Florida
Statewide Assessments

Florida Statewide Assessments

2021–2022

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 (FDOE) (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Dr. Ahmet Turhan, Dr. Yanlin Jiang, Dr. Sherry Li, Dr. Peter Diao, Tyler Lonczak, Matt Gordon, Cameron Clark, Zoe Dai, and Melissa Boyanton. Contributing staff from Pearson are Dr. Jie (Serena) Lin, Dr. Seong Eun (Jane) Hong, Ying Meng, and Ebony Gaines. The major contributors from the FDOE are as follows: Vince Verges, Susie Lee, Jenny Black, Robert Bowman, Jessica Graverholt, Dr. Qian Liu, Racquel Harrell, Sally Donnelly, Travis Barton, Leah Glass, Dr. Stacy Skinner, Dr. Salih Binici, Yachen Luo, Wenyi Li, Jielin Ming, and Saeyan Yun.

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 started to implement the Florida Standards Assessments (FSA) for operational use to replace the Florida Comprehensive Assessment Tests (FCAT) 2.0 in English Language Arts (ELA) and Mathematics during the 2014–2015 school year. Students in Grades 3 and 4 were administered fixed, operational ELA Reading and Mathematics tests on paper. Students in Grades 5 through 10 were administered fixed, operational Reading tests online, and students in Grades 5 through 8 were administered fixed, operational Mathematics tests online. 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 spring 2016, the Grade 4 Reading portion of the ELA assessment transitioned to an online delivery. In spring 2017, the Grades 3 and 4 Mathematics assessment moved online too. Beginning in summer 2017, Algebra 2 was no longer administered. In the grades with online testing, paper forms, in lieu of online forms were administered to students whose Individual Educational Plans (IEPs) or Section 504 Plans indicated such a need. Beginning in spring 2019, some grades and subjects were transitioned to a different mode of delivery per House Bill 7069. Grades 4–6 Reading and Grades 3–6 Mathematics moved from online assessments back to paper assessments, and Grade 7 Writing was transitioned from paper assessment to online assessment in spring 2019. Grade 3 Reading was still universally administered on paper.

The Next Generation Sunshine State Standards (NGSSS) were adopted in 2008 to replace the 1996 Sunshine State Standards. The first operational administration of the Science assessments (in Grades 5 and 8) and Biology 1 end-of-course (EOC) was during the spring 2012 administration window. During the spring 2013 administration window the first operational administration of the U.S. History EOC assessment occurred. In the administrative year of 2014, the first operational Civics EOC assessment was administered.

Since fall 2020, all FSA and NGSSS assessments have been collectively referred to as the Florida Statewide Assessments.

Table 1 displays the complete list of test forms for the spring operational administration.

Table 1: Test Administration

Subject	Administration	Grade/Course
ELA Reading	Paper	3–6
	Online	7–10
	Paper (Accommodated)	
ELA Writing	Paper	4–6
	Online	7–10
	Paper (Accommodated)	

Subject	Administration	Grade/Course
Mathematics	Paper	3–6
	Online	7–8
	Paper (Accommodated)	
Science	Paper	5, 8
EOC	Online	Algebra 1, Geometry, Biology 1, Civics, U.S. History
	Paper (Accommodated)	

With the implementation of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the Florida Statewide Assessments scores.

This volume provides empirical evidence about the reliability and validity of the spring 2022 Florida Statewide Assessments, given its intended uses.

Specifically, 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 marginal reliability. The reliability estimates are presented by grade and subject as well as by demographic subgroup. This section also includes conditional standard errors of measurement (CSEM) and classification accuracy results by grade and subject.
- **Validity.** This volume as well as other volumes of this report provide validity evidence supporting the appropriate inferences from Florida Statewide Assessment scores. 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. Evidence is also 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 Evidence.** By examining the blueprint match between forms and test characteristic curves (TCCs) for both forms, we evaluate comparability of test scores across forms. Comparability of constructs, scores, and technical properties of scores are evaluated and discussed.
- **Test Fairness.** Fairness is statistically analyzed using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

2. PURPOSE OF FLORIDA STATEWIDE ASSESSMENTS

The Florida Statewide Assessments 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, and NCME, 2014) and to ensure that all students have access to the test content via principles of universal design and appropriate accommodations.

The Florida Statewide Assessments yield 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. Scores can also be aggregated to evaluate the performance of subgroups, and both individual and aggregated scores will be compared over time in program evaluation methods.

The Florida Statewide Assessments results serve as the primary indicator for the state's accountability system. The policy and legislative purpose of the Florida Statewide Assessments is described more thoroughly in Volume 1 of this technical report. The test is a standards-based assessment designed to measure student achievement toward the state content standards. Florida Statewide Assessments scores are indications of what students know and can 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 Florida Statewide Assessments requirements.

In the Florida Statewide Assessments administered in 2022, student-level scores included scale scores and raw scores at the reporting category level. The FSA performance cuts were approved by the State Board of Education (SBE) on January 6, 2016, the cut scores of Grades 5 and 8 Science, and Biology 1 were approved by SBE in 2012, and the cut scores of U.S. History and Civics were approved by SBE in 2013 and 2014, respectively. Based on the cut scores of Florida Statewide Assessments approved by SBE, scale scores and achievement levels were reported in spring 2022. Volume 1 Section 8.1 of the Florida Statewide Assessments 2021-2022 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 they are viewed with the usual caution that accompanies the 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

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 test takers' 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 test 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 earlier as $\sigma_X\sqrt{1 - \rho_{XX'}}$, where σ_X is the standard deviation of the scaled score and $\rho_{XX'}$ is a reliability coefficient. Based on the definition of reliability, this formula can be derived.

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

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

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

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

In general, the SEM 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 SEM in the classical test theory is assumed to be homoscedastic error irrespective of the standard deviation of a group.

In contrast, the SEM in IRT vary over the ability continuum. These heterogeneous errors are a function of a test information function that provides different information about test takers 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 have maximum information 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.

3.1 INTERNAL CONSISTENCY

As the Florida Statewide Assessments were administered in a single administration, it is necessary to examine the internal consistency of the tests to support the reliability of the test scores. For the ELA, Mathematics, Science, and EOC Florida Statewide Assessments the reliability coefficients were computed using Cronbach *alpha*, stratified *alpha*, and Feldt-Raju. 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 ELA, Mathematics, Science, and EOC Florida Statewide 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 score variance of the *i*th stratum, and σ_x^2 is the variance of the total test scores. The stratified Cronbach *alpha* coefficient accounts for 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 the 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 divided by the total score variance. 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 7 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. All the items administered in Science and Social Studies are multiple-choice only.

Table 2: Mathematics Item Types and Descriptions

Response Type	Description
Multiplechoice (MC)	Student selects one correct answer from a number of options.
Multipleselect (MS)	Student selects all correct answers from a number of options.
Edittaskchoice (ETC)	Student identifies an incorrect word, phrase, or blank and chooses the replacement from a number of options.
grid (GI)	The student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hottext (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 keypad with a variety of mathematical symbols to create a response.
Tablematch (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Tableinput (TI)	Student types numeric values into a given table.
multi-interaction (MULTI)	An item that contains more than one response types. It could contain more than one of the same response types or a combination of response types.

Table 3: Reading Item Types and Descriptions

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from a number of options.
multipleselect (MS)	Student selects all correct answers from a number of options.
tablematch (MI)	Student checks a box to indicate if information from a column header matches information from a row.
edittaskwithchoice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
hottext (HT)	Student is directed to either select or use drag-and-drop feature to use text to support an analysis or make an inference.
multiplechoice, hottextselectable (Two-part HT)	Student selects the correct answers from Part A and Part B. Part A is a multiple-choice or a multiselect, and Part B is a selectable HT.
evidence-basedselectedresponse (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.

Table 4: NGSSS Science and EOC Item Type and Description

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from a number of options.

Table 5: Mathematics Operational Item Types by Grade

Item Type *	Grade						Algebra 1**	Geometry**
	3	4	5	6	7	8		
MC	27	23	29	30	20	25	33; 33; 32; 31	36; 40; 37
MS	6	8	6	8	2	5	2; 2; 2; 1	3; 1; 1
GI	-	-	-	-	5	3	1; 1; 2; 2	1; 1; 1
HT	-	-	-	-	-	1	0; 0; 0; 1	0; 0; 0
TI	-	-	-	-	1	1	1; 1; 2; 1	0; 0; 0
MI	2	6	4	2	2	1	0; 1; 0; 0	0; 0; 0
EQ	18	16	15	9	18	14	8; 7; 8; 11	8; 9; 9
ETC	1	1	-	6	5	2	9; 8; 8; 7	8; 7; 7
Multi	-	-	-	1	3	4	4; 5; 4; 4	2; 0; 3

* Descriptions for each item type are presented in Table 2.
** Algebra 1 has four core forms and Geometry has three core forms.

Table 6: Reading Operational Item Types by Grade

Item Type *	Grade							
	3	4	5	6	7	8	9	10
MC	37	33	36	37	33	40	34	36
MS	3	6	3	3	8	5	5	5
MI	-	5	4	1	1	-	2	1
ETC	-	-	-	3	4	-	-	-
HT	4	-	3	1	2	1	2	2
EBSR	6	6	4	7	4	6	11	9
Two-Part HT	-	-	-	-	-	-	-	1

* Descriptions for each item type are presented in Table 3.

Table 7: Science and Social Studies Operational Item Types by Grade

Item Type *	Grade		Biology 1**	U.S. History**	Civics**
	5	8			
MC	56	56	56; 56; 56	52; 52; 52	48; 48; 48

* Descriptions for each item type are presented in Table 4.
** Biology 1, U.S. History, and Civics have three core forms.

Table 8 through Table 11 present the Cronbach *alpha* coefficients for Mathematics, ELA, Science, and EOC by grade/course and test form. These tables also include stratified *alpha* and Feldt-Raju coefficients for Mathematics, ELA, and FSA EOC. Please note that both stratified alpha and Feldt-Raju coefficients are not applicable for NGSSS EOC and Science since there are only MC items in these tests.

The Cronbach *alpha* ranged from 0.88 to 0.95 for Mathematics, 0.91 to 0.93 for ELA, 0.89 to 0.94 for EOC, and 0.93 to 0.94 for Science. The stratified *alpha* coefficients ranged from 0.88 to 0.95 for Mathematics, 0.91 to 0.93 for ELA, and 0.89 to 0.94 for FSA EOC. The Feldt-Raju coefficients were between 0.87 and 0.95 for Mathematics, 0.90 and 0.93 for ELA, and 0.90 and 0.95 for FSA EOC. The reliability coefficients by each demographic subgroup and for each reporting category are presented in Appendix A, Reliability Coefficients.

Table 8: Reliability Coefficients (Mathematics)

Grade	Form	Cronbach Alpha	Stratified Alpha	Feldt-Raju
3	Paper	0.95	0.95	0.94
4	Paper	0.95	0.95	0.95
5	Paper	0.94	0.95	0.94
6	Paper	0.94	0.94	0.94
7	Online	0.94	0.94	0.92
	Accommodated	0.88	0.88	0.87
8	Online	0.91	0.91	0.91
	Accommodated	0.88	0.88	0.89

Table 9: Reliability Coefficients (ELA)

Grade	Form	Cronbach Alpha	Stratified Alpha	Feldt-Raju
3	Paper	0.92	0.92	0.92
4	Paper	0.93	0.93	0.92
5	Paper	0.93	0.93	0.92
6	Paper	0.92	0.92	0.91
7	Online	0.93	0.93	0.92
	Accommodated	0.91	0.91	0.90
8	Online	0.93	0.93	0.92
	Accommodated	0.92	0.92	0.90
9	Online	0.93	0.93	0.93
	Accommodated	0.92	0.92	0.91
10	Online	0.93	0.93	0.92
	Accommodated	0.93	0.93	0.92

Table 10: Reliability Coefficients (EOC)

Course	Form*	Cronbach Alpha	Stratified Alpha	Feldt-Raju
Algebra	Online – Core 24	0.93	0.93	0.93
	Online – Core 25	0.93	0.93	0.93
	Online – Core 26	0.94	0.94	0.94
	Online – Core 27	0.93	0.94	0.93
	Accommodated	0.91	0.91	0.92
Geometry	Online – Core 19	0.93	0.93	0.94
	Online – Core 20	0.93	0.93	0.95
	Online – Core 21	0.93	0.93	0.94
	Accommodated	0.89	0.89	0.90
Biology 1	Online – Core 100	0.92	-	-
	Online – Core 200	0.91	-	-
	Online – Core 300	0.92	-	-
	Accommodated	0.90	-	-
Civics	Online – Core 100	0.92	-	-
	Online – Core 200	0.91	-	-
	Online – Core 300	0.92	-	-
	Accommodated	0.90	-	-
U.S. History	Online – Core 100	0.92	-	-
	Online – Core 200	0.91	-	-
	Online – Core 300	0.91	-	-
	Accommodated	0.90	-	-

* Since spring, 2015, 3-4 core forms for Mathematics EOC have been developed annually. Each core form is assigned to a unique number.

Table 11: Reliability Coefficients (Science)

Grade	Form	Cronbach *Alpha*	Stratified *Alpha*	Feldt-Raju
5	Paper	0.93	-	-
8	Paper	0.94	-	-

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 computed using operational items.

Within the item response theory (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 SEM. The SEM 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 farther away from the middle.

The marginal reliability is defined as:

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

where $\sigma_e^2(\hat{\theta})$ is the function generating the SEM and $f(\hat{\theta})$ is the assumed population density. The marginal reliability of a test is computed by integrating θ out of the test information function as follows:

$$\rho = \frac{\sigma_\theta^2 - \bar{\sigma}_e^2}{\sigma_\theta^2}$$

where σ_θ^2 is the true score variance of θ and:

$$\bar{\sigma}_e^2 = \int_{-\infty}^{\infty} \frac{1}{I(\theta)} g(\theta) d\theta$$

where $g(\theta)$ is a density function. If population parameters are assumed normal, then $g(\theta) \sim N(\mu, \sigma^2)$. In the absence of information about the population distribution of θ, a uniform prior is available such that $g(\theta) \sim \mathrm{U}[\mathrm{a}, \mathrm{b}]$ where a and b are the lower and upper limits of the uniform distribution, respectively. The integral is evaluated using Gauss-Hermite quadrature:

$$\bar{\sigma}_e^2 \approx \sum_{q=1}^{Q} \frac{1}{I(\theta_q)} w_q$$

where θ_q is the value at node q and w_q is the weight at node q. The true score variance of θ can be obtained from the marginal maximum likelihood (MML) means procedure.

In IRT, the marginal likelihood is typically maximized to estimate item parameters by integrating θ out of the function and treating population parameters as known. However, suppose the item parameters are treated as fixed but the population parameters are treated as latent. Then, the following marginal likelihood can be maximized with respect to the two latent parameters associated with the normal population distribution:

$$\arg\max L(\mu,\sigma) = \prod_{i=1}^{N}\int_{-\infty}^{\infty}\prod_{j=1}^{K} p\left(x_j\middle|\theta_i,\boldsymbol{\Upsilon_j}\right)g(\theta|\mu,\sigma)d\theta$$

where in this context $p\left(x_j\middle|\theta_i,\boldsymbol{\Upsilon_j}\right)$ is used to mean the probability of individual $i=\{1,2,\dots,N\}$ having observed response x to item $j=\{1,2,\dots,K\}$ given the vector of item parameters, $\boldsymbol{\Upsilon}$. The integral has no closed form and so the function is evaluated using a fixed quadrature routine. Rather than using Gauss-Hermite, Q nodes are chosen from the normal distribution at fixed points and then the integral is evaluated by summation over the Q nodes as:

$$\arg\max L(\mu,\sigma) = \prod_{i=1}^{N}\sum_{q=1}^{Q}\prod_{j=1}^{K} p\left(x_j\middle|\theta_q,\boldsymbol{\Upsilon_j}\right)g\left(\theta_q\middle|\mu,\sigma\right)$$

where θ_q is node q. In this instance, fixed quadrature points allow a smaller number of likelihood evaluations because the values for θ_q are fixed. If Gauss-Hermite were used, the nodes would change as each value of μ and σ are updated and the likelihood calculations would need to be performed at each iteration. Table 12 presents the marginal reliability coefficients for all students. The marginal reliability coefficients for all subjects and grades were ranging from 0.82 to 0.93.

Table 12: Marginal Reliability Coefficient

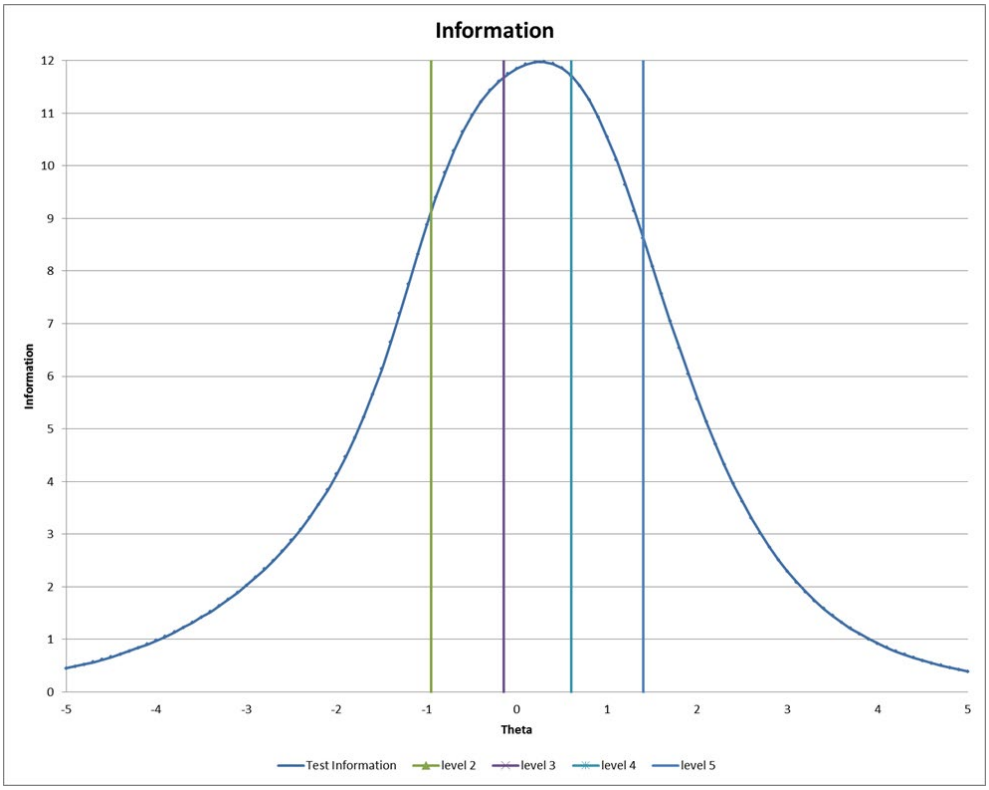
Subject	Grade	Marginal Reliability
ELA	3	0.89
	4	0.91
	5	0.91
	6	0.90
	7	0.91
	8	0.91
	9	0.92
	10	0.91
Mathematics	3	0.92
	4	0.93
	5	0.92
	6	0.88
	7	0.89
	8	0.87
Algebra	Core 24	0.86
	Core 25	0.85
	Core 26	0.84
	Core 27	0.87
Geometry	Core 19	0.85
	Core 20	0.84
	Core 21	0.83
USH	Core 100	0.85
	Core 200	0.85
	Core 300	0.82
Civics	Core 100	0.85
	Core 200	0.86
	Core 300	0.86
Biology 1	Core 100	0.82
	Core 200	0.85
	Core 300	0.86
Science	5	0.89
	8	0.90

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 Florida Statewide Assessments. 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 test takers 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 Florida Statewide Assessments is calculated as:

$$TIF(\theta_i) = \sum_{j=1}^{N_{GPCM}} D^2 a_j^2 \left(\frac{\sum_{s=1}^{m_j} s^2 Exp\left(\sum_{h=1}^{s} Da_j(\theta_i - b_{jh})\right)}{1 + \sum_{s=1}^{m_j} Exp\left(\sum_{h=1}^{s} Da_j(\theta_i - b_{jh})\right)} - \left(\frac{\sum_{s=1}^{m_j} sExp\left(\sum_{h=1}^{s} Da_j(\theta_i - b_{jh})\right)}{1 + \sum_{s=1}^{m_j} Exp\left(\sum_{h=1}^{s} Da_j(\theta_i - b_{jh})\right)} \right)^2 \right) + \sum_{j=1}^{N_{3PL}} D^2 a_j^2 \left(\frac{Q_j}{P_j} \left[\frac{P_j - c_j}{1 - c_j} \right]^2 \right),$$

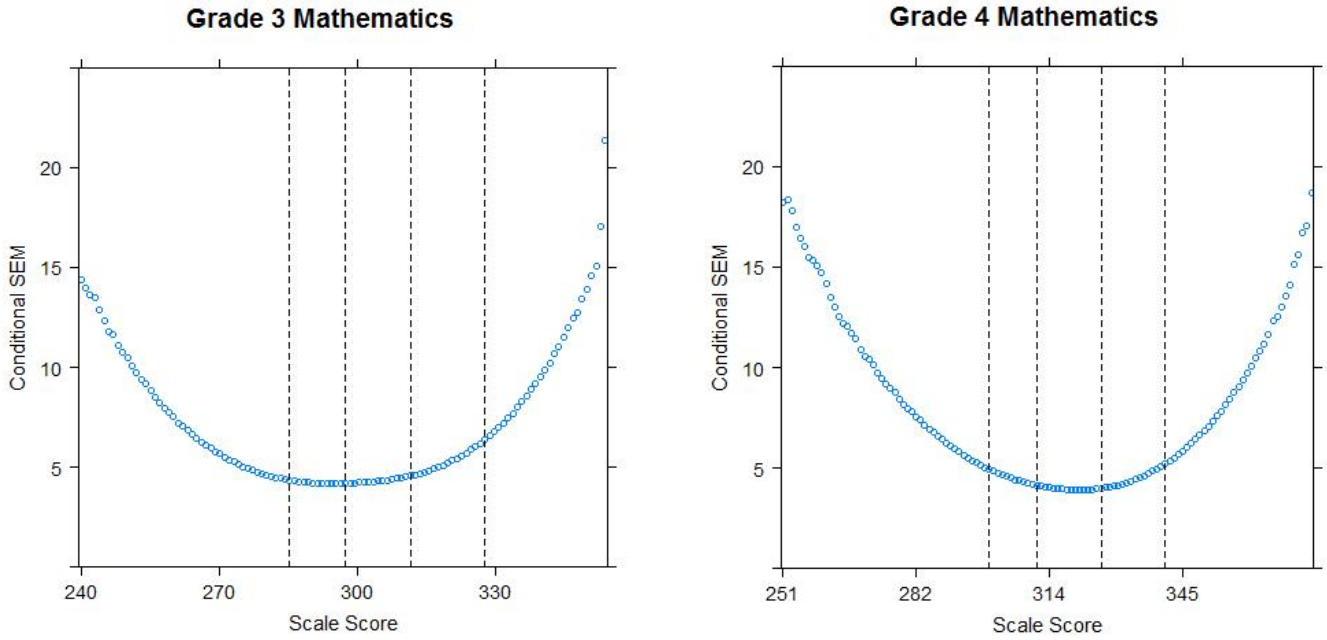
where N_{GPCM} is the number of items that are scored using generalized partial credit model (GPCM) items, N_{3PL} is the number of items scored using 3PL or 2PL model, j indicates item j ($j \in \{1,2,\ldots,N\}$), m_j is the maximum possible score of the item, s indexes step of the item, b_{jh} is the hth step for item j with m total categories. θ_i is the ability of student i.

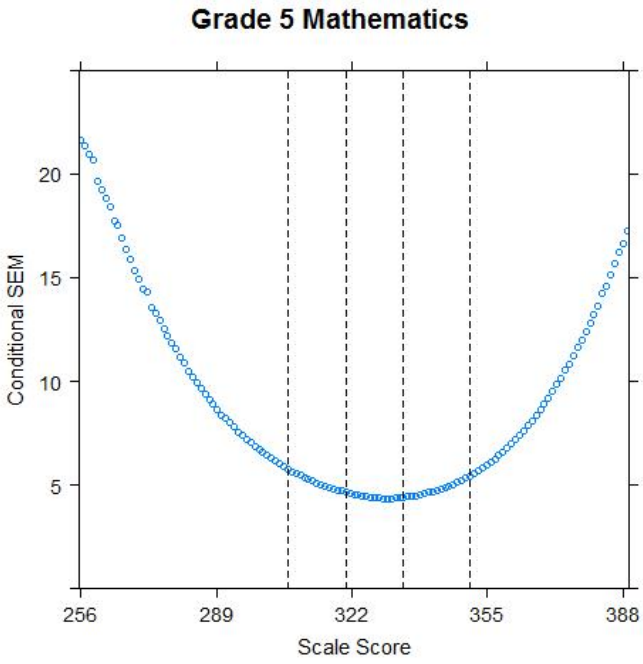
The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

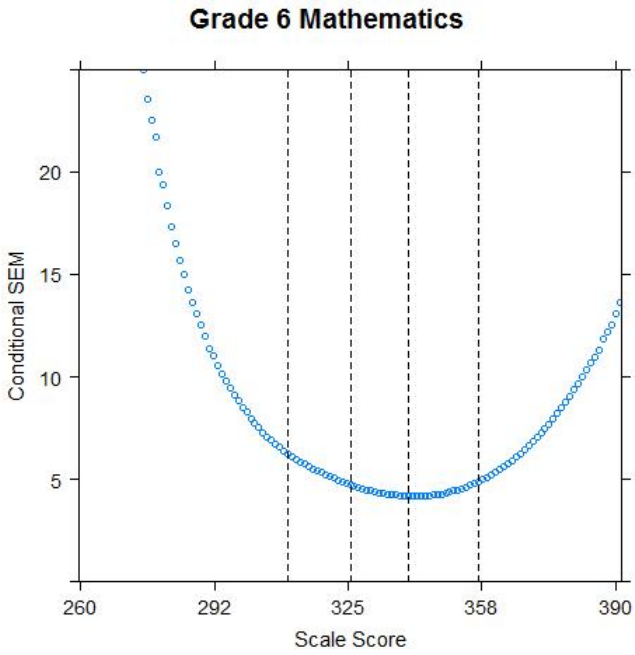
$$se(\theta_i) = \frac{1}{\sqrt{TIF(\theta_i)}}.$$

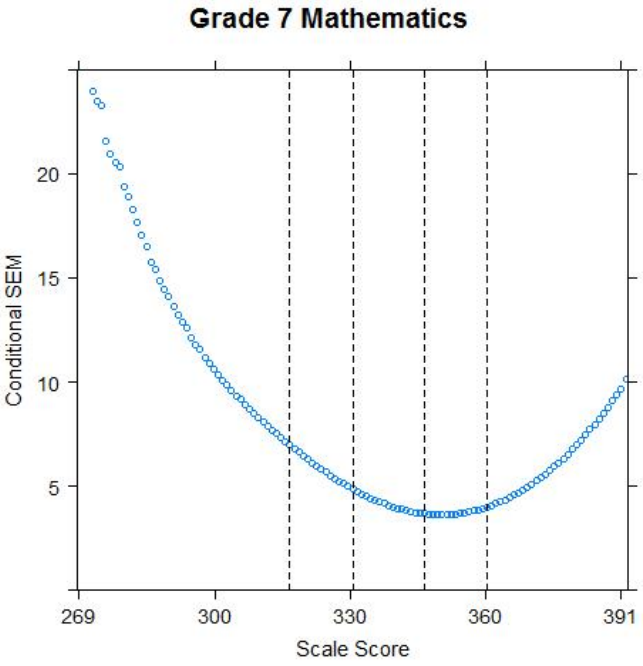
It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2, Figure 3, Figure 4 and Figure 5, respectively, instead of the TIFs, for Mathematics, ELA, EOC, and Science. Vertical lines represent the four performance category cut scores.

Figure 2: Conditional Standard Errors of Measurement (Mathematics)




Grade 5 Mathematics
Conditional SEM
20
15
10
5
256
289
322
355
388
Scale Score


Grade 6 Mathematics
Conditional SEM
20
15
10
5
260
292
325
358
390
Scale Score


Grade 7 Mathematics
Conditional SEM
20
15
10
5
269
300
330
360
391
Scale Score

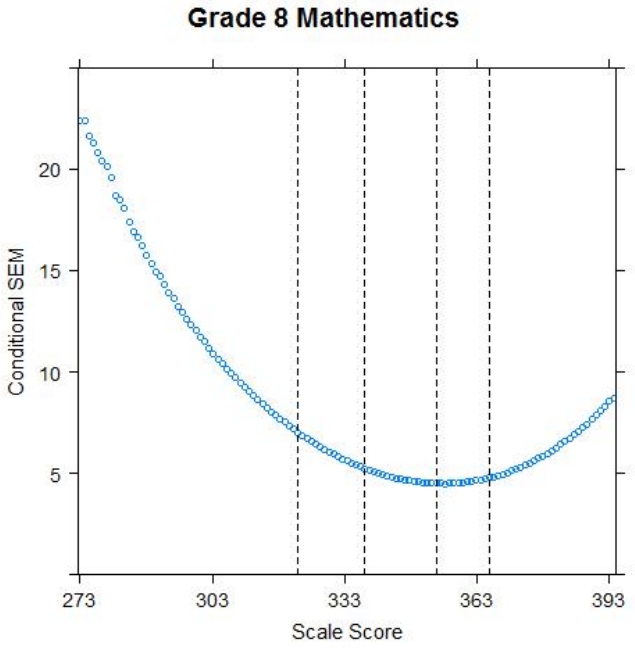
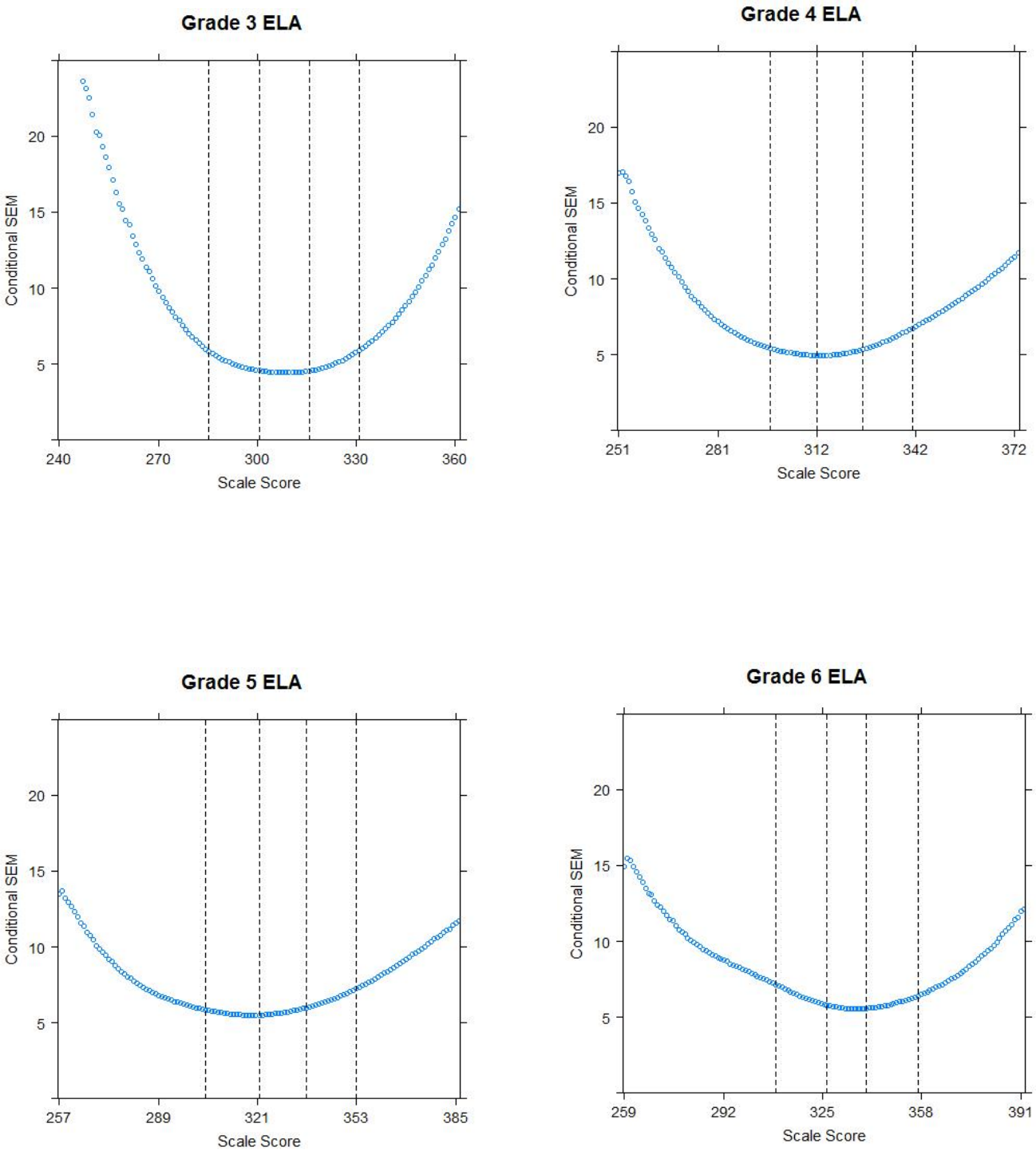
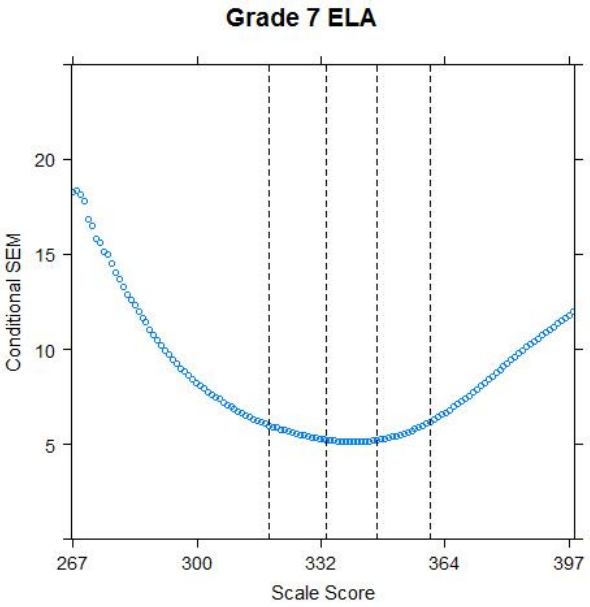
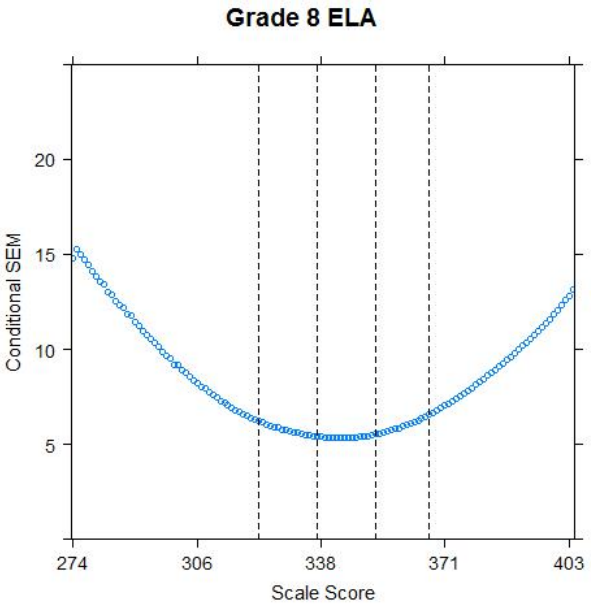
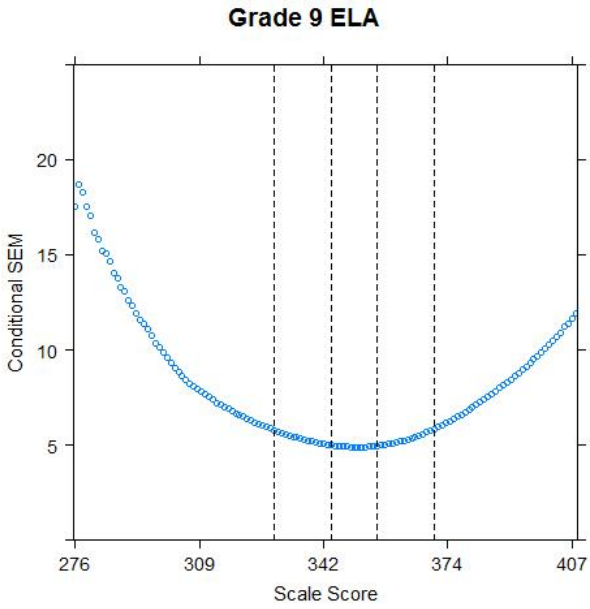

Grade 8 Mathematics
Conditional SEM
20
15
10
5
273
303
333
363
393
Scale Score

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




Grade 7 ELA
Conditional SEM
20
15
10
5
267
300
332
364
397
Scale Score


Grade 8 ELA
Conditional SEM
20
15
10
5
274
306
338
371
403
Scale Score


Grade 9 ELA
Conditional SEM
20
15
10
5
276
309
342
374
407
Scale Score

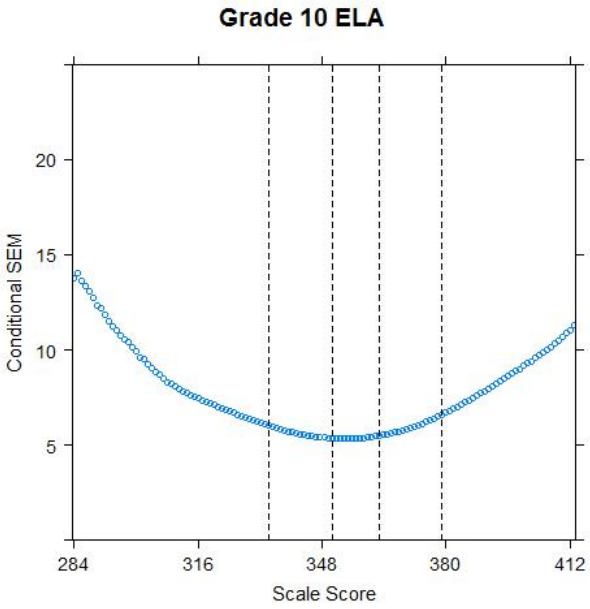
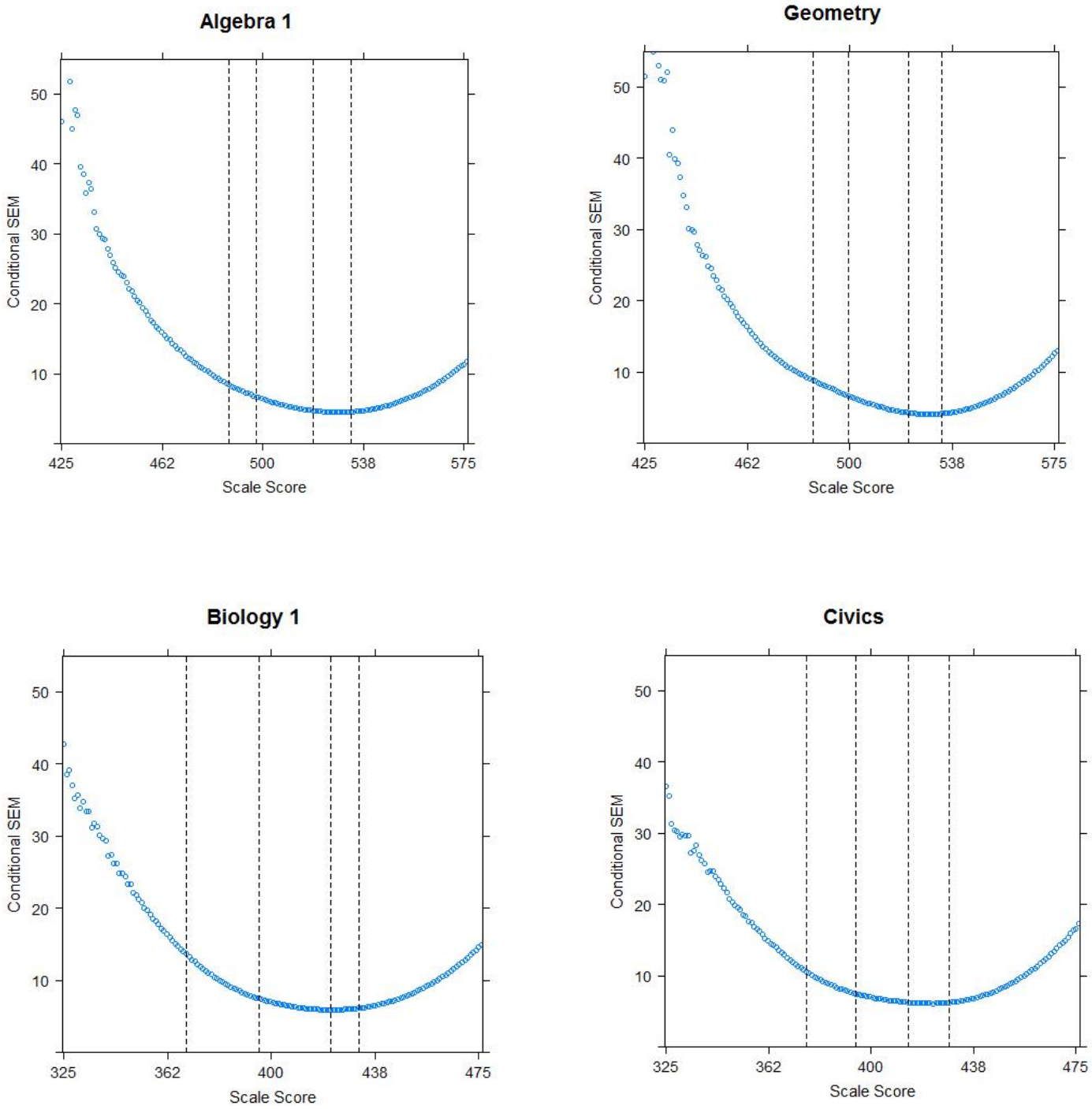

Grade 10 ELA
Conditional SEM
20
15
10
5
284
316
348
380
412
Scale Score

Figure 4: Conditional Standard Errors of Measurement (EOC)



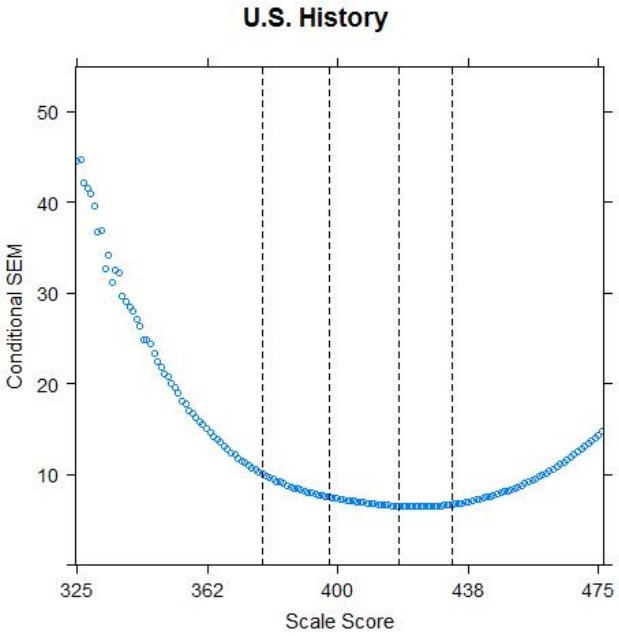
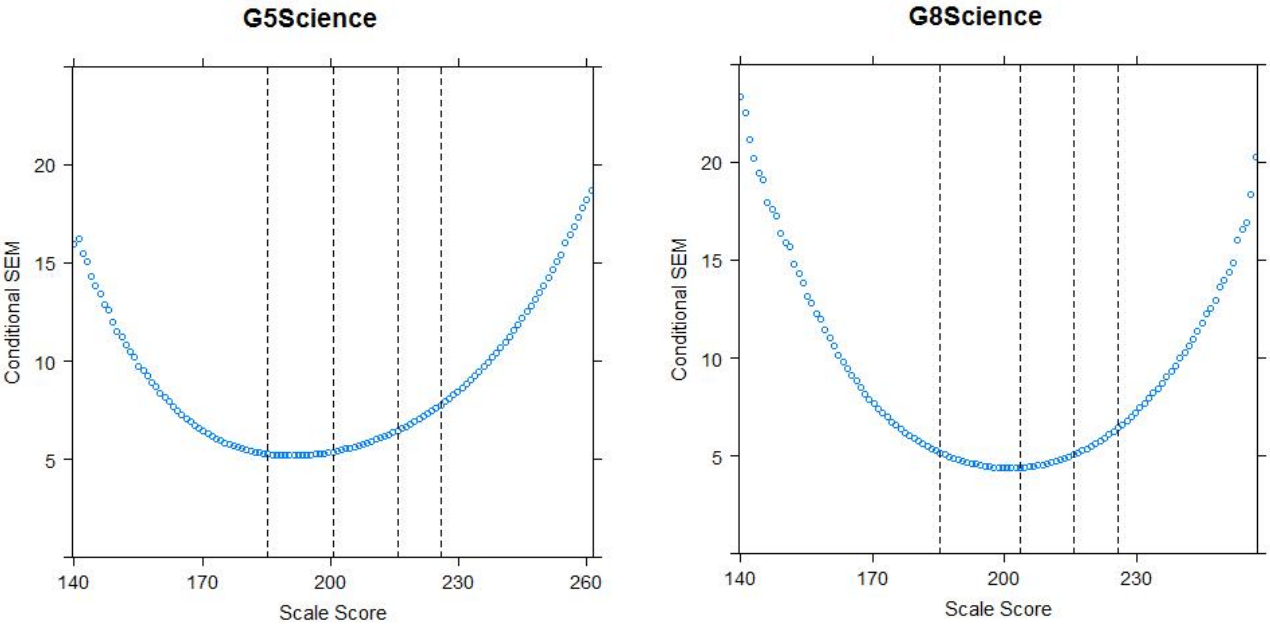


Figure 5: Conditional Standard Errors of Measurement (Science)



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 7 and 8 Mathematics and all EOC tests, the standard error curve is minimized at a higher point along the Florida Statewide Assessments score scale. This suggests the items making up these tests are somewhat challenging relative to the tested population.

Appendix B, Conditional Standard Error of Measurement, includes scale score by scale score CSEM 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. SEM 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, Reliability Coefficients.

3.4 Reliability of Achievement Classification

When students complete the Florida Statewide Assessments, 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 Florida Statewide Assessments standard-setting process.

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

3.4.1 Classification Accuracy

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 two research questions:

1. What is the overall classification accuracy index 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 two research questions:

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

In the first approach, we used students from the spring 2022 Florida Statewide Assessments 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 the IRT-based approach did not include accommodated tests because the sample was too small to compute classification accuracy.

Table 13 provides the sample size, mean, and standard deviation of the observed theta for the data used in the first method described earlier. The theta scores are based on the maximum likelihood estimates (MLEs) obtained from CAI's scoring engine. Table 14 provides the sample size, mean, and standard deviation of the observed theta for the data used in the second method.

Table 13: Descriptive Statistics from Population Data (ELA, Mathematics, Science, and EOC)

ELA				Mathematics				Science and NGSSS EOC			
Grade	N	Average Theta	SD of Theta	Grade	N	Average Theta	SD of Theta	Subject/Core	N	Average Theta	SD of Theta
3	210,199	–0.06	1.12	3	207,359	–0.05	1.17	Science 5	211,713	–0.13	1.21
4	198,365	0.03	1.17	4	194,759	0.01	1.20	Science 8	192,415	–0.11	1.15
5	212,286	0.00	1.19	5	208,616	–0.17	1.25	Bio1 Core 100	74,827	–0.04	1.25
6	196,852	–0.04	1.17	6	183,157	–0.20	1.20	Bio1 Core 200	50,833	0.08	1.21
7	205,856	–0.15	1.19	7	162,822	–0.28	1.18	Bio1 Core 300	80,230	0.08	1.20
8	211,677	–0.22	1.21	8	123,276	–0.42	1.22	Civ Core 100	81,139	0.13	1.23
9	207,409	–0.06	1.15	Alg1 Core 24	65,433	–0.27	1.28	Civ Core 200	64,293	0.34	1.17
10	200,105	–0.09	1.19	Alg1 Core 25	47,570	–0.11	1.25	Civ Core 300	64,087	0.33	1.17
				Alg1 Core 26	47,710	–0.13	1.25	USH Core 100	70,776	0.22	1.18
				Alg1 Core 27	47,694	–0.11	1.24	USH Core 200	53,729	0.28	1.18
				Geo Core 19	76,118	–0.25	1.29	USH Core 300	53,809	0.30	1.18
				Geo Core 20	55,879	–0.15	1.25				
				Geo Core 21	55,882	–0.15	1.24				

* Alg1: Algebra; Geo: Geometry; Bio1: Biology 1; Civ: Civics; USH: U.S. History

Table 14: Descriptive Statistics from Calibration Data (ELA, Mathematics, Science, and EOC)

ELA				Mathematics				Science and NGSSS EOC			
Grade	N	Average Theta	SD of Theta	Grade	N	Average Theta	SD of Theta	Subject/Core	N	Average Theta	SD of Theta
3	28,209	–0.04	1.12	3	27,797	–0.01	1.15	Science 5	210,913	–0.13	1.21
4	26,481	0.03	1.18	4	26,050	0.03	1.19	Science 8	189,882	–0.10	1.15
5	27,973	0.00	1.20	5	28,209	–0.11	1.25	Bio1 Core 100	72,105	–0.03	1.25
6	26,239	–0.01	1.21	6	24,553	–0.15	1.21	Bio1 Core 200	49,015	0.08	1.21
7	24,969	–0.11	1.19	7	25,011	–0.17	1.16	Bio1 Core 300	77,194	0.08	1.20
8	26,200	–0.16	1.22	8	109,277	–0.39	1.21	Civ Core 100	78,904	0.14	1.23
9	24,968	–0.04	1.18	Alg1 Core 24	57,993	–0.29	1.27	Civ Core 200	62,557	0.35	1.17
10	25,087	–0.04	1.18	Alg1 Core 25	42,070	–0.12	1.24	Civ Core 300	62,394	0.34	1.17
				Alg1 Core 26	42,187	–0.14	1.24	USH Core 100	68,546	0.22	1.18
				Alg1 Core 27	42,197	–0.12	1.23	USH Core 200	52,098	0.29	1.18
				Geo Core 19	67,436	–0.24	1.27	USH Core 300	52,181	0.31	1.17
				Geo Core 20	49,365	–0.14	1.23				
				Geo Core 21	49,423	–0.15	1.23				

* Alg1: Algebra; Geo: Geometry; Bio1: Biology 1; Civ: Civics; USH: U.S. History

The observed score approach (Rudner, 2001, 2005) 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:

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

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 SEM of $\hat{\sigma}_i$ and using the asymptotic property of normality of the MLE, $\hat{\theta}_i$, 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:

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

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 2022 Florida Statewide Assessments 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:

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

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

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

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

Table 15 through Table 18 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 0.786 to around 0.802 for Mathematics, 0.752 to 0.775 for ELA, 0.727 to 0.779 for EOC, and is 0.752 for both Science grades.

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 15: Classification Accuracy Index (Mathematics)

Grade	Overall Accuracy Index	Cut Accuracy Index			
		Between Cut 1 and Cut 2	Between Cut 2 and Cut 3	Between Cut 3 and Cut 4	Between Cut 4 and Cut 5
3	0.788	0.956	0.940	0.934	0.958
4	0.802	0.954	0.947	0.944	0.956
5	0.798	0.944	0.942	0.948	0.964
6	0.786	0.923	0.932	0.948	0.974
7	0.791	0.915	0.932	0.959	0.981
8	0.789	0.911	0.927	0.964	0.986

Table 16: Classification Accuracy Index (ELA)

Grade	Overall Accuracy Index	Cut Accuracy Index			
		Between Cut 1 and Cut 2	Between Cut 2 and Cut 3	Between Cut 3 and Cut 4	Between Cut 4 and Cut 5
3	0.775	0.935	0.933	0.937	0.968
4	0.759	0.950	0.932	0.925	0.951
5	0.759	0.951	0.928	0.925	0.954
6	0.752	0.939	0.926	0.929	0.957
7	0.766	0.940	0.931	0.935	0.959
8	0.765	0.937	0.932	0.937	0.958
9	0.769	0.945	0.932	0.933	0.958
10	0.763	0.944	0.926	0.932	0.960

Table 17: Classification Accuracy Index (EOC)

Subject/Core	Overall Accuracy Index	Cut Accuracy Index			
		Between Cut 1 and Cut 2	Between Cut 2 and Cut 3	Between Cut 3 and Cut 4	Between Cut 4 and Cut 5
Algebra 1/Core 24	0.778	0.909	0.920	0.954	0.974
Algebra 1/Core 25	0.770	0.914	0.919	0.948	0.968
Algebra 1/Core 26	0.774	0.912	0.920	0.951	0.970
Algebra 1/Core 27	0.779	0.923	0.925	0.948	0.970
Geometry/Core 19	0.775	0.907	0.920	0.959	0.974
Geometry/Core 20	0.773	0.908	0.920	0.956	0.973
Geometry/Core 21	0.768	0.905	0.916	0.956	0.973
Biology 1/Core 100	0.744	0.917	0.918	0.942	0.952
Biology 1/Core 200	0.748	0.935	0.921	0.936	0.948
Biology 1/Core 300	0.755	0.935	0.925	0.939	0.951
Civics/Core 100	0.738	0.930	0.926	0.931	0.944
Civics/Core 200	0.740	0.947	0.931	0.923	0.934
Civics/Core 300	0.748	0.945	0.933	0.927	0.938
U.S. History/Core 100	0.734	0.929	0.927	0.931	0.941
U.S. History/Core 200	0.734	0.938	0.928	0.926	0.939
U.S. History/Core 300	0.727	0.933	0.926	0.924	0.936

Table 18: Classification Accuracy Index (Science)

Grade/Subject	Overall Accuracy Index	Cut Accuracy Index			
		Between Cut 1 and Cut 2	Between Cut 2 and Cut 3	Between Cut 3 and Cut 4	Between Cut 4 and Cut 5
5	0.752	0.942	0.926	0.930	0.947
8	0.752	0.942	0.924	0.930	0.947

Table 19 through Table 22 provide the FPR and FNR from the IRT-based approach for Mathematics, ELA, EOC, and Science. The FNR and FPR rates for the level 2/3 cut are around 3% to 4% for Mathematics, ELA, EOC, and Science.

Table 19 through Table 22 also provide the overall accuracy rates after accounting for both false positive and false negative rates. For example, the overall accuracy rate of 0.938 for the Level 2/3 cut in Grade 3 Mathematics suggests 93.8% 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.

Table 19: 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	0.023	0.020	0.957	0.030	0.032	0.938	0.029	0.038	0.933	0.016	0.028	0.955
4	0.023	0.021	0.956	0.026	0.027	0.947	0.025	0.032	0.943	0.018	0.028	0.954
5	0.030	0.024	0.946	0.027	0.030	0.943	0.023	0.030	0.947	0.016	0.022	0.962
6	0.036	0.028	0.936	0.029	0.033	0.938	0.021	0.028	0.951	0.012	0.016	0.972
7	0.050	0.034	0.916	0.033	0.033	0.934	0.018	0.024	0.958	0.009	0.013	0.978
8	0.053	0.039	0.908	0.034	0.038	0.928	0.016	0.019	0.965	0.006	0.008	0.986

Table 20: 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	0.034	0.026	0.939	0.030	0.036	0.934	0.027	0.037	0.936	0.013	0.019	0.968
4	0.027	0.024	0.949	0.032	0.037	0.932	0.032	0.044	0.924	0.018	0.032	0.950
5	0.028	0.022	0.951	0.033	0.039	0.927	0.032	0.047	0.922	0.016	0.030	0.954
6	0.035	0.028	0.937	0.032	0.039	0.928	0.029	0.043	0.928	0.017	0.028	0.954
7	0.032	0.028	0.940	0.033	0.038	0.930	0.030	0.037	0.933	0.017	0.027	0.956
8	0.034	0.028	0.938	0.034	0.035	0.931	0.028	0.037	0.936	0.017	0.028	0.954
9	0.028	0.026	0.946	0.030	0.036	0.934	0.028	0.038	0.933	0.017	0.026	0.957
10	0.031	0.026	0.943	0.034	0.041	0.924	0.031	0.041	0.928	0.015	0.028	0.957

Table 21: False Classification Rates and Overall Accuracy Rates (EOC)

	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Subject/Core	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
Algebra 1/Core24	0.048	0.037	0.915	0.036	0.037	0.927	0.018	0.023	0.958	0.010	0.014	0.976
Algebra 1/Core25	0.045	0.038	0.917	0.035	0.039	0.926	0.020	0.028	0.951	0.012	0.017	0.971
Algebra 1/Core26	0.045	0.038	0.917	0.033	0.037	0.930	0.020	0.026	0.955	0.012	0.016	0.972
Algebra 1/Core27	0.041	0.035	0.924	0.035	0.038	0.927	0.021	0.029	0.950	0.012	0.016	0.972
Geometry/Core 19	0.048	0.039	0.913	0.036	0.037	0.927	0.016	0.022	0.962	0.010	0.014	0.977
Geometry/Core 20	0.047	0.038	0.915	0.035	0.038	0.927	0.017	0.023	0.960	0.010	0.015	0.975
Geometry/Core 21	0.051	0.040	0.909	0.035	0.040	0.924	0.017	0.023	0.960	0.010	0.015	0.975
Biology 1/Core 100	0.062	0.031	0.907	0.035	0.036	0.929	0.022	0.035	0.944	0.017	0.030	0.953
Biology 1/Core 200	0.050	0.024	0.926	0.036	0.040	0.924	0.025	0.040	0.935	0.019	0.033	0.947
Biology 1/Core 300	0.049	0.024	0.927	0.035	0.038	0.927	0.024	0.037	0.938	0.018	0.031	0.951

	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Subject/Core	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
Civics/Core 100	0.045	0.028	0.927	0.034	0.037	0.929	0.028	0.041	0.931	0.021	0.037	0.943
Civics/Core 200	0.033	0.022	0.945	0.031	0.038	0.931	0.031	0.049	0.921	0.024	0.043	0.934
Civics/Core 300	0.036	0.024	0.941	0.030	0.036	0.934	0.029	0.045	0.926	0.022	0.040	0.937
U.S. History/Core 100	0.043	0.029	0.928	0.034	0.038	0.928	0.026	0.045	0.929	0.021	0.038	0.941
U.S. History/Core 200	0.037	0.026	0.937	0.033	0.039	0.928	0.029	0.047	0.924	0.021	0.040	0.938
U.S. History/Core 300	0.038	0.027	0.934	0.032	0.039	0.929	0.029	0.050	0.922	0.022	0.043	0.935

Table 22: False Classification Rates and Overall Accuracy Rates (Science)

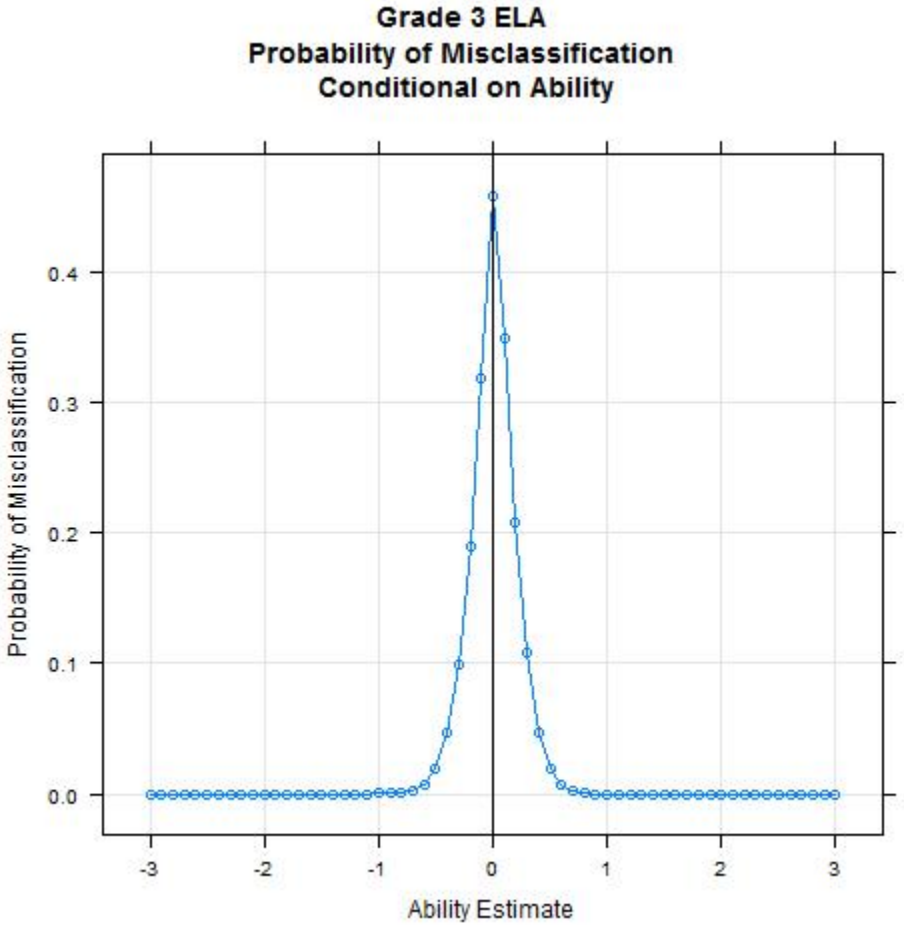
	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Grade	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**	**FPR**	**FNR**	**Accuracy**
5	0.032	0.025	0.943	0.036	0.039	0.926	0.030	0.043	0.927	0.019	0.037	0.944
8	0.030	0.026	0.944	0.030	0.036	0.934	0.026	0.035	0.939	0.018	0.028	0.954

Figure 6 shows a plot exhibiting the probability of misclassification for Grade 3 ELA. The plot displays that students with scores below –0.308 on the theta scale, which corresponds to a scale score of 294, and students with scores above 0.325, corresponding to a scale score of 306, 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 (English Language Learners [ELLs] and Students with Disabilities [SWDs]). The plots of the misclassification probabilities for the Level 1/2 cuts are also included Appendix C for Grade 3 ELA. 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 6: 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 15 through Table 18) 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 Florida Statewide Assessments classification accuracy rates to those of the State of New York, which is comparable in population size (New York State Education Department, 2021). Although New York administers a different testing program, estimated accuracy rates there range from 60% to 70% in ELA and from 66% to 75% in Mathematics (2021). The individual cut accuracy was relatively similar between New York and Florida. Florida showed from 93% to 95% in Mathematics, from 92% to 93% in ELA, from 92% to 93% in EOC, and 93% in Science for the Level 2/3 cut. New York showed from 84% to 88% in ELA and from 86% to 93% in Mathematics for the proficiency cut. The 2019 classification accuracy for the New York EOC tests at the 2/3 cut showed 92% in ELA and 92% to 93% in Mathematics (Algebra 1 and Geometry).

3.4.2 Classification Consistency

Classification accuracy refers to the degree to which a student's true score and observed score would fall within the same performance level (Rudner, 2001). Classification consistency refers to the degree to which test takers are classified into the same performance level assuming the test is administered twice independently (Lee, Hanson, and Brennan, 2002)—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms. In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution. Classification consistency was estimated based on the method in Lee, Hanson, and Brennan (2002).

Similar to accuracy, a 5×5 matrix can be constructed by assuming the test is administered twice independently to the same group of students. The classification consistency index for the individual cuts (CCIC) was estimated as:

$$CCIC = \frac{\sum_{i=1}^{N}(\rho_i(\theta > c)^2 + (1 - \rho_i(\theta > c))^2)}{N}$$

Where c is the cut score required for passing in the same assigned metric, ρ is the probability of being above the cut for student i, N is the total number of students, and θ is true ability in the true-score metric.

Classification consistency with classification accuracy results are presented in Table 23 through Table 26. In cut 1 and cut 2, cut 2 and cut 3, and cut 3 and cut 4 results, all accuracy values are higher than 0.90, and consistency values are around 0.90 or slightly below 0.90. With the higher performance levels, cut 4 and cut 5, most values are around 0.95 or slightly below 0.95. In all performance levels, classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy because the consistency is based on two tests with measurement errors, while the accuracy is based on one test with a measurement error and the true score. The accuracy and consistency rates for each performance level are higher for the levels with smaller standard error.

Table 23. Classification Accuracy and Consistency (Cut 1 and Cut 2)

Grade	ELA		Grade/ Subject	Mathematics		Grade/ Subject	Science and NGSSS EOC	
	Accuracy	Consistency		Accuracy	Consistency		Accuracy	Consistency
3	0.935	0.914	3	0.956	0.940	Science 5	0.942	0.920
4	0.950	0.928	4	0.954	0.937	Science 8	0.942	0.922
5	0.951	0.931	5	0.944	0.924	Bio1_core100	0.917	0.883
6	0.939	0.913	6	0.923	0.910	Bio1_core200	0.935	0.907
7	0.940	0.916	7	0.915	0.883	Bio1_core300	0.935	0.907
8	0.937	0.913	8	0.911	0.871	Civ_core100	0.930	0.900
9	0.945	0.924	Alg1_core24	0.909	0.879	Civ_core200	0.947	0.925
10	0.944	0.920	Alg1_core25	0.914	0.882	Civ_core300	0.945	0.920

Grade	ELA		Grade/ Subject	Mathematics		Grade/ Subject	Science and NGSSS EOC	
	Accuracy	Consistency		Accuracy	Consistency		Accuracy	Consistency
			Alg1_core26	0.912	0.882	USH_core100	0.929	0.901
			Alg1_core27	0.923	0.894	USH_core200	0.938	0.913
			Geo_core19	0.907	0.877	USH_core300	0.933	0.910
			Geo_core20	0.908	0.880			
			Geo_core21	0.905	0.872			

Table 24. Classification Accuracy and Consistency (Cut 2 and Cut 3)

Grade	ELA		Grade/ Subject	Mathematics		Grade/ Subject	Science and NGSSS EOC	
	Accuracy	Consistency		Accuracy	Consistency		Accuracy	Consistency
3	0.933	0.906	3	0.940	0.913	Science 5	0.926	0.895
4	0.932	0.903	4	0.947	0.925	Science 8	0.924	0.907
5	0.928	0.897	5	0.942	0.919	Bio1_core100	0.918	0.898
6	0.926	0.899	6	0.932	0.913	Bio1_core200	0.921	0.893
7	0.931	0.901	7	0.932	0.907	Bio1_core300	0.925	0.897
8	0.932	0.903	8	0.927	0.899	Civ_core100	0.926	0.900
9	0.932	0.906	Alg1_core24	0.920	0.897	Civ_core200	0.931	0.902
10	0.926	0.893	Alg1_core25	0.919	0.895	Civ_core300	0.933	0.906
			Alg1_core26	0.920	0.901	USH_core100	0.927	0.898
			Alg1_core27	0.925	0.897	USH_core200	0.928	0.898
			Geo_core19	0.920	0.897	USH_core300	0.926	0.900
			Geo_core20	0.920	0.897			
			Geo_core21	0.916	0.894			

Table 25. Classification Accuracy and Consistency (Cut 3 and Cut 4)

Grade	ELA		Grade/ Subject	Mathematics		Grade/ Subject	Science and NGSSS EOC	
	Accuracy	Consistency		Accuracy	Consistency		Accuracy	Consistency
3	0.937	0.911	3	0.934	0.906	Science 5	0.930	0.899
4	0.925	0.893	4	0.944	0.920	Science 8	0.930	0.915
5	0.925	0.892	5	0.948	0.925	Bio1_core100	0.942	0.922
6	0.929	0.901	6	0.948	0.931	Bio1_core200	0.936	0.911
7	0.935	0.907	7	0.959	0.942	Bio1_core300	0.939	0.915
8	0.937	0.910	8	0.964	0.952	Civ_core100	0.931	0.904
9	0.933	0.906	Alg1_core24	0.954	0.942	Civ_core200	0.923	0.890
10	0.932	0.900	Alg1_core25	0.948	0.933	Civ_core300	0.927	0.898
			Alg1_core26	0.951	0.937	USH_core100	0.931	0.902
			Alg1_core27	0.948	0.931	USH_core200	0.926	0.895
			Geo_core19	0.959	0.947	USH_core300	0.924	0.892
			Geo_core20	0.956	0.945			
			Geo_core21	0.956	0.944			

Table 26. Classification Accuracy and Consistency (Cut 4 and Cut 5)

Grade	ELA		Grade/ Subject	Mathematics		Grade/ Subject	Science and NGSSS EOC	
	Accuracy	Consistency		Accuracy	Consistency		Accuracy	Consistency
3	0.968	0.958	3	0.958	0.941	Science 5	0.947	0.927
4	0.951	0.936	4	0.956	0.937	Science 8	0.947	0.938
5	0.954	0.940	5	0.964	0.948	Bio1_core100	0.952	0.936
6	0.957	0.940	6	0.974	0.962	Bio1_core200	0.948	0.929
7	0.959	0.942	7	0.981	0.970	Bio1_core300	0.951	0.933
8	0.958	0.940	8	0.986	0.981	Civ_core100	0.944	0.922
9	0.958	0.941	Alg1_core24	0.974	0.967	Civ_core200	0.934	0.910
10	0.960	0.944	Alg1_core25	0.968	0.960	Civ_core300	0.938	0.915
			Alg1_core26	0.970	0.961	USH_core100	0.941	0.920
			Alg1_core27	0.970	0.961	USH_core200	0.939	0.917
			Geo_core19	0.974	0.968	USH_core300	0.936	0.914
			Geo_core20	0.973	0.965			
			Geo_core21	0.973	0.965			

3.5 PRECISION AT CUT SCORES

Table 27 through Table 30 present the mean CSEM at each achievement level by grade and subject. These tables also include achievement level cut scores and associated CSEM. The CSEM at Cut Score is based on TIF curve.

Table 27: 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	6.88		
	2	4.07	285	4
	3	4.16	297	4
	4	5.22	311	5
	5	9.82	327	6
4	1	8.95		
	2	4.48	299	5
	3	4.00	310	4
	4	4.34	325	4
	5	7.79	340	5
5	1	11.14		
	2	5.19	306	6
	3	4.37	320	5
	4	4.74	334	4
	5	7.92	350	5
6	1	17.19		
	2	5.44	310	6
	3	4.33	325	5
	4	4.19	339	4
	5	6.44	356	5
7	1	12.42		
	2	5.91	316	7
	3	4.24	330	5
	4	3.99	346	4
	5	4.98	360	4
8	1	12.27		
	2	6.09	322	7
	3	4.94	337	5
	4	4.74	353	5
	5	5.64	365	5

Table 28: 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	12.20		
	2	5.12	285	6
	3	4.42	300	5
	4	5.03	315	5
	5	7.50	330	6
4	1	8.68		
	2	5.06	297	6
	3	5.00	311	5
	4	5.89	325	6
	5	7.92	340	7
5	1	8.01		
	2	5.67	304	6
	3	5.83	321	6
	4	6.42	336	6
	5	8.34	352	8
6	1	9.59		
	2	6.41	309	8
	3	5.75	326	6
	4	5.95	339	6
	5	7.33	356	7
7	1	9.57		
	2	5.55	318	7
	3	5.04	333	6
	4	5.51	346	6
	5	7.51	360	7
8	1	9.34		
	2	5.80	322	7
	3	5.22	337	6
	4	5.97	352	6
	5	7.90	366	7
9	1	9.13		
	2	5.31	328	6
	3	5.00	343	5
	4	5.20	355	5

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
	5	6.92	370	6
10	1	8.30		
	2	5.62	334	7
	3	5.14	350	6
	4	5.94	362	6
	5	7.73	378	7

Table 29: Achievement Levels and Associated Conditional Standard Error of Measurement (EOC)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Alg1	1	19.76		
	2	7.64	487	8
	3	5.65	497	7
	4	4.76	518	5
	5	5.76	532	5
Geo	1	21.95		
	2	7.84	486	9
	3	5.40	499	7
	4	4.04	521	4
	5	5.45	533	4
Bio1	1	24.97		
	2	9.94	369	15
	3	6.48	395	7
	4	6.02	421	6
	5	7.62	431	6
Civ	1	18.33		
	2	8.81	376	11
	3	6.79	394	7
	4	6.13	413	6
	5	8.33	428	6
USH	1	18.66		
	2	8.61	378	10
	3	6.99	397	8
	4	6.57	417	6
	5	8.28	432	7

Table 30: Achievement Levels and Associated Conditional Standard Error of Measurement (Science)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	7.59		
	2	5.03	185	5
	3	5.81	200	5
	4	7.00	215	6
	5	10.63	225	8
8	1	9.13		
	2	4.58	185	5
	3	4.68	203	4
	4	5.51	215	5
	5	9.56	225	6

3.6 Writing Prompts Inter-Rater Reliability

All Grade 10 Writing prompts and 15% of prompts in Grades 4–9 were hand-scored by two human raters. The basic method to compute inter-rater reliability is percentage agreement. As seen in Table 31, the agreement column shows the exact agreement (when two raters gave the same score), the adjacent ratings (when the difference between two raters was 1), and the non-adjacent ratings (when the difference was larger than 1). In this example, responses 2 and 3 had exact agreement, response 1 had adjacent agreement, and response 4 had non-adjacent agreement.

Table 31: Rater 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:

- **Percentage Exact.** Total number of responses by scorer in which scores are equal divided by the number of responses that were scored twice
- **Percentage 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

- **Percentage 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 32 displays rater-agreement percentages. The percentage of exact agreement between two raters ranged from 74% to 88%. The percentage of adjacent rating was between 11% and 25%. The non-adjacent percentages fell between 0% and 1%. The 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 32: Inter-Rater Reliability

Grade	Dimension	% Exact	% Adjacent	% Not Adjacent	Number of Processed Responses with Scores from Two Raters
4	Purpose, Focus, & Organization	77	22	1	84,108
	Evidence & Elaboration	78	21	1	
	Conventions	83	17	0	
5	Purpose, Focus, & Organization	76	23	0	87,144
	Evidence & Elaboration	77	22	1	
	Conventions	83	17	0	
6	Purpose, Focus, & Organization	74	25	1	80,804
	Evidence & Elaboration	74	25	1	
	Conventions	88	11	1	
7	Purpose, Focus, & Organization	78	22	0	86,200
	Evidence & Elaboration	79	21	0	
	Conventions	82	17	0	
8	Purpose, Focus, & Organization	78	22	0	83,492
	Evidence & Elaboration	78	21	0	
	Conventions	84	16	0	
9	Purpose, Focus, & Organization	79	21	0	82,320
	Evidence & Elaboration	79	21	0	
	Conventions	85	14	0	
10	Purpose, Focus, & Organization	77	22	0	420,886
	Evidence & Elaboration	77	22	0	
	Conventions	83	17	0	

In addition to inter-rater reliability, validity coefficients, percentage exact agreement on validity true scores, and human scores were also calculated. Validity true scores for each dimension were determined by scoring directors, and Test Development Center (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:

- **Percentage Exact.** Total number of responses by scorer where scores are equal divided by the total number of responses that were scored
- **Percentage Adjacent.** Total number of responses by scorer where scores are one point apart divided by the total number of responses that were scored
- **Percentage 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 33 presents final validity coefficients, which were between 77 and 91.

Table 33: Validity Coefficients

Grade	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
4	84	84	85
5	86	87	77
6	78	78	91
7	88	87	88
8	87	87	87
9	89	89	90
10	89	88	87

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:

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

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

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

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 34. They ranged from 0.741 to 0.900. Grade 10 was scored by two human scorers, while only 15% of the students in other grades received two scores.

Table 34: Weighted Kappa Coefficients

Grade	N	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
4	40,828	0.885	0.876	0.861
5	42,644	0.883	0.878	0.859
6	38,717	0.849	0.849	0.877
7	42,691	0.900	0.900	0.861
8	41,394	0.890	0.890	0.863
9	40,793	0.895	0.893	0.871
10	209,385	0.774	0.772	0.741

4. VALIDITY

Validation is the process of collecting evidence to support inferences from assessment results. A prime consideration in validating a test is determining if the test measures what it purports to measure. During the process of evaluating if the test measures the construct of interest, a number of threats to validity must be considered. For example, the test may be biased against a particular group, test scores may be unreliable, students may not be properly motivated to perform on the test, or test content may not span the entire range of the construct to be measured. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring that the test is measuring what it is supposed to measure, it is equally important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Volume 6 (see sections of Appropriate Score Uses and Cautions for Score Use) and Volume 1 (see Scoring section) of this technical report.

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity that has been the subject of considerable research, thought, and debate in the measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result the field has evolved.

This chapter begins with an overview of the major historical perspectives on validity in measurement. Included in this overview is a presentation of a modern perspective that takes an argument-based approach to validity. Following the overview is the presentation of validity evidence for the Florida Statewide Assessments.

4.1 PERSPECTIVES ON TEST VALIDITY

The following sections discuss some of the major conceptualizations of validity used in educational measurement.

4.1.1 Criterion Validity

The basis of criterion validity is the demonstration of a relationship between the test and an external criterion. If the test is intended to measure mathematical ability, for example, then scores from the test should correlate substantially with other valid measures of mathematical ability. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment tasks and the outcome criterion (Cronbach, 1990). In order for the observed relationship between the assessment and the criterion to be a meaningful indicator of criterion validity, the criterion should be relevant to the assessment and be reliable. Criterion validity is typically expressed in terms of the product-moment correlation between the scores of the test and the criterion score.

There are two types of criterion-related evidence: concurrent and predictive. The difference between these types lies in the procedures used for collecting validity evidence. Concurrent

evidence is collected from both the assessment and the criterion at the same time. An example might be found in relating the scores from a district-wide assessment to the American College Testing (ACT) assessment (the criterion). In this example, if the results from the district-wide assessment and the ACT assessment were collected in the same semester of the school year, this would provide concurrent criterion-related evidence. On the other hand, predictive evidence is usually collected at different times; typically, the criterion information is obtained subsequent to the administration of the measure. For example, if ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school, whereas the criterion (e.g., college grade point average) would not be available until the following year.

In ideal situations, the criterion validity approach can provide convincing evidence of a test's validity. However, there are two important obstacles to implementing the approach. First, a suitable criterion must be found. Standards-based tests like the Florida Statewide Assessments are designed to measure student achievement on Florida Statewide Assessments. Finding a criterion representing achievement on the standards may be difficult to do without creating yet another test. It is possible to correlate performance on the Florida Statewide Assessments with other types of assessments, such as the ACT or school assessments. Strong correlations with a variety of other assessments would provide some evidence of validity for the Florida Statewide Assessments, but the evidence would be less compelling if the criterion measures are only indirectly related to the standards.

A second obstacle to the demonstration of criterion validity is that the criterion may need to be validated as well. In some cases, it may be more difficult to demonstrate the validity of the criterion than to validate the test itself. Further, unreliability of the criterion can substantially attenuate the correlation observed between a valid measure and the criterion.

Criterion-related validity evidence on the Florida Statewide Assessments will be collected and reported in an ongoing manner. These data are most likely to come from districts conducting program evaluation research, university researchers and special interest groups researching topics of local interest, as well as the data collection efforts of FDOE.

4.1.2 Content and Curricular Validity

Content validity is a type of test validity that addresses whether the test adequately samples the relevant domain of material it purports to cover (Cronbach, 1990). If a test is made up of a series of tasks that form a representative sample of a particular domain of tasks, then the test is said to have good content validity. For example, a content-valid test of mathematical ability should be composed of tasks allowing students to demonstrate their mathematical ability.

Evaluating content validity is a subjective process based on rational arguments. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity only speaks to the validity of the test itself, not to decisions made based on the test scores. For example, a poor score on a content-valid mathematics test indicates that the student did not demonstrate mathematical ability. But from this alone, one cannot conclusively determine that the student has low mathematical ability. This conclusion could only be reached if it could be shown or argued that the student put forth his or her best effort, the student was not distracted during the test, and the test did not contain a bias preventing the student from scoring well.

Generally, achievement tests such as the Florida Statewide Assessments are constructed so that they have strong content validity. As documented in this volume as well as in Volume 2, tremendous effort is expended by FDOE, the content vendor (CAI and Pearson), and the educator committees to ensure the Florida Statewide Assessments are content-valid. Although content validity has limitations and cannot serve as the only evidence for validation, it is an important piece of evidence for the validation of the Florida Statewide Assessments.

4.1.3 Construct Validity

The term construct validity refers to the degree to which the observed test score is a measure of the underlying characteristic (i.e., the latent construct) of interest. A construct is an individual characteristic assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic is inferred from an assessment result, a generalization or interpretation in terms of a construct is being made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are "good problem-solvers" implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate this is a reasonable and valid use of the results.

Messick (1989) describes construct validity as a "unifying force" in that inferences based on criterion evidence or content evidence can also be framed by the theory of the underlying construct. From this point of view, validating a test is essentially the equivalent of validating a scientific theory. As Cronbach and Meehl (1955) first argued, conducting construct validation requires a theoretical network of relationships involving the test score. Validation not only requires evidence supporting the notion that the test measures the theoretical construct, but it further requires evidence be presented that discredits every plausible alternative hypothesis as well. Because theories can only be supported or falsified, but never proven, validating a test becomes a never-ending process.

Construct-related validity evidence can come from many sources. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) provides the following list of possible sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain, or construct
- Substantial relationships between the assessment results and other measures of the same defined construct
- Little or no relationship between the assessment results and other measures that are clearly not of the defined construct
- Substantial relationships between different methods of measurement regarding the same defined construct
- Relationships to non-assessment measures of the same defined construct

One source of validity evidence suggested by *The Standards* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on

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

Kane (2006) states that construct validity is now widely viewed as a general and all-encompassing approach to accessing test validity. However, in Kane's view there are limitations of the construct validity approach, including the need for strong measurement theories and the general lack of guidance on how to conduct a validity assessment.

4.2 VALIDITY ARGUMENT EVIDENCE FOR THE FLORIDA ASSESSMENTS

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, and NCME, 2014, p.11). Messick (1989, p.13) 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, and NCME, 2014) suggests 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.

4.2.1 Test Purpose

The primary purpose of the Florida Statewide Assessments program is to measure students' achievement of Florida's education standards and classify students into the appropriate achievement levels based on their test scores. Assessment supports instruction and student learning. Assessment results help Florida's educational leadership and stakeholders determine whether the goals of the education system are being met. Assessments help Florida determine whether we have equipped our students with the knowledge and skills they need to be ready for careers and college-level coursework. Florida's educational assessments also provide the basis for student, school, and district accountability systems.

Assessment results are used to determine school and district grades which give citizens a standard way to determine the quality and progress of Florida's education system. While assessment plays a key role in Florida's education system, it is important to remember that testing is not an end in itself, but a means to an end. Florida's assessment and accountability efforts have had a significant positive impact on student achievement over time. Readers can refer to Table 1 in Volume 1 of the *Florida Statewide Assessments 2021–2022 Technical Report* to see the specific required uses and citations for Florida Statewide Assessments.

For the Florida Statewide Assessments program, an argument-based approach to validity (Kane, 2006) is used to ensure that the combined evidence about its assessment system is comprehensive and addresses critical features of the assessments that relate to score interpretations and uses. The

primary claims in Florida Statewide Assessments are represented in the statements below as they relate logically:

- Assessment scores provide a snapshot of information that reflects what students know and can do in relation to academic expectations.
- Students' ability is consistent with the achievement level they are classified into.

Therefore, the following occurs:

- Assessment scores provide information that is helpful for Florida's educational leadership and stakeholders to determine whether the goals of the education system are being met.
- Assessment scores provide information that is helpful for Florida to determine whether it has equipped its students with the knowledge and skills they need to be ready for careers and college-level coursework.
- Assessment scores provide the basis for student, school, and district accountability systems.

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.

The sections below present a summary of the validity argument evidence for the four parts of the interpretive argument: scoring, generalization, extrapolation, and implication. Much of this evidence is presented in greater detail in other volumes in this report. In fact, most of this report can be considered validity evidence for Florida Statewide Assessments. Volume 1: *Annual Technical Report* provides validity evidence on calibration, equating, scaling, scoring, and quality control. Volume 2: *Test Development* provides validity evidence on test specifications, item development, and test construction. Volume 4: *Evidence of Reliability and Validity* provides validity evidence on reliability, content validity, internal structure validity, comparability, and test fairness. Volume 5: *Test Administration* documents evidence on the validity of testing procedures (e.g., standardization of test administration and accommodations) as well as test security procedures. Volume 6: *Score Interpretation Guide* provides validity evidence on the guidance provided to facilitate appropriate interpretation of test scores. Please note that Volume 3 of the *Florida Standards Assessments 2014–2015 Technical Report*: *Setting Achievement Standards* provides evidence on the validity of the process and the results of setting performance standards for Mathematics, ELA, Algebra 1, and Geometry. Similarly, Chapter 5 of the *Florida Statewide Science and EOC Assessments 2019 Technical Report: Performance Standards* provides details on the procedures and results of the standard setting for Grades 5 and 8 Science, Biology 1, Civics, and U.S. History.

Table 35 provides the comprehensive summary of validity evidence in terms of the interpretive argument. The subsequent sections elaborate on this evidence. Relevant volumes or sections in volumes are cited as part of the validity evidence given in Table 35 and the sections that follow.

Table 35: Comprehensive Summary of Validity Evidence

Inferences	Claims	Evidence	Location
Scoring: Students are scored accurately and consistently.	**Model Fit.** The underlying assumptions of the IRT models are met. The assessments are essentially unidimensional.	o Local independence. o Confirmatory factor analysis and correlations among latent factors of reporting categories. o Item-total correlational analysis.	o Volume 4, Section 4.2.2
	Scoring of Performance Tasks. The inter-rater reliability is reasonably high.	o Validity responses are dealt by ScoreBoard throughout the scoring day. o The validity pool includes responses for each possible score point within each domain and will be refreshed as needed to ensure an adequate quantity. The Validity Score Point Distribution Report is run to ensure that the overall score point distribution of the loaded validity reflects the item score point distribution. o Scoring Directors propose and the FDOE reviews and approves all possible validity responses and monitors reports daily to ensure the meaningfulness of the validity statistics. o Inter-rater agreement. o Inter-rater reliability.	o Section 6.2, 2022 Writing Spring and Fall Handscoring Specifications* o Volume 4, Section 3.6
Generalization: The items that students were administered are representative samples of expected performance in the state standards.	**Test Content.** The State's assessments measure the knowledge and skills specified in the State's academic content standards including alignment with academic content standards.	o Content standards, test specifications, test development o Alignment study o Detailed blueprints for each content level by each grade and subject	o Volume 2, Test Development o Volume 2, 2.1.1, Target Blueprints and Volume 4, 4.1.2 o Volume 4, 4.2.2, Alignment study; Appendix D of the *2015–2016 FSA Technical Report* (Appendix G of this volume)
	Validity Related to Cognitive Process. The State's assessments tap the intended cognitive processes appropriate for each grade level as represented in the State's academic content standards.	o Percentages of items by Depth of Knowledge (DOK) levels for each grade and subject o Cognitive lab report.	o Volume 2, 2.1.1, Target Blueprints and Volume 4, 4.1.2 o Volume 4, 4.2.3, Cognitive Laboratories
	Validity Based on Relations to Other Variables. The State has documented adequate validity evidence that the State's assessment scores are related as expected with other variables.	o College-readiness standards map to the NAEP proficient.	o Volume 7 of *2014–2015 Technical Report,* National Benchmarks for State Achievement Standards (Appendix H of this volume)

Inferences	Claims	Evidence	Location
	Test Administration. Implementation of policies and procedures for standardized test administration: • Clear, thorough, and consistent standardized procedures • Training for all individuals responsible for administering the State's assessments • Clearly defined technology and other related requirements for test administration and contingency plans to address possible technology challenges during test administration	o Test development o Test administration o Monitoring of test accommodations	o Volume 2, Test Development o Volume 5, Test Administration o Volume 4, Chapter 4, Validity
	Measurement Error. The measurement error is sufficiently small given the decisions made with the scores.	o CSEM plots o Cronbach Alpha reliability	o Volume 4, 3.3 CSEM o Volume 4, 3.1 Internal Consistency
	Different Student Populations. Scores represent students in schools throughout Florida including participation from Home Education Program students, students with disabilities, ELL students, McKay Scholarship Program students, etc.	o Testing accommodation o Subgroup reliability	o Volume 5, 1.2 o Volume 4, Appendix A, Reliability Coefficients
Extrapolation (Analytic): The achievement level denotes the proficiency required to be on track for college or career readiness across all students.	**Accommodations.** Appropriate accommodations for SWD under IDEA, students covered by Section 504, and ELL.	o List of available accommodations	o Volume 5, 1.2, Testing Accommodations and Appendix B
	Test Administration for Special Populations. Appropriate assessments, with or without appropriate accommodations, are selected for students with disabilities under IDEA, students covered by Section 504, and ELL.	o Description of ELL students and Students with Disabilities o Description of available testing accommodations and practice activities	o Volume 5, 1.1, Eligible Students o Volume 5, 1.2, Testing Accommodations
	Fairness and Accessibility. Assessments are accessible to all students and fair across student groups in the design, development, and analysis of its assessments.	o A description of fairness and accessibility, based on item statistics and content principles of universal design to minimize the impact of construct-irrelevant factors in assessing student achievement	o Volume 4, 6.1, Fairness in Content and 6.2, Statistical Fairness in Item Statistics
	Device Comparability. There are no meaningful differences in the scores for students when the FSA is administered on different devices and platforms.	o Evidence of the comparability of tests across the most frequently used platforms o Score comparability across different devices	o Volume 4, Chapter 4 Validity o Appendix F of the *2017–2018 FSA Technical Report*: Device Comparability (Appendix F of this volume)

Inferences	Claims	Evidence	Location
	Scoring/Scaling. standardized scoring procedures and protocols for assessments that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State's academic achievement standards.	o Computation of the score: - A description of maximum likelihood estimation - Scale score transformation o Score interpretation guide	o Volume 1, Chapter 8, Scoring and Chapter 6, Scaling o Volume 6, 1.1, Overview of Florida's Score Reports o Volume 6, Chapter 4, Appropriate Score Uses and Chapter 5, Cautions for Score Use
Extrapolation: Empirical	**Internal Structure.** Scoring and reporting structures of assessments are consistent with the sub-domain structures of the State's academic content standards on which the intended interpretations and uses of results are based.	o Correlations among reporting category scores o Goodness-of-fit indices for the second-order CFA model o Correlations among latent factors of reporting categories	o Volume 4, 4.2.2, Scoring Validity Evidence
	C**onvergent and discriminant validity.** Assessment scores are related closely with scores obtained from measures assessing similar constructs and are related less closely with scores obtained from measures assessing different constructs for all student groups.	o Correlations between subscores within and across Mathematics, ELA, and Science	o Volume 4, 4.2.4, Extrapolation of Validity Evidence
Implication: The evidence supports the proposed use of test scores.	**Interpretation of Performance Standards.** The State uses technically sound and well-documented process to develop scoring interpretations and performance standards.	o Standard setting study o Achievement Level Descriptions o Classification accuracy and consistency	o Volume 3, Setting Achievement Standards of *2014–2015 Technical Report* (Appendix E of this volume) o Volume 6, 1.3 Achievement Level Descriptions o Volume 6, Appendix D, Achievement Level Descriptions o Volume 4, 3.4, Classification accuracy and consistency
	Scoring/Scaling. Standardized scoring procedures and protocols for its assessments that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State's academic achievement standards.	o Regarding the computation of the score: - A description of maximum likelihood estimation - Scale score transformation o Score interpretation guide	o Volume 1, Chapter 8, Scoring o Volume 6, 1.1 Overview of Florida's Score Report o Volume 6, Chapter 4, Appropriate Score Use and Chapter 5, Cautions for Score Use

**Confidential dDocument*

4.2.2 Scoring Validity Evidence

Scoring validity evidence can be divided into two sections. These sections are the evidence for the scoring of items and the evidence for the fit of items to the model.

Model Fit and Scaling

Item response theory (IRT) models provide a basis for Florida Statewide Assessments. IRT models are used for the selection of items to go on the test, the equating procedures, and the scaling procedures. A failure of model fit would undermine the validity of these procedures. Item fit is examined during test construction. Any item displaying misfit is scrutinized before a decision is made to place the item on the test. Most items on Florida Statewide Assessments display good model fit.

The Standards (AERA, APA, and NCME, 2014) recommends that the source of validity evidence based on internal structure is the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (see Volume 1, Section 5.2, of this technical report). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis.

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

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

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

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

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

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

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

where y_{jk} is the weight for response category k, m is the number of response categories, and $P_{jk}(\hat{\theta}_i)$ is the probability of response category k to item j by test taker 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 36 through Table 39 present summaries of the distributions of the 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 0.10 for $|Q_3|$ (Chen & Thissen, 1997). This provides evidence that the assumption of local independence is met to a reasonable extent, which lends further support to the uni-dimensionality assumption of the IRT models used for Florida Statewide Assessments.

Table 36: Mathematics Q_3 Statistic

Grade	Unconditional Observed Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.427	–0.102	–0.055	–0.018	0.020	0.423
4	0.462	–0.091	–0.048	–0.019	0.020	0.197
5	0.384	–0.109	–0.049	–0.016	0.017	0.329
6	0.374	–0.078	–0.047	–0.016	0.015	0.165
7	0.361	–0.082	–0.044	–0.015	0.013	0.194
8	0.291	–0.100	–0.046	–0.018	0.022	0.206

Table 37: ELA Q_3 Statistic

Grade	Unconditional Observed Correlation	Q3 Distribution					Within Passage Q_3**	
		Minimum	5th Percentile	Median	95th Percentile	Maximum*	Minimum	Maximum
3	0.318	–0.091	–0.043	–0.019	0.013	0.097	–0.039	0.090
4	0.329	–0.108	–0.062	–0.017	0.018	0.826	–0.020	0.120
5	0.334	–0.112	–0.067	–0.015	0.011	0.863	–0.075	0.141
6	0.273	–0.122	–0.060	–0.013	0.014	0.984	–0.046	0.077
7	0.293	–0.130	–0.066	–0.013	0.017	0.917	–0.034	0.098

Grade	Unconditional Observed Correlation	Q3 Distribution					Within Passage Q_3**	
		Minimum	5th Percentile	Median	95th Percentile	Maximum*	Minimum	Maximum
8	0.313	–0.111	–0.062	–0.014	0.016	0.976	–0.029	0.104
9	0.313	–0.117	–0.063	–0.013	0.012	0.964	–0.030	0.245
10	0.308	–0.116	–0.065	–0.012	0.014	0.972	–0.035	0.072

* Maximum Q_3 values of grades 4 through 10 are from elaboration and organization dimensions of the Writing prompt.
** Within Passage Q_3 values are computed for each item pair within a passage.

Table 38: EOC Q_3 Statistic

Course	Unconditional Observed Correlation*	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Alg1	0.343	–0.106	–0.043	–0.013	0.016	0.173
Geo	0.325	–0.113	–0.039	–0.013	0.013	0.282
Bio1	0.149	–0.073	–0.036	–0.010	0.012	0.119
Civ	0.170	–0.076	–0.039	–0.011	0.011	0.181
USH	0.151	–0.089	–0.035	–0.012	0.009	0.169

* Unconditional observed correlations were computed based on each core form and averaged over those core forms.

Table 39: Science Q_3 Statistic

Grade	Unconditional Observed Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
5	0.179	–0.061	–0.037	–0.011	0.008	0.127
8	0.197	–0.067	–0.032	–0.013	0.007	0.165

Confirmatory Factor Analysis

The Florida Statewide Assessments 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 Florida Statewide Assessments 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 (IRT) 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.

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 than the generalized estimating equations (GEE) approach (Reboussin & Liang, 1998) 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 by comparing two nested models (i.e., the null model and the alternative model). 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 the Florida Statewide Assessments.

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 following development 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, $\widehat{\boldsymbol{\Sigma}}$ 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:

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

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

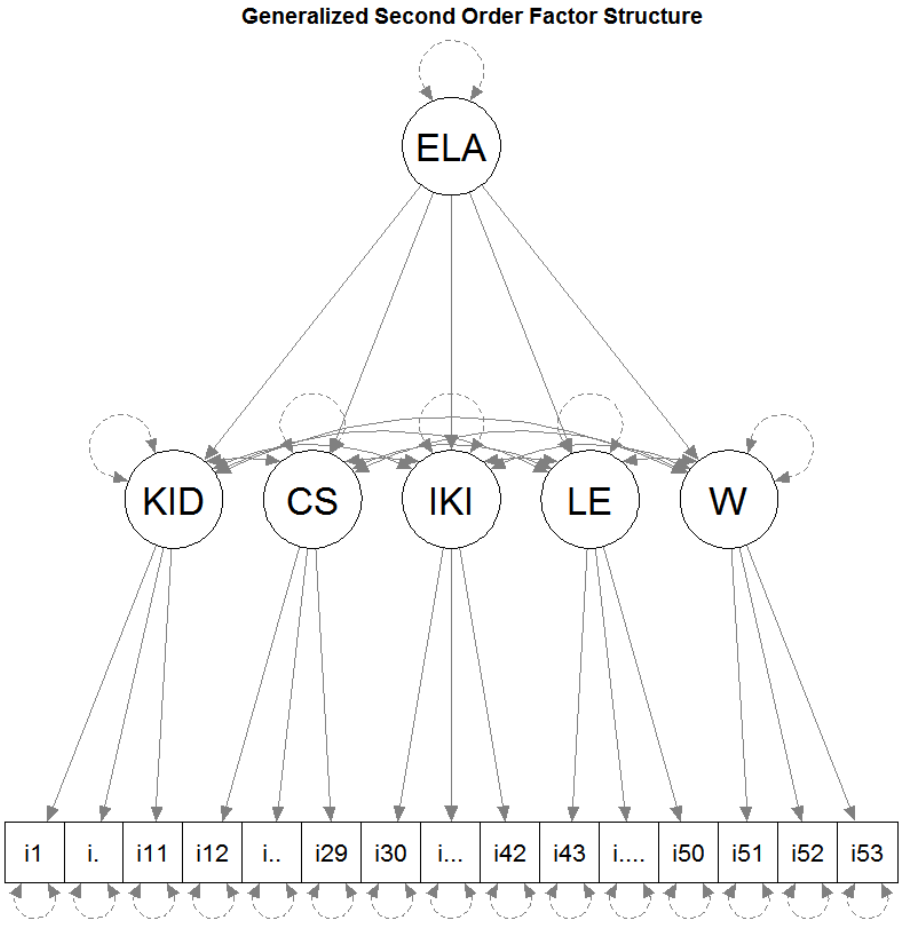
There is a separate factor for each of 3–5 categories for Mathematics, 4–5 reporting categories for ELA, 3–4 categories for EOC, and 4 categories for Science (see Table 45 through Table 48) for reporting category information). Therefore, the number of rows in $\boldsymbol{\Gamma}$ (*k)* differs between subjects, but the general structure of the factor analysis is consistent across subjects.

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 7. 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 Florida Statewide Assessments was to provide evidence that each individual assessment in the Florida Statewide Assessments implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 7: Second-Order Factor Model (ELA)



Results

Several goodness-of-fit statistics from each of the analyses are presented in the following tables. Table 40 presents the summary results obtained from the 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 zero implies better fit and a value of zero 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 0.90 are recognized as acceptable, and values over 0.95 are considered as good fit (Hu & Bentler, 1999). As Hu and Bentler (1999) suggest, the selected cut-off values of the fit index should not be overgeneralized and should be interpreted with caution.

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

Table 40: Goodness-of-Fit Second-Order CFA

ELA					
Grade	**df**	**RMSEA**	**CFI**	**TLI**	**Convergence**
3*	1173	0.02	0.98	0.98	Yes
4	1320	0.02	0.99	0.99	Yes
5*	1322	0.02	0.98	0.98	Yes
6*	1426	0.02	0.98	0.98	Yes
7*	1426	0.02	0.98	0.98	Yes
8*	1426	0.02	0.98	0.98	Yes
9*	1535	0.02	0.98	0.98	Yes
10	1534	0.02	0.99	0.99	Yes
Mathematics					
Grade	**df**	**RMSEA**	**CFI**	**TLI**	**Convergence**
3	1374	0.03	0.97	0.97	Yes
4	1373	0.03	0.98	0.98	Yes
5	1374	0.03	0.97	0.97	Yes
6	1479	0.02	0.98	0.98	Yes
7	1479	0.02	0.98	0.98	Yes
8	1480	0.02	0.96	0.96	Yes
Science					
Grade	**df**	**RMSEA**	**CFI**	**TLI**	**Convergence**
5	1480	0.01	0.99	0.99	Yes
8*	1481	0.01	0.99	0.99	Yes
EOC					
Subject/Form	**df**	**RMSEA**	**CFI**	**TLI**	**Convergence**
Algebra 1 Core 24*	1593	0.03	0.97	0.97	Yes
Algebra 1 Core 25*	1593	0.02	0.98	0.98	Yes
Algebra 1 Core 26	1592	0.02	0.98	0.98	Yes
Algebra 1 Core 27*	1593	0.03	0.97	0.97	Yes
Geometry Core 19*	1593	0.03	0.97	0.96	Yes
Geometry Core 20	1592	0.02	0.97	0.97	Yes

Geometry Core 21*	1593	0.02	0.97	0.97	Yes
Biology 1 Core 100*	1482	0.01	0.99	0.99	Yes
Biology 1 Core 200*	1482	0.01	0.99	0.99	Yes
Biology 1 Core 300	1481	0.02	0.98	0.98	Yes
Civics Core 100*	1077	0.02	0.99	0.99	Yes
Civics Core 200	1076	0.02	0.98	0.98	Yes
Civics Core 300	1076	0.02	0.99	0.99	Yes
USH Core 100*	1273	0.01	0.99	0.99	Yes
USH Core 200*	1272	0.01	0.99	0.99	Yes
USH Core 300*	1272	0.01	0.99	0.99	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 some factors fell slightly below the boundary of zero for Grades 3, 5, 7, 8, and 9 ELA, Grades 3, 6, and 7 Mathematics, Grades 5 and 8 Science, and the EOC subjects when using the Mplus software package. This negative residual variance may be related to the computational implementation of the optimization approach in Mplus, 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). The residual variance was constrained to zero 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, Internal Consistency, items of Florida Statewide Assessments 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 41 through Table 44 we provide the estimated correlations between the reporting categories from the second-order factor model for Mathematics, ELA, EOC, and Science, respectively. In all cases except for Writing, these correlations are very high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

Table 41: Correlations Among Mathematics Factors

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4	Cat5
3	Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1)	1.00				
	Numbers and Operations – Fractions (Cat2)	0.90	1.00			
	Measurement, Data, and Geometry (Cat3)	0.96	0.93	1.00		
4	Operations and Algebraic Thinking (Cat1)	1.00				
	Numbers and Operations in Base Ten (Cat2)	0.97	1.00			
	Numbers and Operations – Fractions (Cat3)	0.95	0.96	1.00		
	Measurement, Data, and Geometry (Cat4)	0.95	0.96	0,94	1.00	
5	Operations, Algebraic Thinking, and Fractions (Cat1)	1.00				
	Numbers and Operations in Base Ten (Cat2)	0.96	1.00			
	Measurement, Data, and Geometry (Cat3)	0.95	0.97	1.00		
6	Ratio and Proportional Relationships (Cat1)	1.00				
	Expressions and Equations (Cat2)	0.98	1.00			
	Geometry (Cat3)	0.93	0.94	1.00		
	Statistics and Probability (Cat4)	0.95	0.96	0.92	1.00	
	The Number System (Cat5)	0.97	0.98	0.94	0.95	1.00
7	Ratio and Proportional Relationships (Cat1)	1.00				
	Expressions and Equations (Cat2)	0.98	1.00			
	Geometry (Cat3)	0.96	0.97	1.00		
	Statistics and Probability (Cat4)	0.96	0.97	0.95	1.00	
	The Number System (Cat5)	0.97	0.98	0.96	0.96	1.00
8	Expressions and Equations (Cat1)	1.00				
	Functions (Cat2)	0.94	1.00			
	Geometry (Cat3)	0.92	0.92	1.00		
	Statistics and Probability and the Number System (Cat4)	0.87	0.88	0.85	1.00	

*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 42: Correlations Among ELA Factors

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4	Cat5
3*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	1.00	1.00			
	Integration of Knowledge and Ideas (Cat3)	1.00	1.00	1.00		
	Language and Editing Task (Cat4)	0.88	0.88	0.88	1.00	
4	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	0.98	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.99	0.98	1.00		
	Language and Editing Task (Cat4)	0.87	0.88	0.88	1.00	
	Text-Based Writing (Cat5)	0.71	0.70	0.70	0.63	1.00
5*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	0.99	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.99	1.00	1.00		
	Language and Editing Task (Cat4)	0.94	0.95	0.95	1.00	
	Text-Based Writing (Cat5)	0.74	0.75	0.75	0.71	1.00
6*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	0.99	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.99	1.00	1.00		
	Language and Editing Task (Cat4)	0.91	0.91	0.91	1.00	
	Text-Based Writing (Cat5)	0.70	0.70	0.70	0.64	1.00
7*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	1.00	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.99	0.99	1.00		
	Language and Editing Task (Cat4)	0.95	0.95	0.94	1.00	
	Text-Based Writing (Cat5)	0.74	0.74	0.73	0.70	1.00
8*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	0.99	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.96	0.97	1.00		
	Language and Editing Task (Cat4)	0.92	0.93	0.90	1.00	
	Text-Based Writing (Cat5)	0.73	0.73	0.71	0.68	1.00
9*	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	1.00	1.00			

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4	Cat5
	Integration of Knowledge and Ideas (Cat3)	0.99	0.99	1.00		
	Language and Editing Task (Cat4)	0.91	0.91	0.91	1.00	
	Text-Based Writing (Cat5)	0.73	0.73	0.72	0.66	1.00
10	Key Ideas and Details (Cat1)	1.00				
	Craft and Structure (Cat2)	0.99	1.00			
	Integration of Knowledge and Ideas (Cat3)	0.99	0.99	1.00		
	Language and Editing Task (Cat4)	0.92	0.92	0.93	1.00	
	Text-Based Writing (Cat5)	0.75	0.75	0.75	0.70	1.00

*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 43: Correlations Among EOC Factors

Course/Form	Reporting Category	Cat1	Cat2	Cat3	Cat4
Algebra 1/Core 24*	Algebra and Modeling (Cat1)	1.00			
	Functions and Modeling (Cat2)	0.98	1.00		
	Statistics and the Number System (Cat3)	0.98	1.00	1.00	
Algebra 1/Core 25*	Algebra and Modeling (Cat1)	1.00			
	Functions and Modeling (Cat2)	0.99	1.00		
	Statistics and the Number System (Cat3)	0.98	0.99	1.00	
Algebra 1/Core 26	Algebra and Modeling (Cat1)	1.00			
	Functions and Modeling (Cat2)	0.98	1.00		
	Statistics and the Number System (Cat3)	0.98	1.00	1.00	
Algebra 1/Core 27*	Algebra and Modeling (Cat1)	1.00			
	Functions and Modeling (Cat2)	0.97	1.00		
	Statistics and the Number System (Cat3)	0.98	0.99	1.00	
Geometry/Core 19*	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	0.98	1.00		
	Modeling with Geometry (Cat3)	0.92	0.94	1.00	
Geometry/Core 20	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	0.99	1.00		
	Modeling with Geometry (Cat3)	0.94	0.95	1.00	

Course/Form	Reporting Category	Cat1	Cat2	Cat3	Cat4
Geometry/Core 21*	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	0.99	1.00		
	Modeling with Geometry (Cat3)	0.94	0.96	1.00	
Biology 1/Core 100*	Molecular and Cellular Biology (Cat1)	1.00			
	Classification, Heredity, and Evolution (Cat2)	0.99	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	0.98	0.99	1.00	
Biology 1/Core 200*	Molecular and Cellular Biology (Cat1)	1.00			
	Classification, Heredity, and Evolution (Cat2)	0.99	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	0.98	1.00	1.00	
Biology 1/Core 300	Molecular and Cellular Biology (Cat1)	1.00			
	Classification, Heredity, and Evolution (Cat2)	0.98	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	0.98	0.99	1.00	
Civics/Core 100*	Origins and Purposes of Law and Government (Cat1)	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	0.98	1.00		
	Government Policies and Political Processes (Cat3)	0.98	0.99	1.00	
	Organization and Function of Government (Cat4)	0.99	0.99	1.00	1.00
Civics/Core 200	Origins and Purposes of Law and Government (Cat1)	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	0.99	1.00		
	Government Policies and Political Processes (Cat3)	1.00	1.00	1.00	
	Organization and Function of Government (Cat4)	1.00	0.99	1.00	1.00
Civics/Core 300	Origins and Purposes of Law and Government (Cat1)	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	0.99	1.00		
	Government Policies and Political Processes (Cat3)	0.98	0.99	1.00	

Course/Form	Reporting Category	Cat1	Cat2	Cat3	Cat4
	Organization and Function of Government (Cat4)	0.99	0.99	0.98	1.00
U.S. History/Core 100*	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	1.00	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	0.99	0.99	1.00	
U.S. History/Core 200*	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	1.00	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	0.99	0.99	1.00	
U.S. History/Core 300*	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	1.00	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	0.99	1.00	1.00	

*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 44: Correlations Among Science Factors

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4
5	Nature of Science (Cat1)	1.00			
	Earth and Space Science (Cat2)	0.95	1.00		
	Physical Science (Cat3)	0.96	0.98	1.00	
	Life Science (Cat4)	0.95	0.97	0.98	1.00
8*	Nature of Science (Cat1)	1.00			
	Earth and Space Science (Cat2)	0.98	1.00		
	Physical Science (Cat3)	0.97	1.00	1.00	
	Life Science (Cat4)	0.97	1.00	0.99	1.00

*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.

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.

Item-Level Analyses

The Standards (AERA, APA, and NCME, 2014) suggests that the relationship between the test content and the intended test construct is one source of evidence for validity. 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 of this technical report 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 Florida Statewide Assessments are presented in this section. 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 test takers.

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 CAI, and the items are delivered by the same engine as is used for delivery of the Smarter Balanced assessment. Hence, the Florida Statewide Assessments 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 Smarter Balanced Assessment Consortium and also in Florida (see Volume 7 of the *Florida Standards Assessments 2014–2015 Technical Reports*). FDOE plans to conduct another set of cognitive lab studies in fall 2023.

The check for unidimensionality can be made at the item level. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called a point-biserial correlation when items are dichotomously scored) is the correlation between an item and the total test score. Conceptually, if an item has a high item-total correlation (that is, 0.30 or above), it indicates that students who performed well

on the test answered the item correctly and students who performed poorly on the test answered the item incorrectly; the item did a good job of discriminating between high-achieving and low-achieving students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate the items on the test require this construct to be answered correctly. We compute both biserial and point-biserial correlations in the Florida Statewide Assessments item bank though point-biserial correlations used in form evaluation. The point-biserial correlations for each operational item are presented in Appendix A of Volume 1 of this technical report.

Justification for the scaling procedures used for the Florida Statewide Assessments can be found in Volume 1 (see Item Calibration and Scaling) of this technical report.

4.2.3 Generalization Validity Evidence

There are two major requirements for validity that allow generalization from observed scale scores to universe scores[1]. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement comes from content validity. Content validity is documented through evidence that the test measures the content standards and benchmarks. The second requirement for validity at the generalization stage is that random measurement error on the test is controlled. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. 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. These sources of evidence are reported in the sections that follow.

Evidence of Content Validity

The Florida Statewide Assessments are based on content standards and benchmarks along with extensive content limits that help define what is to be assessed. Committees of educators collaborate with item development experts, assessment experts, and FDOE staff annually to review new and field-test items so that each test adequately samples the relevant domain of material the test is intended to cover. These review committees participate in this process to verify the content validity of each test.

The sequential committee review process is outlined in Volume 2 of this technical report. In addition to providing information on the difficulty, appropriateness, and fairness of items and performance tasks, committee members provide a check on the alignment between the items and the benchmarks measured. When items are judged to be relevant, that is, representative of the content defined by the standards, this provides evidence to support the validity of inferences made regarding knowledge of this content from the results. When items are judged to be inappropriate

[1] Universe score is defined as the expected value of a person's observed scores over all observations in the universe of generalization, which is analogous to a person's "true score" in classical test theory (Shavelson & Webb, 2006).

for any reason, the committee can either suggest revisions (e.g., rewording an item or reclassifying the item to a more appropriate benchmark) or elect to eliminate the item from the field-test item pool. Items approved are later embedded in live forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the content standards and measurement specifications so that the items measure the appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the test.

Skilled professionals are also involved in establishing evidence of content validity in other ways. Item writers must have at least three years teaching experience in the subject areas for which she or he will be creating items and tasks or two years of experience writing or reviewing items for the subject area. Each team is composed of qualified professionals who also have an understanding of psychometric considerations and sensitivity to racial/ethnic, gender, religious, and socioeconomic issues. Using a varied source of item writers provides a system of checks and balances for item development and review, reducing single-source bias. Since many different people with different backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. The input and review by these assessment professionals provide further support of the item being an accurate measure of the intended content domain.

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

Content Standards

The Florida Statewide Assessments were 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.fldoe.org.

Table 45 through Table 48 present the reporting categories by grade and test, as well as the number of items measuring each category. Table 49 through Table 51 present the number of items in each reporting category for the accommodated forms.

Table 45: 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	10
	Measurement, Data, and Geometry	18
4	Operations and Algebraic Thinking	11
	Numbers and Operations in Base Ten	11
	Numbers and Operations – Fractions	14

Grade*	Reporting Category	Number of Items
	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	12
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

* Reporting categories and the number of items belonging to each reporting category are identical for both online and accommodated forms except for Grade 8 (Table 49).

Table 46: Number of Items for Each ELA Reporting Category

Reporting Category	Grade*							
	3	4	5	6	7	8	9	10
Key Ideas and Details	13	14	18	13	14	14	16	18
Craft and Structure	17	15	14	19	18	18	14	16
Integration of Knowledge and Ideas	12	13	11	10	12	12	16	13
Language and Editing Task	8	8	7	10	8	8	8	7
Text-Based Writing	0	3	3	3	3	3	3	3

* Reporting categories and the number of items belonging to each reporting category are identical for both online and accommodated forms.

Table 47: Number of Items for Each EOC Reporting Category

Course	Reporting Category	Core Form									
		19	20	21	24	25	26	27	100	200	300
Algebra 1	Algebra and Modeling	-	-	-	24	24	24	24	-	-	-
	Functions and Modeling	-	-	-	23	23	23	23	-	-	-
	Statistics and the Number System	-	-	-	11	11	11	11	-	-	-
Geometry	Congruence, Similarity, Right Triangles, and Trigonometry	27	27	27	-	-	-	-	-	-	-
	Circles, Geometric Measurement, and Geometric Properties with Equations	22	22	22	-	-	-	-	-	-	-
	Modeling with Geometry	9	9	9	-	-	-	-	-	-	-
Biology 1	Molecular and Cellular Biology	-	-	-	-	-	-	-	20	20	20
	Classification, Heredity, and Evolution	-	-	-	-	-	-	-	14	14	14
	Organisms, Populations, and Ecosystems	-	-	-	-	-	-	-	22	22	22
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910	-	-	-	-	-	-	-	17	17	17
	Global Military, Political, and Economic Challenges, 1890–1940	-	-	-	-	-	-	-	18	18	18
	The United States and the Defense of the International Peace, 1940–Present	-	-	-	-	-	-	-	17	17	17
Civics	Origins and Purposes of Law and Government	-	-	-	-	-	-	-	12	12	12
	Roles, Rights, and Responsibilities of Citizens	-	-	-	-	-	-	-	12	12	12
	Government Policies and Political Processes	-	-	-	-	-	-	-	12	12	12
	Organization and Function of Government	-	-	-	-	-	-	-	12	12	12

Table 48: Number of Items for Each Science Reporting Category

Grade	Reporting Category	Number of Items
5	Nature of Science	10
	Earth and Space Science	16
	Physical Science	16
	Life Science	14
8	Nature of Science	11
	Earth and Space Science	15
	Physical Science	15
	Life Science	15

Table 49: Number of Items for Each Mathematics Accommodated Reporting Category

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 50: Number of Items for Each ELA Accommodated Reporting Category

Reporting Category	Grade*			
	7	8	9	10
Key Ideas and Details	14	14	16	18
Craft and Structure	18	18	14	16
Integration of Knowledge and Ideas	12	12	16	13
Language and Editing Task	8	8	8	7
Text-Based Writing	3	3	3	3

Table 51: 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
Geometry	Congruence, Similarity, Right Triangles, and Trigonometry	27
	Circles, Geometric Measurement, and Geometric Properties with Equations	22
	Modeling with Geometry	9
Biology 1	Molecular and Cellular Biology	20
	Classification, Heredity, and Evolution	14
	Organisms, Populations, and Ecosystems	22
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910	17
	Global Military, Political, and Economic Challenges, 1890–1940	18
	The United States and the Defense of the International Peace, 1940–Present	17

Course	Reporting Category	Number of Items
Civics	Origins and Purposes of Law and Government	12
	Roles, Rights, and Responsibilities of Citizens	12
	Government Policies and Political Processes	12
	Organization and Function of Government	12

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, of this technical report. The Florida Statewide Assessments 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). TEIs 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.

Test Development

For the 2022 Florida Statewide Assessments administration, Cambium Assessment Inc. (CAI) and Pearson in collaboration with the Florida Department of Education and its Test Development Center (TDC), constructed test forms for Grades 3 through 10 ELA, Grades 3 through 8 Mathematics, Grade 5 and 8 Science, and End-of-Course Assessments (Algebra 1, Geometry, Biology 1, Civics, U.S. History).

The test forms administered in spring 2022 are generated from the test construction activities that happened in summer 2021. During summer 2021, psychometricians and content experts from FDOE, the TDC, and CAI convened virtually for two weeks, and FDOE, the TDC, and Pearson met face to face for one week, to build forms for spring 2022. 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, of this technical report.

The Florida Statewide 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. 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, CAI/Pearson 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

Alignment of Florida Statewide Assessments 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 Florida Statewide Assessments 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 was provided in Volume 4, Appendix D, of the *2015–2016 Florida Standards Assessments Technical Report*.

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).

A third-party, independent alignment study was conducted in 2012 to evaluate the alignment between test items and benchmarks they intend to measure for Grade 5 and 8 Science and Biology 1 EOC assessments. Only benchmarks designated to be assessed on the statewide on-demand assessments were included in the analysis. These benchmarks for the science assessments have not changed since 2012.

Response Processes solicited by Florida Statewide Assessments

The Standards for Educational and Psychological Testing note that "some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers" (p.15). This is true with educational assessments in which the content claims include that items are measured at levels of higher cognitive complexity. Both theoretical and empirical analyses of test

taker processes can be used as evidence for such claims. Cognitive labs, in which researchers question test takers from the student population about their steps in responding to a question, how they solved a question (response strategy) are strong pieces of evidence that the assessments tap the intended cognitive processes appropriate for each grade level as represented in the academic content standards measured.

Florida Statewide Assessments had planned cognitive labs to study the response processes of test takers for Grades 3, 7, and 10 ELA, Grades 3 and 7 Mathematics, Algebra 1, Grade 5 Science, and Biology 1. These grades/subjects are selected because they represent the item types, share similar blueprints and test development procedures. We believe that results from cognitive lab studies from these grades/subjects are generalizable to non-selected grades and non-selected item types.

According to the plan, students will work through a sample of either Mathematics or ELA or Science items in a cognitive lab. Eight students will respond to each item, and their thinking processes will be elicited through a combination of concurrent think-aloud (thinking out loud while reading and responding to an item) and focused probes that are tailored based on the anticipated solution path for a given item.

The cognitive lab interviews will be audio recorded, and the students' responses to the test items will be captured by the Test Delivery System. Following the cognitive lab, the interviewer will review all relevant information and file a report that includes, for each item attempted by the student, a detailed record of the student's think-aloud and responses to probes, as well as a record of the student's test item response.

These reports will be evaluated by content experts to determine whether the evidence for any given item meets the following criteria:

1. Students who receive full credit on an item display—through their think-aloud and responses to probes—defensible evidence that they based their response on the combination of skills and knowledge that make up the "intended construct."
2. Students who do not receive full credit on an item display—through their think aloud and responses to probes—defensible evidence that
    a. they understood (at a general level) what the item was asking them to do, and
    b. they were unable to provide a full-credit response as a result of deficiencies in one or more aspect of the skills or knowledge that make up the "intended construct." For example, they lacked the necessary procedural knowledge for manipulating fractions or they were unable to apply the reasoning skills required by the item.

The planned cognitive lab studies were delayed due to the COVID-19 pandemic and school closings in 2020–2021. These studies are planned to be conducted in fall 2023.

Evidence of Control of Measurement Error

Reliability and the standard error of measurement (SEM) are discussed in an earlier chapter of this volume. Tables reporting the CSEM and coefficient alpha reliability are also included. As discussed earlier, these measures show that Florida Statewide Assessments scores are reliable.

Further evidence is needed to show the IRT model fits well. Item-fit statistics and tests of unidimensionality apply here, as they did in the section describing evidence argument for scoring. As described, these measures indicate good fit of the model.

Validity Evidence for Different Student Populations

It can be argued from a content perspective that the Florida Statewide Assessments are not more or less valid for use with one subpopulation of students relative to another. The Florida Statewide Assessments measure Florida Standards, which are required to be taught to all students. The tests have the same content validity for all students because what is measured on the tests is taught to all students, and all tests are given to all students under standardized conditions.

Great care has been taken so that the items constituting the Florida Statewide Assessments are fair and representative of the content domain expressed in the content standards. Additionally, much scrutiny is applied to the items and their possible impact on demographic subgroups making up the population of the state of Florida. Every effort is made to eliminate items that may have ethnic or cultural biases. As described in Volume 2 of this technical report, item writers are trained on how to avoid economic, regional, cultural, and ethnic biases when writing items. After items are written and passage selections are made, committees of Florida educators are convened by FDOE to examine items for potential subgroup bias. As described in Volume 1, items are further reviewed for potential bias by committees of educators and the FDOE after field-test data are collected. Volume 1 of this technical report delineated the differential item functioning (DIF) analysis which was conducted for all items to detect potential item bias across major gender, ethnic, and special population groups. In fact, DIF analysis is conducted for all items before the item is added to any operational form. DIF summary tables are presented in the appendices of Volume 1 in the *Florida Statewide Assessments 2021-2022 Technical Report*: Appendix A, Operational Item Statistics, for operational items, Appendix B, Anchor Item Statistics, for anchor items, and Appendix C, Field-Test Item Statistics, for field-test items.

In addition, the coefficient alpha reliability was calculated for various demographic subgroups including gender groups (male and female), ethnic groups (white, African American, Hispanic, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and multiracial), ELL and Non-ELL, students with/without disability, and students with/without accommodations (see the reliability in the Appendix A of this volume and classification accuracy in the Reliability chapter of this volume). These reliability measures provide one more piece of evidence for the content validity across demographic subgroups.

4.2.4 Extrapolation Validity Evidence

Validity for extrapolation requires evidence that the universe score is applicable to the larger domain of interest. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. These lines of evidence are detailed below.

Analytic Evidence

The Florida Statewide Assessments create a common foundation to be learned by all students and define the domain of interest. As documented in this report, the Florida Statewide Assessments are designed to measure as much of the domain defined by the standards as possible.

A threat to the validity of the test can arise when the assessment requires competence in a skill unrelated to the construct being measured. For example, students who are ELL may have difficulty fully demonstrating their mathematical knowledge if the mathematics assessment requires fluency in English. The use of accommodation avoids this threat to validity by allowing students who are ELL to demonstrate their mathematical ability on a test that limits the quantity and complexity of English language used in the items. The Florida Statewide Assessments also allow accommodations for students with vision impairment or other special needs. The use of accommodated forms allows accurate measurement of students who would otherwise be unfairly disadvantaged by taking the standard form. Accommodations are discussed in Volume 5 of this technical report. Further, the coefficient alpha reliability measures for the ELL, disability, and accommodation groups (see the reliability and classification accuracy in the Appendix A of this volume), in particular, provide some evidence for the effectiveness of accommodations that would allow meaningful interpretation of results and comparisons across subgroups.

Another threat to test validity could arise when the assessments are administered online on different platforms. Online administration of Florida Statewide Assessments in spring 2022 included Grades 7–8 Mathematics, Grades 7–10 Reading, and all EOC assessments (Algebra 1, Geometry, Biology 1, U.S. History, and Civics). According to the Technology Guidelines of FDOE (2015), "Desktops, laptops, netbooks (Windows, Mac, Chrome, Linux), thin client, and tablets (iPad, Windows and Android) will be compatible devices provided they meet the established hardware, operating system and networking specifications—and are able to address the security requirements." All these devices can be used for EOC administration if the screen size is 9.5 inches or larger. To provide support for the use of multiple devices on Florida EOC assessments, a brief literature review was included about the score comparability across digital devices on large-scale assessments.

Way, Davis, Keng, and Strain-Seymour (2016) pointed out a fundamental consideration in evaluating device comparability: form factor. Form factor is defined as the way students access and manipulate digital content with the devices—the more similar the form factor, the more comparable the scores on those two devices can be expected to be. Form factors for desktop and laptop computers are relatively similar, especially when compared to tablets (e.g., iPad) devices. A lot of earlier research has shown that student performance across desktop and laptop computers is relatively comparable (Keng, Kong, & Bleil, 2011; Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, & Oranje, 2005; Bridgeman, Lennon, & Jackenthal, 2001). Since the current generation of touch-screen tablets became available in 2010, only research after 2010 is cited below to further examine the score comparability between tablet and non-tablet devices.

Olsen (2014) compared performance of grades 1–12 testing on tablet and computer. He found strong positive relationships for student scale scores across devices and concluded that these results provided "strong evidence that STAR Reading Enterprise and STAR Math Enterprise were measuring the same attribute regardless of device type" (p. 2). Although statistically significant differences were reported for some grades for Reading and Mathematics, the device effects were

found favoring computers in some grades and tablets in others. The effect sizes for reading ranged from small to very small.

In their Partnership for Assessment of Readiness for College and Careers (PARCC) spring 2015 digital device comparability study, Steedle, McBride, Johnson, & Keng (2016) found "consistent" and "robust" evidence of comparability between test scores from tablet and non-tablet devices. This study examined performance on eight PARCC assessments Grade 5 Mathematics, Grade 7 Mathematics, Algebra 1, Geometry, Algebra 2, Grade 3 English Language Arts/Literacy (ELA/L), Grade 7 ELA/L, and Grade 9 ELA/L. Students who used tablet and non-tablet devices were matched on demographic information so that two randomly equivalent samples are generated. The item means and IRT difficulty estimates were found similar across devices. While a small number of items were flagged for device effects, they are almost all on high school Mathematics assessments. The raw score and scale score distributions suggested similar overall performance on both performance-based and end-of-year components of the 2015 PARCC assessments.

In addition, IRT true-score equating indicated that students testing on non-tablet devices would be expected to obtain similar scores if they had taken the same test on tablets.

Davis, Kong, McBride, & Morrison (2016) examined the comparability of scores for high school students testing on computers to those testing on tablets. This study addressed construct equivalence and mean differences on Reading, Mathematics, and Science assessments with a variety of item types (multiple-choice and technology-enhanced items). They found no significant mean score differences across devices for any of the three content areas or across any item type evaluated. Construct equivalence also held across devices. Further, Davis, Morrison, Kong, & McBride (2017) extended this research by comparing score distributions across devices for Reading, Mathematics, and Science, and also investigating device effects for gender and ethnicity subgroups. For Mathematics and Science, no significant differences were found between scores that resulted from tablets and computers. For reading, a small device effect favoring tablets was found for the middle to lower part of the score distribution, which might be caused by performance increases of male students testing on tablets. Overall, this study adds to the evidence "for a relatively high degree of comparability between tablets and computers" (p. 35), which is consistent with previous studies reviewed in this section.

In terms of screen size, research suggests that, while the information shown on the screen is held constant, screens of 10 inches or larger are suitable for viewing and interacting with assessments, with little evidence of test performance differences or item-level differences (Keng, Kong, & Bleil, 2011; Davis, Strain-Seymour, & Gay, 2013). This provides further support for Florida EOC assessments to allow the use of tablets with screen size of 9.7 inches or larger.

While it is reassuring that the research generally finds the scores across digital devices to be comparable, DePascale, Dadey, & Lyons (2016) summarized factors that may potentially contribute to the presence of device effects: familiarity, device features (screen size, input mechanism, keyboard), and assessment-specific features (content area). They recommended that when different devices are allowed on an assessment, states should attempt to eliminate or minimize differences in the areas listed above. In particular,

> *differences in devices can be minimized if all students are sufficiently fluent with the functionality of the device on which they are testing; the amount of content that appears on the screen without requiring scrolling is the same across devices; the items are designed*

for comfortable use with fingertip input when touchscreen devices are used (e.g., items are large enough and spaced widely enough); and external keyboards are available for response to essay prompt. (p.17)

Empirical Evidence

Empirical evidence of extrapolation is generally provided by criterion validity when a suitable criterion exists. As discussed before, finding an adequate criterion for a standards-based achievement test can be difficult.

According to *The Standards (*AERA, APA, and NCME, 2014), the convergent and discriminant evidence is one category within the source of validity evidence of the relationship of test scores to external variables. 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 multi-trait-multimethod matrix can be used. Thus, another strategy to examine the convergent and divergent validity could be accomplished by looking at the subscore relationships (by reporting category) within content areas. 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.

Correlations Among Reporting Category Scores

Table 52 through Table 55 present the observed correlation matrix of the reporting category raw scores for each subject area. For Mathematics, the correlations were between 0.59 and 0.85. In ELA, the correlations among the reporting categories range from 0.46 to 0.81. The Language and Editing Task items and Text-Based Writing items exhibited slightly lower correlations with the other reporting categories, ranging from 0.46 to 0.69. For EOC, the correlations fell between 0.69 and 0.85. In Science, the correlations range from 0.70 to 0.80.

Observed correlations from the accommodated forms are presented in Table 56 through Table 58. Note that Grade 5 and 8 Science do not have accommodated forms. The correlations varied between 0.52 and 0.69 for Mathematics, 0.41 and 0.80 for ELA, and 0.60 and 0.78 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 Florida Department of Education cautions each year when scores are released.

Table 59 through Table 65 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 0.67, and the minimum was 0.61, whereas the overall average disattenuated correlation for ELA was 0.84.

Table 52: 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)	10	0.76	1.00			
	Measurement, Data, and Geometry (Cat3)	18	0.85	0.77	1.00		
4	Operations and Algebraic Thinking (Cat1)	11	1.00				
	Numbers and Operations in Base Ten (Cat2)	11	0.83	1.00			
	Numbers and Operations – Fractions (Cat3)	14	0.80	0.81	1.00		
	Measurement, Data, and Geometry (Cat4)	18	0.81	0.81	0.82	1.00	
5	Operations, Algebraic Thinking, and Fractions (Cat1)	21	1.00				
	Numbers and Operations in Base Ten (Cat2)	15	0.84	1.00			
	Measurement, Data, and Geometry (Cat3)	18	0.82	0.82	1.00		
6	Ratio and Proportional Relationships (Cat1)	8	1.00				
	Expressions and Equations (Cat2)	17	0.76	1.00			
	Geometry (Cat3)	8	0.67	0.74	1.00		
	Statistics and Probability (Cat4)	11	0.69	0.75	0.66	1.00	
	The Number System (Cat5)	12	0.72	0.80	0.71	0.70	1.00
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.77	1.00			
	Geometry (Cat3)	13	0.74	0.74	1.00		
	Statistics and Probability (Cat4)	9	0.69	0.69	0.68	1.00	
	The Number System (Cat5)	8	0.74	0.75	0.72	0.66	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	14	0.71	1.00			
	Geometry (Cat3)	15	0.72	0.68	1.00		
	Statistics and Probability and the Number System (Cat4)	10	0.62	0.59	0.63	1.00	

Table 53: Observed Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Key Ideas and Details (Cat1)	13	1.00				
	Craft and Structure (Cat2)	17	0.81	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.77	0.79	1.00		
	Language and Editing Task (Cat4)	8	0.60	0.62	0.59	1.00	
4	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	15	0.78	1.00			
	Integration of Knowledge and Ideas (Cat3)	13	0.80	0.77	1.00		
	Language and Editing Task (Cat4)	8	0.58	0.57	0.57	1.00	
	Text-Based Writing (Cat5)	3	0.58	0.56	0.56	0.49	1.00
5	Key Ideas and Details (Cat1)	18	1.00				
	Craft and Structure (Cat2)	14	0.78	1.00			
	Integration of Knowledge and Ideas (Cat3)	11	0.76	0.71	1.00		
	Language and Editing Task (Cat4)	7	0.69	0.65	0.62	1.00	
	Text-Based Writing (Cat5)	3	0.59	0.56	0.53	0.54	1.00
6	Key Ideas and Details (Cat1)	13	1.00				
	Craft and Structure (Cat2)	19	0.79	1.00			
	Integration of Knowledge and Ideas (Cat3)	10	0.71	0.70	1.00		
	Language and Editing Task (Cat4)	10	0.61	0.61	0.56	1.00	
	Text-Based Writing (Cat5)	3	0.50	0.50	0.46	0.49	1.00
7	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.80	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.74	0.72	1.00		
	Language and Editing Task (Cat4)	8	0.67	0.66	0.61	1.00	
	Text-Based Writing (Cat5)	3	0.58	0.56	0.54	0.54	1.00
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.81	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.73	0.74	1.00		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Language and Editing Task (Cat4)	8	0.62	0.64	0.57	1.00	
	Text-Based Writing (Cat5)	3	0.54	0.55	0.50	0.53	1.00
9	Key Ideas and Details (Cat1)	16	1.00				
	Craft and Structure (Cat2)	14	0.80	1.00			
	Integration of Knowledge and Ideas (Cat3)	16	0.81	0.75	1.00		
	Language and Editing Task (Cat4)	8	0.63	0.59	0.60	1.00	
	Text-Based Writing (Cat5)	3	0.58	0.57	0.54	0.48	1.00
10	Key Ideas and Details (Cat1)	18	1.00				
	Craft and Structure (Cat2)	16	0.80	1.00			
	Integration of Knowledge and Ideas (Cat3)	13	0.76	0.73	1.00		
	Language and Editing Task (Cat4)	7	0.65	0.64	0.60	1.00	
	Text-Based Writing (Cat5)	3	0.57	0.57	0.52	0.52	1.00

Table 54: Observed Correlation Matrix Among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1/Core 24	Algebra and Modeling (Cat1)	24	1.00			
	Functions and Modeling (Cat2)	23	0.84	1.00		
	Statistics and the Number System (Cat3)	11	0.79	0.78	1.00	
Algebra 1/Core 25	Algebra and Modeling (Cat1)	24	1.00			
	Functions and Modeling (Cat2)	23	0.85	1.00		
	Statistics and the Number System (Cat3)	11	0.76	0.76	1.00	
Algebra 1/Core 26	Algebra and Modeling (Cat1)	24	1.00			
	Functions and Modeling (Cat2)	23	0.84	1.00		
	Statistics and the Number System (Cat3)	11	0.78	0.77	1.00	
Algebra 1/Core 27	Algebra and Modeling (Cat1)	24	1.00			
	Functions and Modeling (Cat2)	23	0.83	1.00		
	Statistics and the Number System (Cat3)	11	0.79	0.77	1.00	
Geometry/Core 19	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1.00			

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.83	1.00		
	Modeling with Geometry (Cat3)	9	0.69	0.72	1.00	
Geometry/Core 20	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.85	1.00		
	Modeling with Geometry (Cat3)	9	0.71	0.72	1.00	
Geometry/Core 21	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.84	1.00		
	Modeling with Geometry (Cat3)	9	0.74	0.76	1.00	
Biology 1/Core 100	Molecular and Cellular Biology (Cat1)	20	1.00			
	Classification, Heredity, and Evolution (Cat2)	14	0.78	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.80	0.78	1.00	
Biology 1/Core 200	Molecular and Cellular Biology (Cat1)	20	1.00			
	Classification, Heredity, and Evolution (Cat2)	14	0.76	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.79	0.76	1.00	
Biology 1/Core 300	Molecular and Cellular Biology (Cat1)	20	1.00			
	Classification, Heredity, and Evolution (Cat2)	14	0.77	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.79	0.78	1.00	
U.S. History/Core 100	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.80	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.78	0.78	1.00	
U.S. History/Core 200	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1.00			

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.79	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.76	0.77	1.00	
U.S. History/Core 300	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.79	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.77	0.78	1.00	
Civics/Core 100	Origins and Purposes of Law and Government (Cat1)	12	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.74	1.00		
	Government Policies and Political Processes (Cat3)	12	0.72	0.74	1.00	
	Organization and Function of Government (Cat4)	12	0.73	0.75	0.73	1.00
Civics/Core 200	Origins and Purposes of Law and Government (Cat1)	12	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.74	1.00		
	Government Policies and Political Processes (Cat3)	12	0.70	0.71	1.00	
	Organization and Function of Government (Cat4)	12	0.73	0.72	0.69	1.00
Civics/Core 300	Origins and Purposes of Law and Government (Cat1)	12	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.76	1.00		
	Government Policies and Political Processes (Cat3)	12	0.73	0.75	1.00	
	Organization and Function of Government (Cat4)	12	0.72	0.72	0.71	1.00

Table 55: Observed Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
5	Nature of Science (Cat1)	10	1.00			
	Earth and Space Science (Cat2)	16	0.71	1.00		
	Physical Science (Cat3)	16	0.75	0.77	1.00	
	Life Science (Cat4)	14	0.70	0.76	0.76	1.00
8	Nature of Science (Cat1)	11	1.00			
	Earth and Space Science (Cat2)	15	0.76	1.00		
	Physical Science (Cat3)	15	0.75	0.79	1.00	
	Life Science (Cat4)	15	0.75	0.80	0.79	1.00

Table 56: Observed Correlation Matrix Among Reporting Categories (Mathematics Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
7	Ratio and Proportional Relationships (Cat1)	14	1.00				
	Expressions and Equations (Cat2)	12	0.69	1.00			
	Geometry (Cat3)	13	0.57	0.58	1.00		
	Statistics and Probability (Cat4)	9	0.54	0.59	0.54	1.00	
	The Number System (Cat5)	8	0.62	0.63	0.52	0.54	1.00
8	Expressions and Equations (Cat1)	17	1.00				
	Functions (Cat2)	14	0.66	1.00			
	Geometry (Cat3)	15	0.68	0.61	1.00		
	Statistics and Probability and the Number System (Cat4)	10	0.60	0.52	0.58	1.00	

Table 57: Observed Correlation Matrix Among Reporting Categories (ELA Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
7	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.76	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.71	0.69	1.00		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Language and Editing Task (Cat4)	8	0.54	0.56	0.49	1.00	
	Text-Based Writing (Cat5)	3	0.49	0.52	0.51	0.41	1.00
8	Key Ideas and Details (Cat1)	14	1.00				
	Craft and Structure (Cat2)	18	0.80	1.00			
	Integration of Knowledge and Ideas (Cat3)	12	0.72	0.72	1.00		
	Language and Editing Task (Cat4)	8	0.56	0.61	0.51	1.00	
	Text-Based Writing (Cat5)	3	0.46	0.49	0.41	0.42	1.00
9	Key Ideas and Details (Cat1)	16	1.00				
	Craft and Structure (Cat2)	14	0.77	1.00			
	Integration of Knowledge and Ideas (Cat3)	16	0.79	0.73	1.00		
	Language and Editing Task (Cat4)	8	0.59	0.56	0.55	1.00	
	Text-Based Writing (Cat5)	3	0.55	0.54	0.52	0.46	1.00
10	Key Ideas and Details (Cat1)	18	1.00				
	Craft and Structure (Cat2)	16	0.79	1.00			
	Integration of Knowledge and Ideas (Cat3)	13	0.76	0.74	1.00		
	Language and Editing Task (Cat4)	7	0.64	0.63	0.63	1.00	
	Text-Based Writing (Cat5)	3	0.50	0.51	0.47	0.42	1.00

Table 58: Observed Correlation Matrix Among Reporting Categories (EOC Accommodated Forms)

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1	Algebra and Modeling (Cat1)	24	1.00			
	Functions and Modeling (Cat2)	23	0.78	1.00		
	Statistics and the Number System (Cat3)	11	0.74	0.72	1.00	
Geometry	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1.00			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.73	1.00		
	Modeling with Geometry (Cat3)	9	0.60	0.61	1.00	
Biology 1	Molecular and Cellular Biology (Cat1)	20	1.00			

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Classification, Heredity, and Evolution (Cat2)	14	0.69	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.74	0.68	1.00	
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.78	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.74	0.74	1.00	
Civics	Origins and Purposes of Law and Government (Cat1)	12	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.68	1.00		
	Government Policies and Political Processes (Cat3)	12	0.66	0.70	1.00	
	Organization and Function of Government (Cat4)	12	0.66	0.66	0.69	1.00

Table 59: 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				
	Numbers and Operations – Fractions (Cat2)	10	0.92	1			
	Measurement, Data, and Geometry (Cat3)	18	0.96	0.96	1		
4	Operations and Algebraic Thinking (Cat1)	11	1				
	Numbers and Operations in Base Ten (Cat2)	11	0.99	1			
	Numbers and Operations – Fractions (Cat3)	14	0.95	0.95	1		
	Measurement, Data, and Geometry (Cat4)	18	0.96	0.96	0.96	1	
5	Operations, Algebraic Thinking, and Fractions (Cat1)	21	1				
	Numbers and Operations in Base Ten (Cat2)	15	0.97	1			
	Measurement, Data, and Geometry (Cat3)	18	0.96	0.97	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
6	Ratio and Proportional Relationships (Cat1)	8	1				
	Expressions and Equations (Cat2)	17	0.99	1			
	Geometry (Cat3)	8	0.93	0.93	1		
	Statistics and Probability (Cat4)	11	0.98	0.96	0.91	1	
	The Number System (Cat5)	12	0.97	0.98	0.94	0.94	1
7	Ratio and Proportional Relationships (Cat1)	14	1				
	Expressions and Equations (Cat2)	12	0.97	1			
	Geometry (Cat3)	13	0.95	0.97	1		
	Statistics and Probability (Cat4)	9	0.94	0.96	0.96	1	
	The Number System (Cat5)	8	0.96	1.00	0.97	0.96	1
8	Expressions and Equations (Cat1)	17	1				
	Functions (Cat2)	14	0.96	1			
	Geometry (Cat3)	15	0.92	0.92	1		
	Statistics and Probability and the Number System (Cat4)	10	0.85	0.85	0.87	1	

Table 60: Disattenuated Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Key Ideas and Details (Cat1)	13	1				
	Craft and Structure (Cat2)	17	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	12	1.00	1.00	1		
	Language and Editing Task (Cat4)	8	0.89	0.90	0.89	1	
4	Key Ideas and Details (Cat1)	14	1				
	Craft and Structure (Cat2)	15	0.98	1			
	Integration of Knowledge and Ideas (Cat3)	13	1.00	0.98	1		
	Language and Editing Task (Cat4)	8	0.86	0.86	0.85	1	
	Text-Based Writing (Cat5)	3	0.69	0.68	0.68	0.71	1
5	Key Ideas and Details (Cat1)	18	1				
	Craft and Structure (Cat2)	14	1.00	1			

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Integration of Knowledge and Ideas (Cat3)	11	1.00	1.00	1		
	Language and Editing Task (Cat4)	7	0.93	0.94	0.94	1	
	Text-Based Writing (Cat5)	3	0.69	0.71	0.70	0.71	1
6	Key Ideas and Details (Cat1)	13	1				
	Craft and Structure (Cat2)	19	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	10	1.00	1.00	1		
	Language and Editing Task (Cat4)	10	0.86	0.87	0.89	1	
	Text-Based Writing (Cat5)	3	0.61	0.62	0.63	0.66	1
7	Key Ideas and Details (Cat1)	14	1				
	Craft and Structure (Cat2)	18	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	12	1.00	1.00	1		
	Language and Editing Task (Cat4)	8	0.93	0.95	0.93	1	
	Text-Based Writing (Cat5)	3	0.67	0.68	0.69	0.73	1
8	Key Ideas and Details (Cat1)	14	1				
	Craft and Structure (Cat2)	18	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	12	0.97	0.97	1		
	Language and Editing Task (Cat4)	8	0.89	0.90	0.87	1	
	Text-Based Writing (Cat5)	3	0.64	0.65	0.62	0.72	1
9	Key Ideas and Details (Cat1)	16	1				
	Craft and Structure (Cat2)	14	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	16	1.00	0.99	1		
	Language and Editing Task (Cat4)	8	0.89	0.90	0.89	1	
	Text-Based Writing (Cat5)	3	0.66	0.69	0.65	0.67	1
10	Key Ideas and Details (Cat1)	18	1				
	Craft and Structure (Cat2)	16	0.99	1			
	Integration of Knowledge and Ideas (Cat3)	13	1.00	0.98	1		
	Language and Editing Task (Cat4)	7	0.91	0.91	0.92	1	
	Text-Based Writing (Cat5)	3	0.67	0.69	0.67	0.70	1

Table 61: Disattenuated Correlation Matrix Among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1/Core 24	Algebra and Modeling (Cat1)	24	1			
	Functions and Modeling (Cat2)	23	0.99	1		
	Statistics and the Number System (Cat3)	11	0.97	0.97	1	
Algebra 1/Core 25	Algebra and Modeling (Cat1)	24	1			
	Functions and Modeling (Cat2)	23	0.99	1		
	Statistics and the Number System (Cat3)	11	0.97	0.99	1	
Algebra 1/Core 26	Algebra and Modeling (Cat1)	24	1			
	Functions and Modeling (Cat2)	23	0.98	1		
	Statistics and the Number System (Cat3)	11	0.98	0.99	1	
Algebra 1/Core 27	Algebra and Modeling (Cat1)	24	1			
	Functions and Modeling (Cat2)	23	0.98	1		
	Statistics and the Number System (Cat3)	11	1.00	0.99	1	
Geometry/Core 19	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.97	1		
	Modeling with Geometry (Cat3)	9	0.91	0.96	1	
Geometry/Core 20	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.99	1		
	Modeling with Geometry (Cat3)	9	0.94	0.96	1	
Geometry/Core 21	Congruence, Similarity, Right Triangles, and Trigonometry (Cat1)	27	1			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.99	1		
	Modeling with Geometry (Cat3)	9	0.94	0.97	1	
Biology 1/Core 100	Molecular and Cellular Biology (Cat1)	20	1			
	Classification, Heredity, and Evolution (Cat2)	14	0.90	1		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.93	0.94	1	
Biology 1/Core 200	Molecular and Cellular Biology (Cat1)	20	1			

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Classification, Heredity, and Evolution (Cat2)	14	0.88	1		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.92	0.92	1	
Biology 1/Core 300	Molecular and Cellular Biology (Cat1)	20	1			
	Classification, Heredity, and Evolution (Cat2)	14	0.90	1		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.92	0.94	1	
U.S. History/Core 100	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.94	1		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.92	0.93	1	
U.S. History/Core 200	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.92	1		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.89	0.91	1	
U.S. History/Core 300	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.93	1		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.90	0.92	1	
Civics/Core 100	Origins and Purposes of Law and Government (Cat1)	12	1			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.92	1		
	Government Policies and Political Processes (Cat3)	12	0.89	0.89	1	
	Organization and Function of Government (Cat4)	12	0.90	0.90	0.89	1
Civics/Core 200	Origins and Purposes of Law and Government (Cat1)	12	1			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.90	1		

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Government Policies and Political Processes (Cat3)	12	0.84	0.86	1	
	Organization and Function of Government (Cat4)	12	0.87	0.87	0.86	1
Civics/Core 300	Origins and Purposes of Law and Government (Cat1)	12	1			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.91	1		
	Government Policies and Political Processes (Cat3)	12	0.88	0.90	1	
	Organization and Function of Government (Cat4)	12	0.87	0.87	0.87	1

Table 62: Disattenuated Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
5	Nature of Science (Cat1)	10	1			
	Earth and Space Science (Cat2)	16	0.83	1		
	Physical Science (Cat3)	16	0.88	0.91	1	
	Life Science (Cat4)	14	0.83	0.90	0.91	1
8	Nature of Science (Cat1)	11	1			
	Earth and Space Science (Cat2)	15	0.90	1		
	Physical Science (Cat3)	15	0.89	0.94	1	
	Life Science (Cat4)	15	0.89	0.95	0.94	1

Table 63: Disattenuated Correlation Matrix Among Reporting Categories (Mathematics Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
7	Ratio and Proportional Relationships (Cat1)	14	1				
	Expressions and Equations (Cat2)	12	1.00	1			
	Geometry (Cat3)	13	0.91	0.91	1		
	Statistics and Probability (Cat4)	9	0.93	1.00	1.00	1	
	The Number System (Cat5)	8	0.98	0.99	0.89	0.98	1
8	Expressions and Equations (Cat1)	17	1				

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Functions (Cat2)	14	0.97	1			
	Geometry (Cat3)	15	0.95	0.96	1		
	Statistics and Probability and the Number System (Cat4)	10	0.92	0.89	0.95	1	

Table 64: Disattenuated Correlation Matrix Among Reporting Categories (ELA Accommodated Forms)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
7	Key Ideas and Details (Cat1)	14	1				
	Craft and Structure (Cat2)	18	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	12	0.99	1.00	1		
	Language and Editing Task (Cat4)	8	0.81	0.90	0.80	1	
	Text-Based Writing (Cat5)	3	0.59	0.66	0.66	0.57	1
8	Key Ideas and Details (Cat1)	14	1				
	Craft and Structure (Cat2)	18	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	12	1.00	0.98	1		
	Language and Editing Task (Cat4)	8	0.83	0.90	0.82	1	
	Text-Based Writing (Cat5)	3	0.55	0.59	0.53	0.59	1
9	Key Ideas and Details (Cat1)	16	1				
	Craft and Structure (Cat2)	14	1.00	1			
	Integration of Knowledge and Ideas (Cat3)	16	1.00	1.00	1		
	Language and Editing Task (Cat4)	8	0.91	0.93	0.90	1	
	Text-Based Writing (Cat5)	3	0.63	0.67	0.65	0.67	1
10	Key Ideas and Details (Cat1)	18	1				
	Craft and Structure (Cat2)	16	0.97	1			
	Integration of Knowledge and Ideas (Cat3)	13	1.00	1.00	1		
	Language and Editing Task (Cat4)	7	0.91	0.92	0.99	1	
	Text-Based Writing (Cat5)	3	0.58	0.60	0.61	0.58	1

Table 65: Disattenuated Correlation Matrix Among Reporting Categories (EOC Accommodated Forms)

Course	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1	Algebra and Modeling (Cat1)	24	1			
	Functions and Modeling (Cat2)	23	0.99	1		
	Statistics and the Number System (Cat3)	11	0.96	0.96	1	
Geometry	Congruence, Similarity, Right Triangles and Trigonometry (Cat1)	27	1			
	Circles, Geometric Measurement, and Geometric Properties with Equations (Cat2)	22	0.93	1		
	Modeling with Geometry (Cat3)	9	0.95	0.99	1	
Biology 1	Molecular and Cellular Biology (Cat1)	20	1			
	Classification, Heredity, and Evolution (Cat2)	14	0.80	1		
	Organisms, Populations, and Ecosystems (Cat3)	22	0.86	0.82	1	
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	17	1			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	18	0.92	1		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	17	0.88	0.88	1	
Civics	Origins and Purposes of Law and Government (Cat1)	12	1			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	12	0.84	1		
	Government Policies and Political Processes (Cat3)	12	0.82	0.84	1	
	Organization and Function of Government (Cat4)	12	0.81	0.80	0.84	1

Convergent and Discriminant Validity

According to Standard 1.14 of the Standards for Educational and Psychological Testing (AERA, APA, and NCME, 1999), it is necessary to provide evidence of convergent and discriminant validity evidence. It is a part of validity evidence demonstrating that assessment scores are related as expected with criterion and other variables for all student groups. However, a second, independent test measuring the same constructs as Mathematics, ELA, Science, and EOC assessments in the State of Florida, which could easily permit for a cross-test set of correlations, was not available. Therefore, the correlations between subscores within and across Mathematics, ELA, and Science were examined alternatively. The a-priori expectation is that subscores within the same subject (e.g., Mathematics) will correlate more positively than subscore correlations

across subjects (e.g., Mathematics and ELA). These correlations are based on a small number of items (e.g., typically around 8 to 12); as a consequence, the observed score correlations will be smaller in magnitude as a result of the very large measurement error at the subscore level. For this reason, both the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within subject and across subjects for Grade 3–8 Mathematics and ELA. In general, the pattern is consistent with the a-priori expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct with a few small notes on the Writing dimensions. The Writing dimensions are based on a single essay, which is scored as three test items. Hence, the correlations between Writing and other dimensions, both in the observed score and even disattenuated scores, are somewhat unstable given the large measurement error. Table 66 through Table 77 show the observed and disattenuated score correlations between Mathematics and ELA subscores for Grades 3–8, where students took both subjects.

Table 66: Grade 3 Observed Score Correlations

Subject	Reporting Category	Mathematics			ELA			
		Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep4
Mathematics	Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1)	1.00	0.76	0.85	0.62	0.65	0.63	0.56
	Numbers and Operations – Fractions (Cat2)		1.00	0.77	0.58	0.60	0.58	0.51
	Measurement, Data, and Geometry (Cat3)			1.00	0.66	0.68	0.65	0.57
ELA	Key Ideas and Details (Cat1)				1.00	0.81	0.77	0.60
	Craft and Structure (Cat2)					1.00	0.79	0.62
	Integration of Knowledge and Ideas (Cat3)						1.00	0.59
	Language and Editing Task (Cat4)							1.00

Table 67: Grade 3 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics			ELA			
		Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep4
Mathematics	Operations, Algebraic Thinking, and Numbers in Base Ten (Cat1)	1.00	0.91	0.96	0.74	0.75	0.76	0.77
	Numbers and Operations – Fractions (Cat2)		1.00	0.95	0.75	0.76	0.77	0.77
	Measurement, Data, and Geometry (Cat3)			1.00	0.80	0.81	0.81	0.81
ELA	Key Ideas and Details (Cat1)				1.00	1.00	1.00	0.88
	Craft and Structure (Cat2)					1.00	1.00	0.89
	Integration of Knowledge and Ideas (Cat3)						1.00	0.89
	Language and Editing Task (Cat4)							1.00

Table 68: Grade 4 Observed Score Correlations

Subject	Reporting Category	Mathematics				ELA				
		Rep1	Rep2	Rep3	Rep4	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Operations and Algebraic Thinking (Cat1)	1.00	0.83	0.80	0.81	0.69	0.65	0.69	0.54	0.55
	Numbers and Operations in Base Ten (Cat2)		1.00	0.81	0.81	0.65	0.62	0.66	0.53	0.54
	Numbers and Operations – Fractions (Cat3)			1.00	0.82	0.63	0.60	0.64	0.50	0.50
	Measurement, Data, and Geometry (Cat4)				1.00	0.67	0.64	0.67	0.53	0.52
ELA	Key Ideas and Details (Cat1)					1.00	0.78	0.80	0.57	0.57
	Craft and Structure (Cat2)						1.00	0.76	0.56	0.55
	Integration of Knowledge and Ideas (Cat3)							1.00	0.56	0.56
	Language and Editing Task (Cat4)								1.00	0.49
	Text-Based Writing (Cat5)									1.00

Table 69: Grade 4 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics				ELA				
		Rep1	Rep2	Rep3	Rep4	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Operations and Algebraic Thinking (Cat1)	1.00	0.99	0.95	0.96	0.83	0.81	0.85	0.80	0.65
	Numbers and Operations in Base Ten (Cat2)		1.00	0.95	0.95	0.79	0.77	0.82	0.77	0.63
	Numbers and Operations – Fractions (Cat3)			1.00	0.96	0.76	0.74	0.78	0.72	0.58
	Measurement, Data, and Geometry (Cat4)				1.00	0.80	0.79	0.82	0.76	0.61
ELA	Key Ideas and Details (Cat1)					1.00	0.98	1.00	0.85	0.69
	Craft and Structure (Cat2)						1.00	0.98	0.85	0.68
	Integration of Knowledge and Ideas (Cat3)							1.00	0.85	0.68
	Language and Editing Task (Cat4)								1.00	0.70
	Text-Based Writing (Cat5)									1.00

Table 70: Grade 5 Observed Score Correlations

Subject	Reporting Category	Mathematics			ELA					Science			
		Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4
Mathematics	Operations, Algebraic Thinking, and Fractions (Cat1)	1.00	0.84	0.82	0.65	0.60	0.58	0.60	0.53	0.66	0.67	0.66	0.62
	Numbers and Operations in Base Ten (Cat2)		1.00	0.82	0.65	0.60	0.58	0.60	0.54	0.65	0.68	0.68	0.62
	Measurement, Data, and Geometry (Cat3)			1.00	0.68	0.62	0.60	0.61	0.53	0.68	0.71	0.71	0.66
ELA	Key Ideas and Details (Cat1)				1.00	0.78	0.75	0.69	0.58	0.73	0.70	0.74	0.71
	Craft and Structure (Cat2)					1.00	0.71	0.64	0.56	0.67	0.65	0.68	0.66
	Integration of Knowledge and Ideas (Cat3)						1.00	0.62	0.53	0.65	0.63	0.67	0.64
	Language and Editing Task (Cat4)							1.00	0.53	0.63	0.60	0.63	0.59
	Text-Based Writing (Cat5)								1.00	0.52	0.50	0.53	0.48
Science	Nature of Science (Cat1)									1.00	0.70	0.74	0.70
	Earth and Space Science (Cat2)										1.00	0.77	0.76
	Physical Science (Cat3)											1.00	0.76
	Life Science (Cat4)												1.00

Table 71: Grade 5 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics			ELA					Science			
		Rep1	Rep2	Rep3	Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4
Mathematics	Operations, Algebraic Thinking, and Fractions (Cat1)	1.00	0.97	0.96	0.77	0.76	0.76	0.79	0.60	0.83	0.82	0.79	0.76
	Numbers and Operations in Base Ten (Cat2)		1.00	0.97	0.76	0.76	0.76	0.80	0.62	0.82	0.83	0.81	0.77
	Measurement, Data, and Geometry (Cat3)			1.00	0.81	0.80	0.80	0.82	0.62	0.87	0.88	0.86	0.82
ELA	Key Ideas and Details (Cat1)				1.00	1.00	1.00	0.93	0.68	0.93	0.87	0.90	0.89
	Craft and Structure (Cat2)					1.00	1.00	0.94	0.70	0.93	0.87	0.90	0.88
	Integration of Knowledge and Ideas (Cat3)						1.00	0.93	0.69	0.93	0.87	0.90	0.89
	Language and Editing Task (Cat4)							1.00	0.70	0.90	0.84	0.87	0.82
	Text-Based Writing (Cat5)								1.00	0.65	0.61	0.63	0.58
Science	Nature of Science (Cat1)									1.00	0.94	0.97	0.93
	Earth and Space Science (Cat2)										1.00	0.97	0.98
	Physical Science (Cat3)											1.00	0.97
	Life Science (Cat4)												1.00

Table 72: Grade 6 Observed Score Correlations

Subject	Reporting Category	Mathematics					ELA				
		Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Ratio and Proportional Relationships (Cat1)	1.00	0.76	0.67	0.69	0.72	0.60	0.59	0.56	0.51	0.43
	Expressions and Equations (Cat2)		1.00	0.74	0.75	0.79	0.66	0.66	0.62	0.56	0.48
	Geometry (Cat3)			1.00	0.66	0.71	0.53	0.53	0.51	0.45	0.40
	Statistics and Probability (Cat4)				1.00	0.70	0.64	0.64	0.60	0.53	0.45
	The Number System (Cat5)					1.00	0.61	0.61	0.58	0.55	0.48
ELA	Key Ideas and Details (Cat1)						1.00	0.78	0.70	0.60	0.49
	Craft and Structure (Cat2)							1.00	0.69	0.60	0.49
	Integration of Knowledge and Ideas (Cat3)								1.00	0.55	0.45
	Language and Editing Task (Cat4)									1.00	0.48
	Text-Based Writing (Cat5)										1.00

Table 73: Grade 6 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics					ELA				
		Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Ratio and Proportional Relationships (Cat1)	1.00	0.99	0.93	0.98	0.97	0.80	0.81	0.85	0.76	0.56
	Expressions and Equations (Cat2)		1.00	0.93	0.96	0.98	0.81	0.82	0.86	0.77	0.56
	Geometry (Cat3)			1.00	0.90	0.94	0.70	0.71	0.75	0.66	0.50
	Statistics and Probability (Cat4)				1.00	0.94	0.85	0.86	0.89	0.79	0.58
	The Number System (Cat5)					1.00	0.78	0.80	0.83	0.78	0.59
ELA	Key Ideas and Details (Cat1)						1.00	1.00	0.99	0.85	0.59
	Craft and Structure (Cat2)							1.00	1.00	0.86	0.61
	Integration of Knowledge and Ideas (Cat3)								1.00	0.87	0.61
	Language and Editing Task (Cat4)									1.00	0.65
	Text-Based Writing (Cat5)										1.00

Table 74: Grade 7 Observed Score Correlations

Subject	Reporting Category	Mathematics					ELA				
		Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Ratio and Proportional Relationships (Cat1)	1.00	0.77	0.74	0.69	0.73	0.65	0.62	0.60	0.55	0.50
	Expressions and Equations (Cat2)		1.00	0.74	0.69	0.74	0.60	0.56	0.55	0.50	0.44
	Geometry (Cat3)			1.00	0.68	0.71	0.57	0.55	0.53	0.49	0.44
	Statistics and Probability (Cat4)				1.00	0.66	0.58	0.55	0.53	0.48	0.41
	The Number System (Cat5)					1.00	0.58	0.55	0.54	0.50	0.44
ELA	Key Ideas and Details (Cat1)						1.00	0.77	0.72	0.63	0.54
	Craft and Structure (Cat2)							1.00	0.69	0.62	0.53
	Integration of Knowledge and Ideas (Cat3)								1.00	0.57	0.50
	Language and Editing Task (Cat4)									1.00	0.51
	Text-Based Writing (Cat5)										1.00

Table 75: Grade 7 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics					ELA				
		Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4	Rep5
Mathematics	Ratio and Proportional Relationships (Cat1)	1.00	0.97	0.95	0.94	0.96	0.79	0.79	0.81	0.77	0.59
	Expressions and Equations (Cat2)		1.00	0.97	0.96	1.00	0.75	0.73	0.76	0.72	0.53
	Geometry (Cat3)			1.00	0.96	0.97	0.73	0.73	0.74	0.72	0.53
	Statistics and Probability (Cat4)				1.00	0.96	0.78	0.77	0.79	0.75	0.53
	The Number System (Cat5)					1.00	0.75	0.74	0.77	0.74	0.55
ELA	Key Ideas and Details (Cat1)						1.00	0.98	0.96	0.88	0.64
	Craft and Structure (Cat2)							1.00	0.96	0.90	0.65
	Integration of Knowledge and Ideas (Cat3)								1.00	0.87	0.64

Subject	Reporting Category	Mathematics					ELA				
		Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4	Rep5
ELA	Language and Editing Task (Cat4)									1.00	0.68
	Text-Based Writing (Cat5)										1.00

Table 76: Grade 8 Observed Score Correlations

Subject	Reporting Category	Mathematics				ELA					Science			
		Rep1	Rep2	Rep3	Rep4	Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4
Mathematics	Expressions and Equations (Cat1)	1.00	0.71	0.72	0.62	0.49	0.49	0.45	0.41	0.36	0.51	0.54	0.56	0.53
	Functions (Cat2)		1.00	0.67	0.58	0.48	0.48	0.44	0.41	0.36	0.49	0.51	0.53	0.51
	Geometry (Cat3)			1.00	0.63	0.47	0.48	0.43	0.41	0.36	0.49	0.53	0.55	0.53
	Statistics and Probability and the Number System (Cat4)				1.00	0.46	0.47	0.41	0.41	0.38	0.46	0.52	0.52	0.53
ELA	Key Ideas and Details (Cat1)					1.00	0.74	0.63	0.52	0.44	0.61	0.61	0.58	0.61
	Craft and Structure (Cat2)						1.00	0.64	0.54	0.46	0.61	0.62	0.59	0.62
	Integration of Knowledge and Ideas (Cat3)							1.00	0.46	0.38	0.53	0.54	0.51	0.54
	Language and Editing Task (Cat4)								1.00	0.46	0.44	0.49	0.47	0.50
	Text-Based Writing (Cat5)									1.00	0.37	0.42	0.40	0.43
Science	Nature of Science (Cat1)										1.00	0.67	0.64	0.66
	Earth and Space Science (Cat2)											1.00	0.72	0.73
	Physical Science (Cat3)												1.00	0.71
	Life Science (Cat4)													1.00

Table 77: Grade 8 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics				ELA					Science			
		Rep1	Rep2	Rep3	Rep4	Rep1	Rep2	Rep3	Rep4	Rep5	Rep1	Rep2	Rep3	Rep4
Mathematics	Expressions and Equations (Cat1)	1.00	0.96	0.91	0.85	0.62	0.61	0.60	0.60	0.44	0.66	0.68	0.71	0.67
	Functions (Cat2)		1.00	0.91	0.85	0.64	0.64	0.63	0.62	0.45	0.68	0.68	0.71	0.68
	Geometry (Cat3)			1.00	0.87	0.60	0.60	0.58	0.60	0.44	0.64	0.67	0.70	0.67
	Statistics and Probability and the Number System (Cat4)				1.00	0.63	0.63	0.60	0.63	0.49	0.64	0.71	0.72	0.71
ELA	Key Ideas and Details (Cat1)					1.00	0.92	0.83	0.75	0.52	0.78	0.77	0.74	0.76
	Craft and Structure (Cat2)						1.00	0.84	0.77	0.54	0.77	0.77	0.74	0.77
	Integration of Knowledge and Ideas (Cat3)							1.00	0.69	0.48	0.72	0.72	0.69	0.71
	Language and Editing Task (Cat4)								1.00	0.63	0.65	0.70	0.67	0.71
	Text-Based Writing (Cat5)									1.00	0.46	0.50	0.48	0.51
Science	Nature of Science (Cat1)										1.00	0.86	0.83	0.84
	Earth and Space Science (Cat2)											1.00	0.90	0.92
	Physical Science (Cat3)												1.00	0.89
	Life Science (Cat4)													1.00

4.2.5 Implication Validity Evidence

The Standards (AERA, APA, and NCME, 2014) suggests that test-criterion relationships belong to the source of validity evidence of the relationship of test scores to external variables. The 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.

There are inferences made at different levels based on the Florida Statewide Assessments. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others. For example, the Science assessments report individual student scores, but some students may feel that few ramifications of the test directly affect them; such students may fail to put forth their full effort. Although this report documents in detail evidence showing that the Science assessment is a valid measure of student achievement on the Standards, individual and school-level scores are not valid if students do not take the test seriously. The incorporation of graduation requirements associated with the Grade 10 Reading and Algebra 1 assessments increases the consequences of the test for high school students; this may mitigate concerns about student motivation affecting test validity. Also, as students are made fully aware of the potential Every Student Succeeds Act (ESSA) ramifications of the test results for their school, this threat to validity should diminish.

One of the most important inferences to be made concerns the student's achievement level, especially for accountability tests. Even if the total-correct score can be validated as an appropriate measure of the standards, it is still necessary that the scaling and achievement-level designation procedures be validated. Because scaling and standard setting are both critical processes for the success of Florida Statewide Assessments, separate volumes are devoted to them. Volume 3 of the *Florida Standards Assessments 2014–2015 Technical Report* discusses the details concerning performance standards, and Volume 1 of this technical report discusses scaling. These volumes serve as documentation of the validity argument for these processes.

At the aggregate level (i.e., school, district, or statewide), the implication validity of school accountability assessments can be judged by the impact the testing program has on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage of students classified as proficient. As mentioned before, there exists a potential for negative impacts on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects as well.

Summary of Validity Evidence

Florida Statewide Assessments scores provide information reflecting what students know and can do in relation to the academic expectations. They are summative measures of a student's performance in a subject at one point in time. They provide a snapshot of the student's overall achievement, not a detailed accounting of the student's understanding of specific content areas defined by the standards. However, the scores help parents begin to understand their child's academic performance as it relates to Florida Statewide Assessments, they provide information to

educators and suggest areas needing further evaluation of student performance. The results can also be used for intervention needed for students struggling with Florida Statewide Assessments. In addition to being helpful in evaluating the strengths and weaknesses of a particular academic program or curriculum, the test results can be used to answer a variety of questions about a student, educational program, school, or district. It is important to be cautious for interpretation of score use, such as understanding measurement error, using scores at extreme ends of distributions, interpreting score means, using reporting category information, and program evaluation implications. Chapter 5 of Volume 6 of the *Florida Statewide Assessments 2021–2022 Technical Report* narrated the details in cautions of score use.

This volume as well as other volumes of this technical report provide validity evidence supporting the appropriate inferences from Florida Statewide Assessments scores. In general, the validity evidence provides supports to the primary claim that Florida Statewide Assessments scores provide information reflecting what students know and can do in relation to the academic expectations defined in terms of academic content and achievement standards. Validity arguments based on rationale and logic are strongly supported for Florida Statewide Assessments. The empirical validity evidence for the scoring and the generalization validity arguments for these assessments are also quite strong. Reliability indices, model fit, and dimensionality studies provide consistent results, indicating the Florida Statewide Assessments are properly scored, and scores can be generalized to the universe score.

5. EVIDENCE OF COMPARABILITY

As the Florida Statewide Assessments were administered in multiple modes (both online and paper-based), 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. Florida Statewide Assessments had online assessments for Grades 7 through 10 ELA, Grades 7 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 Grade 3 Reading, Grades 4 through 6 ELA, and Grades 3 through 6 Mathematics, 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 78. In EOC, the first core form (Core 24 for Algebra 1 and Core 19 for Geometry) was administered as the accommodated version. In the accommodated forms, 5 items in the Algebra 1 Core 24 form and 3 items in Geometry Core 19 form were replaced with items that are feasible to administer in paper.

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

For the 2021–2022 Florida Statewide Assessments, the paper-based 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 *2021–2022 Florida Statewide Assessments Technical Report.*

5.2 COMPARABILITY OF FLORIDA STATEWIDE ASSESSMENTS TEST SCORES OVER TIME

The comparability of Florida Statewide Assessments scores over time was ensured via two methods. First, during test construction, both a content and statistical perspective were implemented. All test items in both FSA and NGSSS were placed onto forms were aligned to the same standards and test blueprint specifications. In addition, spring 2021 form statistics were used as targets for both numerical and graphical summaries for the spring 2022 forms. See Section 4 of Volume 2 of this technical report for details about both the content and statistical methods. Second, during spring 2022 calibrations, equating was performed in order to place item parameters estimates from spring 2022 onto IRT equated bank scale. The equating procedure and results are presented in Volume 1, Section 6.2.

5.3 COMPARABILITY OF ONLINE AND ACCOMMODATED TEST SCORES

In a review of literature on the issue of score comparability between online and accommodated (paper-based) forms, DePascale, Dadey, & Lyons (2016) cites Winter (2010) on the definition for

score comparability. Specifically, Winter (2010) notes that comparability requires that a test and its variations must

- measure the same set of knowledge and skills at the same level of content-related complexity (i.e., comparable constructs);
- produce scores at the desired level (i.e., type) of specificity that reflect the same degree of achievement on those constructs (i.e., comparable scores); and
- have similar technical properties in relation to the level of score reported (i.e., comparable technical properties of scores).

Note that variations of a form refer not only to the online versus paper or accommodated distinction, but also to online tests administered across devices and platforms.

Paper-based test forms were offered as a special accommodation for students who qualified, according to their Individual Educational Plans (IEP) or Section 504 Plans, in place of online test forms for Grades 7 and 8 Mathematics, Grades 7 through 10 ELA, and EOC. 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-based 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 (TEI) 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 several items were replaced in the accommodated forms. However, this is not concerning since all of the items are on the same IRT scale. In Mathematics EOC, TCCs for the accommodated forms are above those of the online forms, slightly shifted upward compared to the TCCs for the online forms. As seen in Table 78, several TEI were replaced in accommodated forms of Grade 7 and 8 Mathematics, Algebra 1, and Geometry. All ELA and NGSSS EOC accommodated form items were the same as online form items. TCCs for Mathematics forms which have replaced items in accommodated forms are presented in Appendix D, Test Characteristic Curves.

Table 78: Number of Item Replacements for the Accommodated Forms

Mathematics	Number of Items Replaced
Grade 7	10
Grade 8	11
Algebra 1	5
Geometry	3

A device comparability study was conducted to provide evidence of the comparability of the FSA across the most frequently used platforms. Score comparability across different devices can be examined to assess whether student performance on the Florida Statewide Assessments differs between students conditional on the device. The device effects were examined via regression and a likelihood ratio test to compare the regression models. The study showed that there are no

systematic differences in the scores for students when administered the Florida Statewide Assessments on different devices. The details of the study can be found in Appendix F of this volume of the *Florida Statewide Assessments 2021–2022 Technical Report* (previously Appendix F of Volume 4 of the *Florida Standards Assessments 2017–2018 Technical Report).*

5.4 COMPARABILITY OF CONSTRUCTS

To make a claim about comparable constructs, as Winter (2010) suggests, it is important to provide evidence to show that (1) assessed content should be comparable across different versions of the assessment; and (2) testing administration devices do not introduce construct-irrelevant variance into score estimates. In the following, evidence is summarized that shows how Florida has applied the known findings in the research literature and followed best practices in the field to minimize construct irrelevant variance and reduce threats to score comparability during test design, development, and administration.

When an online test is converted to paper, the following steps occur and are documented in the Florida Statewide Assessments technical report. First and foremost, the paper version is constructed to the exact same test specifications and, in many cases, the items between the online and paper forms are the same. Some technology-enhanced items are replaced on the paper versions with items intended to render on paper. They are chosen to essentially mirror the online items they are replacing such that the paper form measures the same construct in a similar way. NGSSS tests have multiple-choice items only, so the paper versions used for accommodations contain items that are identical to items on the online forms and are presented in the same format, as well.

5.5 COMPARABILITY OF SCORES

As described in this technical report, Florida tests use maximum likelihood estimation for scoring and report scale scores, performance levels, and reporting category scores. This applies to all versions of the assessment (e.g., online, paper, with and without accommodations). To ensure that paper accommodated forms produce comparable scores comparable to online forms, we conducted equating to place paper accommodated forms onto the IRT calibrated item pool. This process is described in this technical report. The essence is that the paper accommodated items that are common with the online form use item parameters from the online calibrations, and all other items use item parameters from previous online administrations. Since both online and paper accommodated forms are scored using the same IRT calibrated item pool and the forms are statistically parallel, the scores obtained from taking paper accommodated form are comparable to those obtained from the online form.

As for research on score comparability, some important messages emerged from a comprehensive review of literature by DePascale, Dadey, & Lyons (2016): (1) the majority of comparability studies have found their computer and paper-based tests to be comparable overall (e.g., Davis, Kong, & McBride, 2016; Davis, Orr, Kong, & Lin, 2015); (2) Research on device comparability shows a generally high degree of score comparability across digital devices on large-scale assessments, and factors that may potentially contribute to the presence of device effects include familiarity, device features (e.g., screen size, input mechanism, keyboard), and assessment-specific features (e.g., content area); and (3) there are clear, practical steps throughout the assessment cycle

that states and their assessment contractors can take to be proactive in identifying, anticipating, and avoiding potential threats to score comparability due to devices.

As described in Section 5.4, Comparability of Constructs, numerous processes have been implemented in the design, development, and administration of Florida assessments that mirror best practices recommended by research. As an empirical check, we also conducted a study to investigate comparability of the Florida Statewide Assessments across the most frequently used platforms. The details of the study can be found in the Appendix F of this volume of the *Florida Statewide Assessments 2021–2022 Technical Report* (previously the Appendix F of Volume 4 of the *Florida Standards Assessments 2017–-2018 Technical Report)*. As expected, the study showed no systematic differences in the scores for students when administered the Florida Statewide Assessments on different devices. This lends confirmatory evidence from empirical data that the processes implemented seem to be effective in minimizing threats to score comparability across devices.

5.6 COMPARABILITY OF TECHNICAL PROPERTIES OF SCORES

For state-mandated accountability assessments, score comparability almost invariably refers to comparability of scale scores. This is true for Florida assessments, as we expect scale scores from different versions of the assessment to be used interchangeably. Given that scale scores are at a finer grain size than achievement-level classifications, showing the comparability of scale scores implies that aggregate scores or classifications derived from them, like performance levels, are also comparable (DePascale, Dadey, & Lyons, 2016). In the following, we provide evidence that technical properties of scale scores are comparable between online and paper accommodated assessments.

6. FAIRNESS AND ACCESSIBILITY

6.1 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
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.

Section 2.1 in Volume 5 of this technical report discusses unique accommodations, appropriate accommodations, appropriate selection and use of accommodations, and appropriate implementation of accommodations in Florida Statewide Assessments.

The use of alternative formats and accommodations for individuals with visual disabilities raises concerns about fairness and validity. Due to the small sample sizes associated with visually impaired students with disabilities, it is not feasible to conduct empirical analyses based on Florida data to investigate the effects of this accommodation. Therefore, we rely on research findings in the literature for this investigation. In a review of literature in Shaftel et al (2015), it seems that findings were mixed on DIF research with respect to visually-impaired students. Zebehazy, Zigmond, & Zimmerman (2012) investigated DIF of test items on Pennsylvania's Alternate System of Assessment (PASA) for students with visual impairments and results indicated DIF among the functional vision groups when compared to a matched group of sighted students. By contrast, Stone, Cook, Laitusis, & Cline (2010) conducted a similar study and found only one item at each grade showed large DIF favoring students without visual impairments, supporting the accessibility and validity of alternate formats for students with visual disabilities. Shaftel et al (2015) conducted DIF research comparing students with and without disabilities and concluded that results were encouraging in terms of demonstrating that the different item types, when designed and developed with accessibility in mind, did not disadvantage any particular student group.

6.2 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 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 of this technical report. 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, female), or negatively (i.e., –A, –B, or–C), signifying that the item favored the reference group (e.g., White, 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, of this technical report.

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 *2021–2022 Florida Statewide Assessments Technical Report*. The DIF statistics for each test item are presented in the appendices of Volume 1 of the *2021–2022 Florida Statewide Assessments Technical Report.*

6.3 SUMMARY

This volume, as well as other volumes of this technical report, is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. In general, the validity evidence provides support to the primary claim that Florida Statewide Assessment scores provide information reflecting what students know and can do in relation to the academic expectations defined in terms of academic content and achievement standards.

The overall results of this volume 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.
- **Comparability.** Evidence is provided to support score comparability across forms over time and between online and paper accommodated forms.
- **Test Fairness.** Evidence is provided to support test fairness based on content alignment reviews and statistical analysis.

7. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), and 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.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS-RR-01-23). Princeton, NJ: Educational Testing Service.

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., Bollen, K., Paxton, P., Curran P., & Kirby, J. (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.

Cronbach, L. J., & Meehl, P. E. (1955*). Construct validity in psychological tests. Psychological Bulletin*, 52(4), 281–302.

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th Ed.), Harper & Row, NY.

Davis, L.L., Kong, X., & McBride, M. (2015, April). *Device comparability of tablets and computers for assessment purposes.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Davis, L. L., Kong, X., McBride, Y., & Morrison, K. (2016). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education. 30*(1), 16-26.

Davis, L.L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment*, 20, 180-198.

Davis, L. L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs*. Retrieved from https://docplayer.net/19176443-Testing-on-tablets-part-ii-of-a-series-of-usability-studies-on-the-use-of-tablets-for-k-12-assessment-programs.html

Davis, Morrison, Kong, & McBride (2017). Disaggregated Effects of Device on Score Comparability. *Educational Measurement: Issues and Practice, 36*, 35–45.

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Council of Chief State School Officers, Retrieved from https://ccsso.org/sites/default/files/2018-07/CCSSO%20TILSA%20Score%20Comparability%20Across%20Devices.pdf

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.*

Florida Department of Education. (2015). *Florida Standards Assessments 2014-2015 Technical Report.*

Florida Department of Education. (2019). Florida Statewide Science and EOC Assessments 2019 Technical Report: Performance Standards

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

Hu, L. T., & 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.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Keng, L., Kong, X.J., & Bleil, B. (2011). *Does size matter? A study on the use of netbooks in K-12 assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Lazarus, S. S., & Thurlow, M. L. (2016). *2015-16 high school assessment accommodations policies: An analysis of ACT, SAT, PARCC, and Smarter Balanced (NCEO Report 403).*

Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.info/Resources/publications/OnlinePubs/Report403/default.html

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.

Lee, W.C., Hanson, B.A., & Brennan, R.L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26(4)*, 412–432.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessing in teaching* (7th ed.). New Jersey: Prentice-Hall Inc.

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., & 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 https://www.p12.nysed.gov/irs/ela-math/

Olsen, J. B. (2014). Score comparability for web and iPad delivered adaptive tests. Paper presented at the meeting of the National Council on Measurement in Education, Philadelphia, PA.

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.

Reboussin, B. A., & Liang, K. Y. (1998). An estimating equations approach for the liscomp model. *Psychometrika*, *63*(2), 165-182. https://doi.org/10.1007/BF02294773

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

Rudner, L. M. (2005) Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13), 1–4.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., and Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series (NCES 2005–457).* U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from: http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf

Shaftel, J., Benz, S., Boeth, E., Gahm, J., He, D, Loughran, J., Mellen, M. Meyer, E., Minor, E., & Overland, E. (2015). *Accessibility for Technology-Enhanced Assessments (ATEA) Report of Project Activities*. Research Report. University of Kansas.

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

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *PARCC spring 2015 digital devices comparability research study.* Retrieved from https://files.eric.ed.gov/fulltext/ED599032.pdf

Stone, E., Cook, L., Laitusis, C. C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education*, 23, 132-152.

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 https://nceo.umn.edu/docs/onlinepubs/synth44.pdf.

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.

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement* (pp. 260-284). New York: Routledge.

Winter, P. (2010). Comparability and test variations. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations (pp. 1-11).* Washington, DC: Council of Chief State School Officers.

Zebehazy, K. T., Zigmond, N., & Zimmerman, G. J. (2012). Ability or access-ability: differential item functioning of items on alternate performance-based assessment tests for students with visual impairments. *Journal of Visual Impairment & Blindness*, 106, 325-338.