay-scoring engine uses a statistical rubric with great success, as measured by the rater agreements observed relative to the human-to-human rater agreements. This technology is similar to all essay-scoring systems in the field. Although some systems replace the statistical process with a "neural network" algorithm, that algorithm functions like the statistical model. Not all descriptions of essay-scoring algorithms are as transparent as AIR's, but whenever a training set is used for the machine to "learn a rubric," the same technology is being used.

The engine is designed to employ a "training set," a set of essays scored with maximally valid scores that are used to form the basis of the prediction model. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Moreover, an ideal training sample over-represents higher- and lower-scoring papers and is selected according to a scientific sampling design with known probabilities of selection.

The training process of the scoring engine has two phases. The first phase requires oversampled, high- and low-scoring papers, leaving an equally weighted representative sample for the second phase. The first phase is used to identify concepts that are proportionately represented in higher-scoring papers. Here, concepts are defined as words and their synonyms, as well as clusters of words used meaningfully in proximity.

The second phase takes a series of measures on each essay in the remaining training set. These measures include latent semantic analysis (LSA) measures based on the concepts identified in the first phase; other semantic measures indicate the coherence of concepts within and across paragraphs and a range of word-use and syntactic measures. The LSA is similar to a data reduction method identifying common concepts within the narrative and reducing the data to a configurable number of LSA dimensions.

For each trait in the rubric, the system estimates an appropriate statistical model where these LSA and other syntactic characteristics described above serve as the independent variables, and the final, resolved score serves as the dependent variable in an ordered probit regression. This model, along with its final parameter estimates, is used to generate a predicted or "proxy" score. The probability of scoring in the *p*th category is compared to a random draw from the uniform distribution, and a final score point of 1 through 4 is determined from this comparison.

In addition to the training set, an independent random sample of responses is drawn for the cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are handscored, and the LSA and other syntactic characteristics of the papers are computed. Subsequently, a second machine score is generated by applying the model coefficients obtained from the ordered probit in the training set. This forms a predicted score for the papers in the cross-validation set for each dimension in the rubric, which can then be used to evaluate the agreement rates between the human and Autoscore engine.

When implementing the scoring engine, we expect that the computer-to-human agreement rates to be at least as high as the human-to-human agreement rates obtained from the double-scored process. If the engine yields scores with rater agreement rates that are at least as high as the human rater agreement rates, then the scoring engine can be deployed for operational scoring. If the computer-to-human agreement rates are not at least as high as the human-to-human rates, then

adjustments to the scoring engine statistical model are necessary in order to find a scoring model that yields rater agreement rates that match the human-to-human rates.

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 1,000 papers, and the remaining records were used for validation. In addition, due to the small number of records for validation in grade 8, 900 for training and 370 for validation were also implemented. The total number of LSA dimensions and the sample size for validation are presented in Table 24. Table 24 also shows that the scoring engine produced comparable results with human scores.

Table 24: Percent Agreement in Handscoring and Scoring Engine

Grade	Dimension	Handscoring from DRC			AIR Scoring Engine				
		% Exact	% Adjacent	% Not Adjacent	LSA	% Exact	% Adjacent	% Not Adjacent	N for Validation
8	Purpose, Focus, & Organization	73.56	25.70	0.74	50	74.05	25.14	0.81	370
	Evidence & Elaboration	76.37	22.96	0.67	200	72.59	27.04	0.37	270
	Conventions	77.63	21.85	0.52	50	81.89	17.84	0.27	370
9	Purpose, Focus, & Organization	77.25	22.75	0.00	50	78.02	21.98	0.00	464
	Evidence & Elaboration	80.00	20.00	0.00	100	79.96	20.04	0.00	464
	Conventions	77.45	22.55	0.00	100	78.66	20.91	0.43	464

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 25, Table 26, and Table 27 present the reporting categories by grade and test, as well as the number of items measuring each category. Table 28, Table 29, and Table 30 present the number of items by each reporting category for the accommodated forms.

Table 25: Number of Items for Each ELA Reporting Category

Reporting Category	Grade							
	3	4	5	6	7	8	9	10
Key Ideas and Details	10	12	14	12	15	14	17	14
Craft and Structure	17	17	15	22	17	18	18	16
Integration of Knowledge and Ideas	11	12	12	10	10	12	11	11
Language and Editing Task	8	7	8	8	10	8	8	12
Text-Based Writing		1	1	1	1	1	1	1

* Reporting categories and the number of items belonging to each reporting category are identical for both online and accommodated forms except for grade 5 (see Table 28).

Table 26: 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	9
	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	18
6	Ratio and Proportional Relationships	8
	Expressions and Equations	17
	Geometry	8
	Statistics and Probability	11
	The Number System	11
7	Ratio and Proportional Relationships	14
	Expressions and Equations	12
	Geometry	13
	Statistics and Probability	9
	The Number System	8
8	Expressions and Equations	17
	Functions	13
	Geometry	15
	Statistics & Probability and the Number System	10

Table 27: Number of Items for Each EOC Reporting Category

Course	Reporting Category	Core Form				
		3	4	5	6	7
Algebra 1	Algebra and Modeling			24	24	24
	Functions and Modeling			23	23	23
	Statistics and the Number System			11	11	11
Algebra 2	Algebra and Modeling	21	21			
	Functions and Modeling	21	21			
	Statistics, Probability, and the Number System	16	16			
Geometry	Congruence, Similarity, Right Triangles and Trigonometry	27	27			
	Circles, Geometric Measurement and Geometric Properties with Equations	22	22			
	Modeling with Geometry	9	9			

Table 28: Number of Items for Each ELA Accommodated Reporting Category

Grade	Reporting Category	Number of Items
5	Key Ideas and Details	13
	Craft and Structure	16
	Integration of Knowledge and Ideas	12
	Language and Editing Task	8
	Text-Based Writing	1

Table 29: Number of Items for Each Mathematics Accommodated Reporting Category

Grade	Reporting Category	Number of Items
5	Operations, Algebraic Thinking, and Fractions	21
	Numbers and Operations in Base Ten	15
	Measurement, Data, and Geometry	18
6	Ratio and Proportional Relationships	8
	Expressions and Equations	17
	Geometry	8
	Statistics and Probability	11
	The Number System	11

Grade	Reporting Category	Number of Items
7	Ratio and Proportional Relationships	14
	Expressions and Equations	12
	Geometry	13
	Statistics and Probability	9
	The Number System	8
8	Expressions and Equations	17
	Functions	14
	Geometry	15
	Statistics & Probability and the Number System	10

Table 30: Number of Items for Each EOC Accommodated Reporting Category

Course	Reporting Category	Number of Items
Algebra 1	Algebra and Modeling	24
	Functions and Modeling	23
	Statistics and the Number System	11
Algebra 2	Algebra and Modeling	21
	Functions and Modeling	21
	Statistics, Probability, and the Number System	16
Geometry	Congruence, Similarity, Right Triangles and Trigonometry	27
	Circles, Geometric Measurement and Geometric Properties with Equations	22
	Modeling with Geometry	9

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 2016 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 2015, when all parties met face-to-face to select items 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 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 2016, 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 Mathematics and grade 3 Reading. The following tests and grades included field test items:

- Grades 4 through 10 in ELA;
- Grades 5 through 8 in Mathematics; and
- End-of-Course Assessments (Algebra 1, Algebra 2, Geometry).

Field testing was conducted during the spring as 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 reflected 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 was completed. The study found that items were fully aligned with the intended content and that items in FSA test forms demonstrated a good representation of the standards—the Language Arts Florida Standards (LAFS) and the Mathematics Florida Standards (MAFS). A full report on alignment is provided in Appendix D.

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 25, Table 26, and Table 27 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 data well 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 31 through Table 33 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 0.42 to 0.8. The Language and Editing Task items and Text-Based Writing items exhibited slightly lower correlations with the other reporting categories ranging from 0.42 to 0.64. For Mathematics, the correlations were between 0.65 and 0.83. For EOC, the correlations between the three subscales fell between 0.75 and 0.83. Observed correlations from the accommodated forms are presented in Table 34 through Table 36. The correlations varied between 0.32 and 0.78 for ELA, 0.46 and 0.79 for Mathematics, and 0.49 and 0.76 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 37 through Table 42 display disattenuated correlations. Disattenuated values greater than 1.00 are reported as 1.00*. In ELA, the Writing dimension had the lowest correlations among the five reporting categories. For the Writing dimension, the average value was 0.71 and the minimum was 0.66, whereas the overall average disattenuated correlation for ELA was 0.90.

Table 31: Observed 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)	17	0.75	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.64	0.67	1.00		
	Language and Editing Task (Cat4)	8	0.59	0.64	0.52	1.00	
4	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	17	0.72	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.69	0.70	1.00		
	Language and Editing Task (Cat4)	7	0.48	0.49	0.47	1.00	
	Text-Based Writing (Cat5)	1	0.54	0.55	0.52	0.42	1.00
5	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	15	0.72	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.71	0.67	1.00		
	Language and Editing Task (Cat4)	8	0.60	0.59	0.55	1.00	
	Text-Based Writing (Cat5)	1	0.55	0.53	0.50	0.52	1.00
6	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	22	0.80	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.67	0.71	1.00		
	Language and Editing Task (Cat4)	8	0.56	0.60	0.50	1.00	
	Text-Based Writing (Cat5)	1	0.58	0.61	0.53	0.51	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
7	Key Ideas and Details (Cat1)	15	1.00				
	Craft and Structure (Cat2)	17	0.77	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.67	0.69	1.00		
	Language and Editing Task (Cat4)	10	0.55	0.56	0.48	1.00	
	Text-Based Writing (Cat5)	1	0.54	0.54	0.48	0.46	1.00
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.76	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.70	0.72	1.00		
	Language and Editing Task (Cat4)	8	0.52	0.55	0.52	1.00	
	Text-Based Writing (Cat5)	1	0.56	0.59	0.55	0.48	1.00
9	Key Ideas and Details (Cat1)	17	1.00				
	Craft and Structure (Cat2)	18	0.76	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.70	0.71	1.00		
	Language and Editing Task (Cat4)	8	0.55	0.57	0.52	1.00	
	Text-Based Writing (Cat5)	1	0.58	0.59	0.53	0.47	1.00
10	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	16	0.73	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.66	0.62	1.00		
	Language and Editing Task (Cat4)	12	0.56	0.55	0.52	1.00	
	Text-Based Writing (Cat5)	1	0.58	0.57	0.53	0.54	1.00

Table 32: Observed 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)	9	0.73	1.00			
	Measurement, Data, and Geometry (Cat3)	19	0.76	0.68	1.00		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Operations and Algebraic Thinking (Cat1)	11	1.00				
	Numbers and Operations in Base Ten (Cat2)	11	0.73	1.00			
	Numbers and Operations – Fractions (Cat3)	14	0.78	0.75	1.00		
	Measurement, Data, and Geometry (Cat4)	18	0.74	0.72	0.77	1.00	
5	Operations, Algebraic Thinking, and Fractions (Cat1)	21	1.00				
	Numbers and Operations in Base Ten (Cat2)	15	0.83	1.00			
	Measurement, Data, and Geometry (Cat3)	18	0.83	0.79	1.00		
6	Ratio and Proportional Relationships (Cat1)	8	1.00				
	Expressions and Equations (Cat2)	17	0.77	1.00			
	Geometry (Cat3)	8	0.70	0.77	1.00		
	Statistics and Probability (Cat4)	11	0.68	0.75	0.70	1.00	
	The Number System (Cat5)	11	0.69	0.77	0.68	0.65	1.00
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.82	1.00			
	Geometry (Cat3)	13	0.79	0.75	1.00		
	Statistics and Probability (Cat4)	9	0.69	0.68	0.66	1.00	
	The Number System (Cat5)	8	0.75	0.76	0.70	0.66	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	13	0.68	1.00			
	Geometry (Cat3)	15	0.72	0.66	1.00		
	Statistics & Probability and the Number System (Cat4)	10	0.73	0.66	0.71	1.00	

Table 33: Observed Correlation Matrix among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 1/Core 5	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.80	1.00	
	Statistics and the Number System (Cat3)	11	0.78	0.75	1.00

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 1/Core 6	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.83	1.00	
	Statistics and the Number System (Cat3)	11	0.77	0.76	1.00
Algebra 1/Core 7	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.82	1.00	
	Statistics and the Number System (Cat3)	11	0.76	0.75	1.00
Algebra 2/Core 3	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.82	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.79	0.76	1.00
Algebra 2/Core 4	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.81	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.80	0.76	1.00
Geometry/Core 3	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	0.83	1.00	
	Modeling with Geometry (Cat3)	9	0.76	0.75	1.00
Geometry/Core 4	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	0.83	1.00	
	Modeling with Geometry (Cat3)	9	0.78	0.79	1.00

Table 34: Observed Correlation Matrix among Reporting Categories (ELA Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	17	0.65	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.61	0.61	1.00		
	Language and Editing Task (Cat4)	7	0.40	0.44	0.36	1.00	
	Text-Based Writing (Cat5)	1	0.46	0.49	0.46	0.32	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
5	Key Ideas and Details (Cat1)	13	1.00				
	Craft and Structure (Cat2)	16	0.67	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.61	0.60	1.00		
	Language and Editing Task (Cat4)	8	0.48	0.58	0.43	1.00	
	Text-Based Writing (Cat5)	1	0.49	0.52	0.43	0.46	1.00
6	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	22	0.75	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.60	0.63	1.00		
	Language and Editing Task (Cat4)	8	0.54	0.58	0.47	1.00	
	Text-Based Writing (Cat5)	1	0.52	0.55	0.48	0.47	1.00
7	Key Ideas and Details (Cat1)	15	1.00				
	Craft and Structure (Cat2)	17	0.75	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.67	0.68	1.00		
	Language and Editing Task (Cat4)	10	0.57	0.60	0.50	1.00	
	Text-Based Writing (Cat5)	1	0.48	0.52	0.40	0.44	1.00
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.69	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.68	0.69	1.00		
	Language and Editing Task (Cat4)	8	0.47	0.57	0.52	1.00	
	Text-Based Writing (Cat5)	1	0.48	0.52	0.48	0.51	1.00
9	Key Ideas and Details (Cat1)	17	1.00				
	Craft and Structure (Cat2)	18	0.78	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.70	0.68	1.00		
	Language and Editing Task (Cat4)	8	0.52	0.53	0.50	1.00	
	Text-Based Writing (Cat5)	1	0.55	0.55	0.49	0.44	1.00
10	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	16	0.71	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.63	0.56	1.00		
	Language and Editing Task (Cat4)	12	0.54	0.51	0.46	1.00	
	Text-Based Writing (Cat5)	1	0.51	0.53	0.41	0.47	1.00

Table 35: Observed 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	0.79	1.00			
	Measurement, Data, and Geometry (Cat3)	18	0.78	0.79	1.00		
6	Ratio and Proportional Relationships (Cat1)	8	1.00				
	Expressions and Equations (Cat2)	17	0.70	1.00			
	Geometry (Cat3)	8	0.61	0.66	1.00		
	Statistics and Probability (Cat4)	11	0.54	0.54	0.46	1.00	
	The Number System (Cat5)	11	0.70	0.77	0.58	0.49	1.00
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.75	1.00			
	Geometry (Cat3)	13	0.69	0.71	1.00		
	Statistics and Probability (Cat4)	9	0.59	0.64	0.57	1.00	
	The Number System (Cat5)	8	0.70	0.75	0.65	0.60	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	14	0.62	1.00			
	Geometry (Cat3)	15	0.67	0.59	1.00		
	Statistics & Probability and the Number System (Cat4)	10	0.68	0.55	0.64	1.00	

Table 36: Observed Correlation Matrix among Reporting Categories (EOC Accommodated Forms)

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 1	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.66	1.00	
	Statistics and the Number System (Cat3)	11	0.58	0.49	1.00
Algebra 2	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.68	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.72	0.67	1.00

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Geometry	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	0.76	1.00	
	Modeling with Geometry (Cat3)	9	0.64	0.59	1.00

Table 37: 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)	17	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.99	0.98	1.00		
	Language and Editing Task (Cat4)	8	0.86	0.88	0.83	1.00	
4	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	17	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.99	1.00	1.00		
	Language and Editing Task (Cat4)	7	0.97	0.99	0.97	1.00	
	Text-Based Writing (Cat5)	1	0.71	0.72	0.70	0.79	1.00
5	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	15	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	1.00*	0.99	1.00		
	Language and Editing Task (Cat4)	8	0.89	0.90	0.85	1.00	
	Text-Based Writing (Cat5)	1	0.71	0.71	0.67	0.73	1.00
6	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	22	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.99	0.99	1.00		
	Language and Editing Task (Cat4)	8	0.89	0.91	0.89	1.00	
	Text-Based Writing (Cat5)	1	0.73	0.73	0.74	0.78	1.00
7	Key Ideas and Details (Cat1)	15	1.00				
	Craft and Structure (Cat2)	17	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	1.00*	1.00*	1.00		
	Language and Editing Task (Cat4)	10	0.84	0.86	0.83	1.00	
	Text-Based Writing (Cat5)	1	0.67	0.66	0.66	0.66	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	1.00*	0.98	1.00		
	Language and Editing Task (Cat4)	8	0.84	0.86	0.87	1.00	
	Text-Based Writing (Cat5)	1	0.71	0.72	0.72	0.72	1.00
9	Key Ideas and Details (Cat1)	17	1.00				
	Craft and Structure (Cat2)	18	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	1.00*	1.00*	1.00		
	Language and Editing Task (Cat4)	8	0.85	0.86	0.87	1.00	
	Text-Based Writing (Cat5)	1	0.73	0.73	0.72	0.68	1.00
10	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	16	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.97	0.93	1.00		
	Language and Editing Task (Cat4)	12	0.82	0.82	0.82	1.00	
	Text-Based Writing (Cat5)	1	0.73	0.74	0.71	0.73	1.00

Table 38: 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)	9	0.90	1.00			
	Measurement, Data, and Geometry (Cat3)	19	0.94	0.91	1.00		
4	Operations and Algebraic Thinking (Cat1)	11	1.00				
	Numbers and Operations in Base Ten (Cat2)	11	0.97	1.00			
	Numbers and Operations – Fractions (Cat3)	14	0.98	0.98	1.00		
	Measurement, Data, and Geometry (Cat4)	18	0.93	0.94	0.95	1.00	

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)	18	0.96	0.93	1.00		
6	Ratio and Proportional Relationships (Cat1)	8	1.00				
	Expressions and Equations (Cat2)	17	0.98	1.00			
	Geometry (Cat3)	8	0.94	0.96	1.00		
	Statistics and Probability (Cat4)	11	0.93	0.94	0.93	1.00	
	The Number System (Cat5)	11	0.90	0.93	0.87	0.84	1.00
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.98	1.00			
	Geometry (Cat3)	13	0.95	0.91	1.00		
	Statistics and Probability (Cat4)	9	0.93	0.93	0.89	1.00	
	The Number System (Cat5)	8	0.95	0.97	0.89	0.94	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	13	0.92	1.00			
	Geometry (Cat3)	15	0.90	0.90	1.00		
	Statistics & Probability and the Number System (Cat4)	10	0.96	0.94	0.94	1.00	

Table 39: Disattenuated Correlation Matrix among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 1/Core 5	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.98	1.00	
	Statistics and the Number System (Cat3)	11	1.00*	1.00*	1.00
Algebra 1/Core 6	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.99	1.00	
	Statistics and the Number System (Cat3)	11	0.99	1.00*	1.00
Algebra 1/Core 7	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	0.98	1.00	
	Statistics and the Number System (Cat3)	11	0.98	1.00*	1.00

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 2/Core 3	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.96	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.94	0.94	1.00
Algebra 2/Core 4	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.97	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.96	0.94	1.00
Geometry/Core 3	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	0.96	1.00	
	Modeling with Geometry (Cat3)	9	0.96	0.97	1.00
Geometry/Core 4	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	0.97	1.00	
	Modeling with Geometry (Cat3)	9	0.94	0.97	1.00

Table 40: Disattenuated Correlation Matrix among Reporting Categories (ELA Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	17	0.96	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.97	0.96	1.00		
	Language and Editing Task (Cat4)	7	0.82	0.88	0.79	1.00	
	Text-Based Writing (Cat5)	1	0.62	0.65	0.65	0.59	1.00
5	Key Ideas and Details (Cat1)	13	1.00				
	Craft and Structure (Cat2)	16	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.97	0.94	1.00		
	Language and Editing Task (Cat4)	8	0.77	0.92	0.73	1.00	
	Text-Based Writing (Cat5)	1	0.65	0.69	0.61	0.66	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
6	Key Ideas and Details (Cat1)	12	1.00				
	Craft and Structure (Cat2)	22	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	1.00*	1.00*	1.00		
	Language and Editing Task (Cat4)	8	0.86	0.85	0.87	1.00	
	Text-Based Writing (Cat5)	1	0.68	0.66	0.73	0.66	1.00
7	Key Ideas and Details (Cat1)	15	1.00				
	Craft and Structure (Cat2)	17	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	1.00*	1.00*	1.00		
	Language and Editing Task (Cat4)	10	0.82	0.87	0.80	1.00	
	Text-Based Writing (Cat5)	1	0.60	0.66	0.55	0.58	1.00
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.97	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	1.00*	0.99	1.00		
	Language and Editing Task (Cat4)	8	0.77	0.90	0.88	1.00	
	Text-Based Writing (Cat5)	1	0.62	0.65	0.64	0.75	1.00
9	Key Ideas and Details (Cat1)	17	1.00				
	Craft and Structure (Cat2)	18	1.00*	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	1.00*	1.00*	1.00		
	Language and Editing Task (Cat4)	8	0.77	0.79	0.83	1.00	
	Text-Based Writing (Cat5)	1	0.68	0.68	0.69	0.62	1.00
10	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	16	0.98	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.96	0.86	1.00		
	Language and Editing Task (Cat4)	12	0.79	0.76	0.76	1.00	
	Text-Based Writing (Cat5)	1	0.64	0.69	0.59	0.64	1.00

Table 41: 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	0.93	1.00			
	Measurement, Data, and Geometry (Cat3)	18	0.94	0.94	1.00		
6	Ratio and Proportional Relationships (Cat1)	8	1.00				
	Expressions and Equations (Cat2)	17	0.94	1.00			
	Geometry (Cat3)	8	0.91	0.90	1.00		
	Statistics and Probability (Cat4)	11	0.93	0.85	0.81	1.00	
	The Number System (Cat5)	11	0.92	0.93	0.79	0.78	1.00
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.96	1.00			
	Geometry (Cat3)	13	0.91	0.89	1.00		
	Statistics and Probability (Cat4)	9	0.88	0.92	0.84	1.00	
	The Number System (Cat5)	8	0.94	0.96	0.86	0.90	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	14	0.93	1.00			
	Geometry (Cat3)	15	0.85	0.88	1.00		
	Statistics & Probability and the Number System (Cat4)	10	0.95	0.90	0.89	1.00	

Table 42: Disattenuated Correlation Matrix among Reporting Categories (EOC Accommodated Forms)

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Algebra 1	Algebra and Modeling (Cat1)	24	1.00		
	Functions and Modeling (Cat2)	23	1.00*	1.00	
	Statistics and the Number System (Cat3)	11	1.00*	1.00*	1.00
Algebra 2	Algebra and Modeling (Cat1)	21	1.00		
	Functions and Modeling (Cat2)	21	0.92	1.00	
	Statistics, Probability, and the Number System (Cat3)	16	0.95	0.93	1.00

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3
Geometry	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1.00		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	22	1.00*	1.00*	
	Modeling with Geometry (Cat3)	9	1.00*	1.00*	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. Mplus is commonly used for collecting validity evidence on the internal structure of assessments. Weighted least squares means and variance adjusted (WLSMV) was employed as the estimation method 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).

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 (θ), 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:

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

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:

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

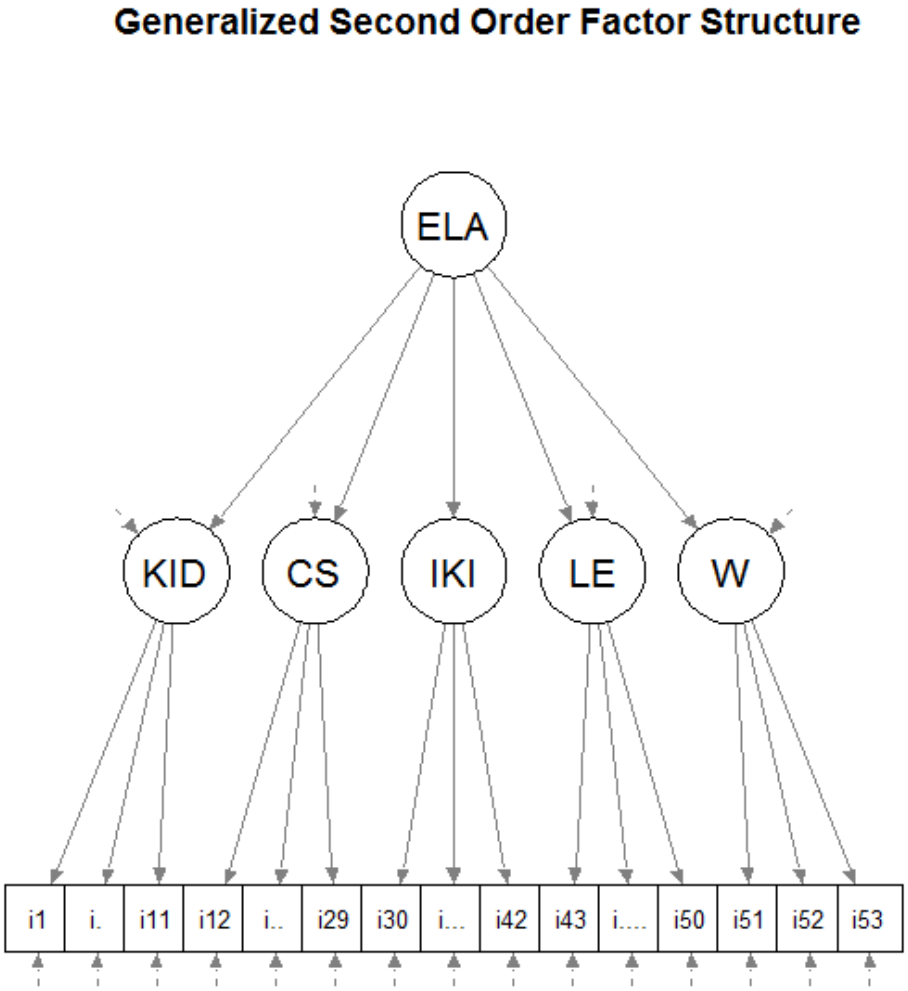
where $\hat{\boldsymbol{\Sigma}}$ is the implied correlation matrix among test items, $\boldsymbol{\Lambda}$ is the *p x k* matrix of first-order factor loadings relating item scores to first-order factors, $\boldsymbol{\Gamma}$is the *k x 1* matrix of second-order factor loadings relating the first-order 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 $\boldsymbol{\Phi} \rightarrow \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$. 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 25, Table 26, and Table 27 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 6. 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 6: 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 43 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 CFI and TLI were equal to or greater than 0.95.

Table 43: Goodness-of-Fit Second-Order CFA

ELA					
Grade	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3	986	0.02	0.98	0.98	Yes
4	1219	0.02	0.99	0.99	Yes
5*	1270	0.02	0.99	0.99	Yes
6	1425	0.02	0.99	0.99	Yes
7	1425	0.02	0.99	0.99	Yes
8*	1426	0.02	0.99	0.99	Yes
9	1534	0.02	0.99	0.99	Yes
10	1479	0.02	0.98	0.98	Yes
Mathematics					
Grade	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
3	1374	0.03	0.96	0.96	Yes
4	1373	0.02	0.98	0.98	Yes
5	1374	0.03	0.97	0.97	Yes
6	1425	0.03	0.98	0.98	Yes
7	1479	0.03	0.98	0.98	Yes
8	1426	0.03	0.95	0.95	Yes
EOC					
Subject/Form	*df*	*RMSEA*	*CFI*	*TLI*	*Convergence*
Alg 1/Core 5	1592	0.03	0.96	0.96	Yes
Alg 1/Core 6	1592	0.03	0.97	0.97	Yes
Alg 1/Core 7	1592	0.03	0.97	0.96	Yes
Alg 2/Core 3	1592	0.03	0.96	0.96	Yes
Alg 2/Core 4	1592	0.03	0.96	0.96	Yes
Geo/Core 3	1592	0.03	0.97	0.97	Yes
Geo/Core 4*	1593	0.03	0.97	0.97	Yes

*For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.

The second-order factor model converged for all tests. However, the residual variance for one factor fell slightly below the boundary of 0 for grades 5 and 8 ELA and Geometry core 4 when using the M-Plus software package. For purposes of exploration, the same model was implemented using the "lavaan" package (Rosseel, 2012) in R or a comparable model was implemented using MPlus for these tests and the model converged without yielding a negative residual variance. This negative residual variance may be related to the computational implementation of the optimization approach in M-Plus, it may be a flag related to model misspecification, or it may be related to other causes (Van Driel, 1978; Chen, Bollen, Paxton, Curran & Kirby, 2001). For parsimony with the other tests, we selected to remain within the M-Plus environment, but constrained the residual variance to 0 for these tests. This is equivalent to treating the parameter as fixed which does not necessarily conform to our a-priori hypothesis.

As indicated in Section 3.1, FSA items are operationally calibrated by IRTPRO software; however, factor analyses presented here were conducted with Mplus software. 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 uses WLSMV based on limited information maximum likelihood and chooses 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 does not include the same flexibility. However, CFA results presented here still indicated good fit indices even though pseudo-guessing was constrained to zero or not taken into account.

In Table 44, Table 45, and Table 46, we provide the estimated correlations between the reporting categories from the second-order factor model for ELA, Mathematics, 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 44: Correlations among ELA Factors

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4	Cat5
3	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.997	1			
	Integration of Knowledge and Ideas (Cat3)	0.98	0.99	1		
	Language and Editing Task (Cat4)	0.87	0.88	0.86	1	
4	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.99	1			
	Integration of Knowledge and Ideas (Cat3)	0.98	0.98	1		
	Language and Editing Task (Cat4)	0.91	0.92	0.91	1	
	Text-Based Writing (Cat5)	0.71	0.72	0.71	0.66	1

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4	Cat5
5*	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.999	1			
	Integration of Knowledge and Ideas (Cat3)	0.98	0.98	1		
	Language and Editing Task (Cat4)	0.93	0.94	0.92	1	
	Text-Based Writing (Cat5)	0.72	0.72	0.71	0.67	1
6	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.98	1			
	Integration of Knowledge and Ideas (Cat3)	0.99	0.99	1		
	Language and Editing Task (Cat4)	0.90	0.90	0.91	1	
	Text-Based Writing (Cat5)	0.75	0.75	0.75	0.69	1
7	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.99	1			
	Integration of Knowledge and Ideas (Cat3)	0.997	0.99	1		
	Language and Editing Task (Cat4)	0.89	0.88	0.88	1	
	Text-Based Writing (Cat5)	0.69	0.69	0.69	0.62	1
8*	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.99	1			
	Integration of Knowledge and Ideas (Cat3)	0.99	0.99	1		
	Language and Editing Task (Cat4)	0.87	0.86	0.87	1	
	Text-Based Writing (Cat5)	0.72	0.71	0.72	0.63	1
9	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.99	1			
	Integration of Knowledge and Ideas (Cat3)	0.997	0.995	1		
	Language and Editing Task (Cat4)	0.83	0.83	0.83	1	
	Text-Based Writing (Cat5)	0.72	0.72	0.72	0.60	1
10	Key Ideas and Details (Cat1)	1				
	Craft and Structure (Cat2)	0.97	1			
	Integration of Knowledge and Ideas (Cat3)	0.95	0.94	1		
	Language and Editing Task (Cat4)	0.87	0.86	0.84	1	
	Text-Based Writing (Cat5)	0.75	0.74	0.73	0.66	1

*For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.

Table 45: 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)	0.88	1			
	Measurement, Data, and Geometry (Cat3)	0.94	0.90	1		
4	Operations and Algebraic Thinking (Cat1)	1				
	Numbers and Operations in Base Ten (Cat2)	0.97	1			
	Numbers and Operations – Fractions (Cat3)	0.97	0.97	1		
	Measurement, Data, and Geometry (Cat4)	0.94	0.93	0.94	1	
5	Operations, Algebraic Thinking, and Fractions (Cat1)	1				
	Numbers and Operations in Base Ten (Cat2)	0.95	1			
	Measurement, Data, and Geometry (Cat3)	0.95	0.92	1		
6	Ratio and Proportional Relationships (Cat1)	1				
	Expressions and Equations (Cat2)	0.98	1			
	Geometry (Cat3)	0.93	0.95	1		
	Statistics and Probability (Cat4)	0.91	0.93	0.89	1	
	The Number System (Cat5)	0.95	0.96	0.92	0.90	1
7	Ratio and Proportional Relationships (Cat1)	1				
	Expressions and Equations (Cat2)	0.97	1			
	Geometry (Cat3)	0.94	0.94	1		
	Statistics and Probability (Cat4)	0.94	0.93	0.91	1	
	The Number System (Cat5)	0.95	0.94	0.92	0.91	1
8	Expressions and Equations (Cat1)	1				
	Functions (Cat2)	0.89	1.00			
	Geometry (Cat3)	0.89	0.87	1.00		
	Statistics & Probability and the Number System (Cat4)	0.93	0.90	0.90	1	

Table 46: Correlations among EOC Factors

Course/Form	Reporting Category	Cat1	Cat2	Cat3
Algebra 1/Core 5	Algebra and Modeling (Cat1)	1		
	Functions and Modeling (Cat2)	0.97	1	
	Statistics and the Number System (Cat3)	0.98	0.98	1
Algebra 1/Core 6	Algebra and Modeling (Cat1)	1		
	Functions and Modeling (Cat2)	0.99	1	
	Statistics and the Number System (Cat3)	0.97	0.99	1
Algebra 1/Core 7	Algebra and Modeling (Cat1)	1		
	Functions and Modeling (Cat2)	0.97	1	
	Statistics and the Number System (Cat3)	0.97	0.99	1
Algebra 2/Core 3	Algebra and Modeling (Cat1)	1		
	Functions and Modeling (Cat2)	0.96	1	
	Statistics, Probability, and the Number System (Cat3)	0.95	0.95	1
Algebra 2/Core 4	Algebra and Modeling (Cat1)	1		
	Functions and Modeling (Cat2)	0.96	1	
	Statistics, Probability, and the Number System (Cat3)	0.97	0.93	1
Geometry/Core 3	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	1		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	0.96	1	
	Modeling with Geometry (Cat3)	0.97	0.97	1
Geometry/Core 4*	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	1		
	Circles, Geometric Measurement and Geometric Properties with Equations (Cat2)	0.967	1	
	Modeling with Geometry (Cat3)	0.981	0.986	1

*For these tests, the second-order model was run by constraining the residual variance of a certain factor to zero due to non-significant negative residual variance.

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):

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

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:

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

where u_{ij} is the item score of the *i*th examinee for item *j*, $T_j(\hat{\theta}_i)$ 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*-1)/2, Q_3 statistics will be produced. The Q_3 values are expected to be small. Table 47, Table 48, and Table 49 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 47: 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	0.023	-0.090	-0.049	-0.020	0.011	0.198	-0.062	0.198
4	0.036	-0.293	-0.089	-0.014	0.059	0.707	-0.127	0.180
5	0.033	-0.186	-0.080	-0.015	0.047	0.753	-0.084	0.180
6	0.040	-0.271	-0.098	-0.015	0.066	0.721	-0.099	0.400
7	0.037	-0.195	-0.092	-0.015	0.058	0.817	-0.108	0.124
8	0.038	-0.261	-0.094	-0.010	0.065	0.793	-0.092	0.249
9	0.030	-0.194	-0.076	-0.014	0.042	0.784	-0.112	0.105
10	0.029	-0.236	-0.069	-0.012	0.049	0.870	-0.064	0.179

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

Table 48: Mathematics Q_3 Statistic

Grade	Average Zero-Order Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.025	-0.207	-0.050	-0.019	0.021	0.256
4	0.022	-0.099	-0.046	-0.017	0.018	0.187
5	0.052	-0.262	-0.123	-0.017	0.099	0.391
6	0.049	-0.390	-0.120	-0.015	0.089	0.297
7	0.052	-0.351	-0.134	-0.013	0.105	0.398
8	0.047	-0.277	-0.111	-0.020	0.083	0.370

Table 49: EOC Q_3 Statistic

Course	Average Zero-Order Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Algebra 1	0.079	-0.583	-0.202	-0.012	0.174	0.823
Algebra 2	0.055	-0.365	-0.137	-0.015	0.118	0.443
Geometry	0.054	-0.316	-0.135	-0.015	0.128	0.519

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 4 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 Mathematics and grade 3 Reading, there were no accommodated forms, as these tests were universally administered on paper. For other grades, the number of items replaced between the online and paper accommodated forms is provided in Table 50. In EOC, the first core form (Core 5 for Algebra 1 and Core 3 for Algebra 2 and Geometry) was administered as the accommodated version.

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

For the 2015–2016 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 2016 FSA Technical Report.

6.2 COMPARABILITY OF FSA TEST SCORES OVER TIME

The comparability of FSA scores over time was ensured via two methods. First, during test construction both a content and statistical perspective were implemented. The FSA test items placed onto forms in both spring 2015 and spring 2016 were aligned to the same standards and test blueprint specifications. In addition, spring 2015 form statistics were used as targets for both numerical and graphical summaries for the spring 2016 forms. See Section 4 of Volume 2 for details about both the content and statistical methods. Second, during spring 2016 calibrations, equating was performed in order to place item parameters estimates from spring 2016 onto the 2015 baseline scale. The equating procedure and results are presented in Volume 1 Section 6.2.

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-and-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 test characteristic curves (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 EOC, 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 50, only one item was replaced in ELA Grade 5. TCCs for both ELA and Mathematics are presented in Appendix D.

Table 50: Number of Item Replacements for the Accommodated Forms

Mathematics	Number of Items Replaced	ELA	Number of Items Replaced
Grade 5	9	Grade 4	0
Grade 6	16	Grade 5	1
Grade 7	16	Grade 6	0
Grade 8	17	Grade 7	0
Algebra 1	17	Grade 8	0
Algebra 2	21	Grade 9	0
Geometry	21	Grade 10	0

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 2015–2016 FSA Annual Technical Report. The DIF statistics for each test item are presented in the appendices of Volume 1 of the 2015–2016 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.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87*(3), 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, F., Kenneth A. Bollen, P. Paxton, P. Curran, and J. Kirby. 2001. ‘‘Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies.’’ Sociological Methods & Research 29:468-508.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 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: Holt, Rinehart, and Winston.

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.

Florida Department of Education. (2013). *Florida Statewide Assessments 2013 Yearbook.*

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6).

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*(3), 381–389.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*(3), 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.

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*(1), 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. (2012). Mplus user's guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

New York State Education Department (2014). *New York State testing program 2014: English language arts and mathematics grades 3–8*. Retrieved from http://www.p12.nysed.gov/assessment/reports/2014/gr38cc-tr14.pdf

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

Phillips, G. W. (2016). *National benchmarks for state achievement standards*. Washington, DC: American Institutes for Research.

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/.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14).

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

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. Retrieved October 2002, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

van Driel, Otto P. 1978. ‘‘On Various Causes of Improper Solutions in Maximum Likelihood Factor Analysis.’’ Psychometrika 43:225-43.

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 assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 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.