State of Florida

Florida Assessment of Student Thinking (FAST), Benchmarks for Excellent Student Thinking (B.E.S.T.), and Science & Social Studies Statewide Assessments

2023–2024

Volume 4
Evidence of Reliability and Validity

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from Cambium Assessment, Inc.: Dr. Myvan Bui, Dr. Sherry Li, Dr. Yuan Hong, Dr. Peter Diao, Dr. Susan Lottridge, Dr. Amy Burkhardt, Matt Gordon, Zoe Dai, and Oliver Brown. The major contributors from the FDOE are as follows: Vince Verges, Susie Lee, Racquel Harrell, Sally Donnelly, Dr. Shakia Johnson, Kristina Lamb, Jenny Black, Dr. Esra Kocyigit, Dr. Salih Binici, Wenyi Li, Jielin Ming, Saeyan Yun, and Yiting Yao.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. Introduction and Overview of Reliability and Validity Evidence

During the 2022–2023 school year, Florida transitioned from the fixed-form Florida Standards Assessments (FSA) to the computer-adaptive FAST (Florida Assessment of Student Thinking), and B.E.S.T. (Benchmarks for Excellent Student Thinking) assessments. The FSA previously replaced the Florida Comprehensive Assessment Tests (FCAT) 2.0 in English language arts (ELA) and mathematics during the 2014–2015 school year. FAST is administered as a progress monitoring assessment and includes Voluntary Prekindergarten (VPK) through grade 10 ELA and VPK through grade 8 mathematics assessments. B.E.S.T. assessments that are not part of the FAST progress monitoring program include grades 4–10 writing and end-of-course (EOC) assessments in Algebra 1 and Geometry. The science and social studies assessments include science grades 5 and 8, and end-of-course (EOC) assessments for Biology 1, Civics, and US History. These transitioned to computer-adaptive assessments during the 2023–2024 school year. They are also not part of the FAST progress monitoring program.

For FAST progress monitoring assessments, students participate three times per year: in this case, once at the beginning of the year (PM1, August 14–September 29, 2023), once in the middle of the year (PM2, December 4, 2023–January 26, 2024), and once at the end of the year (PM3, May 1–May 31, 2024).

- PM1 is designed to provide a baseline score so teachers can track student progress in learning the B.E.S.T. standards from PM1 to PM2.
- PM2 occurs after an opportunity to learn the grade-level standards. This test administration provides a mid-year score to compare to the baseline score from PM1.
- PM3 produces summative scores that will accurately measure student mastery of the B.E.S.T. standards at the end of the school year. While PM1 and PM2 are for informational purposes only, PM3 is used for school accountability in grade 3 and higher beginning with the 2023–2024 school year. Assessments in grades pre-K–2 are not currently part of the state's accountability system.

This technical report describes the FAST assessments for grades 3–10 ELA and grades 3–8 mathematics, B.E.S.T. assessments, and science and social studies assessments. The details of the VPK to grade 2 assessments in reading and mathematics are provided in Renaissance's Star Assessments for Math, Reading, and Early Literacy Technical Manuals.

In addition to the online computer-adaptive test (CAT), Florida also has accommodated forms. Accommodated forms were administered to students in lieu of the online forms if such a need was indicated on their Individualized Education Program (IEP) or Section 504 Plan. Accommodated forms used online parameters for scoring purposes, and no calibrations were performed on the accommodated forms. For each grade, one accommodated form was given. Additional accommodations guidelines can be found in Volume 5 of this technical report.

Table 1 displays the complete list of tests for the spring operational administration. DEI stands for Data Entry Interface and is used for grades 3–10 FAST accommodated ELA and mathematics assessments, as well as the science and social studies assessments. When a paper-based version is provided as an accommodation, student responses from the paper-based tests were transcribed into the DEI to ensure timely results. TTS stands for text-to-speech and is used for the science and social studies assessments. It is a part of the Test Delivery System (TDS) and allows for authorized individuals to submit answers for students for immediate reporting.

Table 1: Test Administration

Subject	Administration	Grade/Course
ELA Reading	Online	3–10
	DEI (Accommodated)	
Mathematics	Online	3–8
	DEI (Accommodated)	
Science	Online	5, 8
	DEI (Accommodated)	
	TTS (Accommodated)	
Mathematics EOC	Online	Algebra 1, Geometry
	DEI (Accommodated)	
Science & Social Studies EOC	Online	Biology, Civics, U.S. History
	DEI (Accommodated)	
	TTS (Accommodated)	

With the implementation of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from Florida's assessment scores. This volume provides empirical evidence about the reliability and validity of the spring assessments given their intended uses.

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

- **Reliability.** The precision of individual test scores is critically important to valid test score interpretation and is provided along with test scores as part of overall and subscale-level reporting. The precision of test scores varies with respect to the information value of the test at each ability location. Marginal reliability was computed in order to take into account the varying measurement errors across ability ranges. The reliability estimates are presented by grade and subject as well as by demographic subgroup. This section also includes conditional standard errors of measurement (CSEMs) 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 the 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.

- **Comparability Evidence.** By examining the blueprint match between forms administered by the CAT and accommodated forms, and test characteristic curves (TCCs), 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's Statewide Assessments

The Florida's statewide, standardized assessments are standards-based, summative assessments 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* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) and to ensure that all students have access to the test content via principles of universal design and appropriate accommodations.

The 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.

Florida's statewide assessment results serve as the primary indicator for the state's accountability system. The policy and legislative purpose of the assessments are described more thoroughly in Volume 1 of this technical report. The tests are standards-based assessments designed to measure student achievement toward the state content standards. The 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 English language arts (ELA) (scores inform district promotion decisions), grade 10 ELA, and Algebra 1 (assessment graduation requirements), the assessment is never the sole determinant in making these decisions.

For the adaptive tests, simulation reports were examined to track the compliance of the test structure to the assessment requirements. For accommodated fixed forms, test items were selected prior to the test administration to ensure that the test construction aligned to the approved blueprint.

The FAST and B.E.S.T. performance cuts were approved by the State Board of Education (SBE) on October 18, 2023. These FAST and B.E.S.T. cut scores, approved by SBE, scale scores, and achievement levels were used in spring 2024. Volume 3 of the *Florida B.E.S.T. 2022–2023 Technical Report* describes the standard setting and how each of these cut scores was set. The cut scores of grades 5 and 8 science and Biology 1 were approved by the State Board of Education in 2012, and the cut scores of U.S. History and Civics were approved in 2013 and 2014, respectively. Chapter 5: Performance Standards from the *Florida Statewide Science and EOC Assessments 2019 Technical Report* describes the standard setting and cut score establishment for science and social studies. These volumes are not updated annually as they do not contain activities that change based on the administration year.

Volume 1 of this technical report, Section 7 Scoring, describes how the scoring is performed and how the cut scores are used in scoring. Student-level scores included scale scores at the overall and reporting category level. The scale scores for reporting categories were used to indicate student performance classification on each of the reporting categories. 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

Test score reliability is traditionally estimated using both classical and item response theory (IRT) approaches. Classical indicators of reliability, such as Cronbach's alpha or test-retest reliability, provide a single estimate of the reliability of test scores, assuming that reliability is constant across the entire range of scores. However, the precision of test scores can vary across different levels of the latent trait being measured. For example, most fixed-form assessments target test information near important cut scores or near the population mean so that test scores are most precise in targeted locations. Because adaptive tests target test information near each student's ability level, the precision of test scores may increase, especially for lower- and higher-ability students. The precision of individual test scores is critically important to valid test score interpretation and is provided along with test scores as part of all student-level reporting. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional standard errors of measurement (CSEMs), which are estimated at different points on the ability scale for all students.

3.1 Marginal Reliability

The regular summative and progress monitoring assessments are adaptive testing administrations. Because there is no fixed form in adaptive testing, marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range.

Marginal reliability ($\overline{\rho}$) is defined as

$$\overline{\rho} := [\overline{\sigma}^2 - \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \widehat{CSEM}_i^{\ 2}]/\overline{\sigma}^2,$$

where N is the number of students; $\widehat{CSEM}_i$ is the estimated CSEM for student i based on the Hessian at the maximum likelihood estimate (MLE) score $\hat{\theta}_i$,

$$\overline{\sigma}^2 := \frac{1}{N-1}\sum_{\mathrm{i=1}}^{\mathrm{N}} (\hat{\theta}_i - \overline{\mu})^2$$

is the estimated variance of the student theta scores $\hat{\theta}_i$, and $\overline{\mu}$ is the estimated mean of the student theta scores. The higher the reliability coefficient, the greater the precision of the test.

Table 2 to Table 6 present the marginal reliability coefficients for all students for the spring summative and PM3 tests (PM1 and PM2 test administrations are in Appendix A). The reliability coefficients for all subjects and grades range from 0.72 to 0.92 for regular forms and 0.63 to 0.87 for accommodated (including Data Entry Interface [DEI] and text-to-speech [TTS]). Appendix A: Reliability Coefficients, provides further breakdown, including reliability coefficients for demographic subgroups and reporting categories.

In the regular tests, it is noted that overall marginal reliabilities are much lower than the ideal 0.85 or higher in mathematics grades 7 and 8. High-scoring students from these groups tend to take the Algebra 1 and Geometry tests. Thus, the ability distribution in these populations is typically restricted at the upper end of the scale, depressing the reliabilities. For all tests, students are also restricted at the lower end due to the high number of students who need to be scored with the truncated lowest obtainable score allowed (see Volume 1, Section 7 Scoring), which would also contribute to lowering the reliabilities. Furthermore, the adaptive algorithm needs to cater to the lower range in PM1/PM2, so the available bank in PM3 will tend to be lower in number where students were already taking the test at their ability level in the first two administrations. The PM3 bank was limited by the items not already seen in the first two administrations. That is, there is an insufficient match between student ability and difficulty of the test item bank for all students at this early stage of item pool development.

It is also noted that some overall reliabilities for some demographic subgroups are quite low and likely due to a combination of three factors: restriction of score range ($\bar{\sigma}^2$) resulting from subgroup populations who tend to score in a narrower lower range and a higher CSEM for those same subgroups because of the mismatch of overall test difficulty to those lower scoring populations. It is expected that with more item development (and more focused item development on the easier end), reliabilities will improve in future test administrations because the adaptive algorithm will be able to select easier items for those subgroups. There are limits based on the content standards themselves, however—for example, in general, reading items will be too challenging for English language learner (ELL) students. Marginal reliabilities are sample dependent, as they are based on the observed scores. In theory, if the reliabilities had been calculated on ELL students with all abilities, the reliability would be much higher. However, by their nature, the ELL subgroup will have very restricted ELA ability range.

Marginal reliabilities are particularly low at the reporting category level and for the PM1 and PM2 test administrations in Appendix A. This is not unexpected. Each reporting category has a very small number of items (8–19) (see Volume 2, Appendix G). Furthermore, although PM1 and PM2 are considered progress monitoring tests, they administer summative-type items to students before they have had a chance to learn the material. Both factors would depress reliabilities. This mismatch between student abilities and the item difficulty distributions in the bank can be seen in Appendix F. This issue is also present for accommodated forms, especially TTS forms where the mismatch with the constructed form is very pronounced, contributing to much lower marginal reliabilities. Table 4 to Table 6 contain overall marginal reliabilities for accommodated forms, which are generally lower. The sample size for accommodated forms is smaller, which would contribute to the difference. These reported reliabilities are for the sample dependent observed thetas (abilities) rather than the theoretical marginal reliabilities (based on the test information function), and the observed tends to be lower. Further discussions about the comparability of online and accommodated forms can be found in Section 5.3, Comparability of Online and Accommodated Tests, of this volume.

In summary, for the marginal reliability issues, there are definite areas of possible improvement in the depth and breadth of the item bank (especially at the easier level). To address this, CAI and FDOE psychometric teams continuously analyze item bank characteristics and work with the Content teams in relation to future item development plans to improve item bank features, including to better match all skill levels. The bank item pools grow each year through new item field testing. Further details of the item development plan for Cambium Assessment, Inc. (CAI) are provided in Volume 2, Test Development, of this technical report.

Table 2: Marginal Reliability, ELA and Mathematics

Subject	Grade	Number of Items	Marginal Reliability	N-Count	Scale Score Mean	Scale Score SD	SEM (Mean of CSEM)
ELA Reading	3	705	0.85	215,574	200.93	22.48	7.15
	4	665	0.83	212,165	211.52	22.73	7.98
	5	671	0.88	203,412	221.64	21.80	6.80
	6	573	0.84	205,054	223.96	23.72	8.03
	7	623	0.84	214,938	228.47	25.01	8.24
	8	624	0.86	209,835	234.86	25.07	8.30
	9	652	0.86	216,621	239.87	24.17	8.01
	10	681	0.87	215,657	245.21	23.83	7.94
Mathematics	3	674	0.92	214,927	201.53	21.69	5.62
	4	674	0.91	207,096	213.46	21.59	5.65
	5	681	0.91	197,191	223.18	22.60	5.92
	6	676	0.91	194,855	229.88	21.42	5.49
	7	619	0.79	144,768	230.34	22.73	8.34
	8	635	0.72	114,710	235.03	22.96	9.99
Algebra		596	0.89	228,344	400.30	29.20	7.87
Geometry		518	0.91	213,902	402.56	27.39	6.27

Table 3: Marginal Reliability, Science and Social Studies

Subject	Number of Items	Marginal Reliability	N-Count	Scale Score Mean	Scale Score SD	SEM (Mean of CSEM)
Biology 1	1,021	0.85	199,788	406.74	29.27	9.82
Civics	866	0.85	188,377	408.36	30.83	10.00
U.S. History	864	0.85	183,226	409.52	30.19	10.21
Grade 5 Science	1,139	0.89	174,486	203.06	23.57	7.56
Grade 8 Science	951	0.85	178,331	199.70	23.61	7.87

Table 4: Marginal Reliability, Accommodated Forms, ELA and Mathematics

Subject	Grade	Marginal Reliability	N-Count	Scale Score Mean	Scale Score SD	SEM (Mean of CSEM)
ELA Reading	3	0.77	895	195.26	20.97	7.87
	4	0.73	958	202.54	21.60	8.99
	5	0.76	800	211.63	22.22	8.70
	6	0.74	573	212.87	23.42	9.40
	7	0.75	330	221.39	24.25	9.95
	8	0.81	329	228.00	23.38	8.80
	9	0.80	385	231.45	23.48	8.98
	10	0.85	397	235.09	21.86	7.58
Mathematics	3	0.87	895	192.34	23.73	7.05
	4	0.86	940	206.27	21.41	6.72
	5	0.84	805	212.82	23.84	7.39
	6	0.87	564	219.30	22.14	7.04
	7	0.71	284	221.67	27.76	12.00
	8	0.63	243	223.56	25.47	12.88
Algebra		0.65	407	386.85	34.12	16.51
Geometry		0.65	365	390.21	28.21	13.53

Table 5: Marginal Reliability, Science and Social Studies TTS

Subject	Marginal Reliability	N-Count	Scale Score Mean	Scale Score SD	SEM (Mean of CSEM)
Biology 1	0.75	15,494	386.19	27.36	11.42
Civics	0.75	26,176	386.30	27.62	11.58
U.S. History	0.74	9,359	387.63	28.15	12.11
Grade 5 Science	0.85	28,471	185.21	20.48	7.37
Grade 8 Science	0.73	23,173	183.64	20.48	8.63

Table 6: Marginal Reliability, Science and Social Studies DEI

Subject	Marginal Reliability	N-Count	Scale Score Mean	Scale Score SD	SEM (Mean of CSEM)
Biology 1	0.81	335	394.25	27.60	10.02
Civics	0.83	316	394.32	32.11	11.16
U.S. History	0.84	361	401.32	28.80	9.96
Grade 5 Science	0.87	789	189.76	21.43	7.19
Grade 8 Science	0.83	305	191.02	23.88	8.19

3.2 Standard Error of Measurement

Except for B.E.S.T. writing (raw score reported), the Florida statewide assessments are based on the three-parameter logistic (3PL) model. For ELA and mathematics, they also use the two-parameter logistic model (2PL) and generalized partial-credit model (GPCM) of IRT models. Theta scores and standard errors of measurement are generated using "pattern scoring" as described here.

Likelihood Function

The likelihood function for generating MLEs is based on a mixture of item types and can therefore be expressed as

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N_{MC}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{exp \sum_{k=0}^{z_i} Da_i(\theta - \delta_{ki})}{\sum_{j=0}^{m_i} exp \sum_{k=0}^{j} Da_i(\theta - \delta_{ki})}$$

$$P_i = c_i + \frac{1 - c_i}{1 + exp\,[-Da_i(\theta - b_i)]}$$

$$Q_i = 1 - P_i$$

where c_i is the lower asymptote of the item response curve (i.e., the pseudo-guessing parameter), a_i is the slope of the item response curve (i.e., the discrimination parameter), $b_i$is the location parameter, z_i is the observed response to the item, i indexes the item, j indexes the step of the item, m_i is the maximum possible score point (starting from 0), δ_{ki} is the kth step for item i with m total categories, and $D = 1.7$. MC and CR refer to multiple-choice and constructed-response items, respectively.

We subsequently find $\arg\max_{\theta} log\big(L(\theta)\big)$ as the student's theta (i.e., MLE) given the set of items administered to the student.

Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded and an MLE cannot be generated. The extreme cases are handled as follows:

i. Assign the lowest obtainable theta (LOT) value of -3 to a raw score of 0.
ii. Assign the highest obtainable theta (HOT) value of 3 to a perfect score.
iii. Generate MLE for every other case and apply the following rule:
    a. If MLE is lower than -3, assign theta to -3.
    b. If MLE is higher than 3, assign theta to 3.

Numerically Differentiated Hessian of Log-Likelihood

The CSEM is computed using the pattern of responses of the operational items on the adaptively administered test. In this context, the CSEM at the MLE is computed using the inverse of the square root of the negative of the Hessian of the log-likelihood function, which is based on the estimates of the item parameters in the test along with the actual pattern of responses. The formula used is

$$CSEM(\hat{\theta}) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta}\right)}},$$

where

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta} - b_{ik}\right)\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta} - b_{ik}\right)\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta} - b_{ik}\right)\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta} - b_{ik}\right)\right)} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i) Q_i}{(1 - c_i)^2} \left(1 - \frac{z_i c_i}{P_i^2}\right)$$

where N_{GPCM} is the number of items that are scored using GPCM items, and N_{3PL} is the number of items scored using the 3PL or 2PL model, $\hat{\theta}$ is the estimated ability of the student, and D, a_i, c_i, P_i, Q_i, z_i, b_{ik} are defined as before. Through the use of the Newton-Rhapson method during maximum likelihood estimation, this Hessian is numerically approximated at $\hat{\theta}$.

CSEM at Extreme Scores

When the MLE is not available (such as for extreme score cases) or the MLE is censored to the lowest obtainable theta score (LOT) or highest obtainable theta score (HOT), the CSEM for student *s* is estimated by

$$\mathrm{CSEM}(\hat{\theta}_s) = \frac{1}{\sqrt{I(\hat{\theta}_s)}}$$

where $I\left(\hat{\theta}_s\right)$ is the test information for student *s*. The FAST assessments include items that are scored using the 3PL, 2PL, and GPCM models from IRT. The 2PL can be visualized as either a 3PL item with no guessing parameter or a dichotomously scored GPCM item. The test information is calculated as:

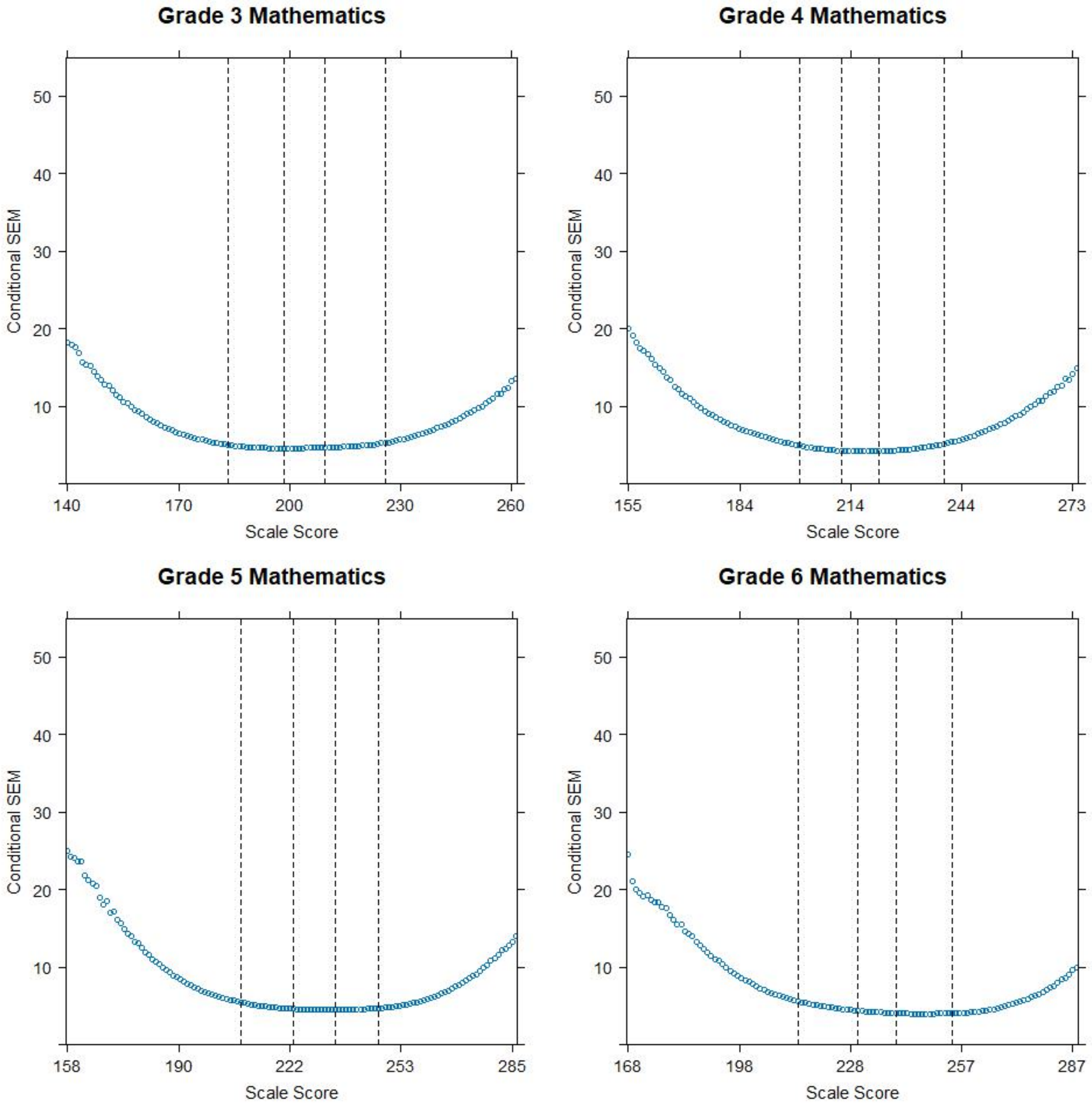
$$I\left(\hat{\theta}_s\right) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta}_s - b_{ik}\right)\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta}_s - b_{ik}\right)\right)} - \left(\frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta}_s - b_{ik}\right)\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i\left(\hat{\theta}_s - b_{ik}\right)\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{Q_i}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2 \right)$$

where, N_{GPCM} is the number of items that are scored using GPCM items, and N_{3PL} is the number of items scored using a 3PL or 2PL model.

For standard error of LOT/HOT scores, theta in the formula on the previous page is replaced with the LOT/HOT values. Finally, CSEM is limited to 1.5 on the theta scale as a global requirement.

These standard error plots are presented in Figure 1 to Figure 4, instead of the test information functions (TIFs). Vertical lines represent the four performance category cut scores. This information is presented for comparison with accommodated forms in Section 5.5, Comparability of Scores, of this volume.

Figure 1: Conditional Standard Errors of Measurement (Mathematics)



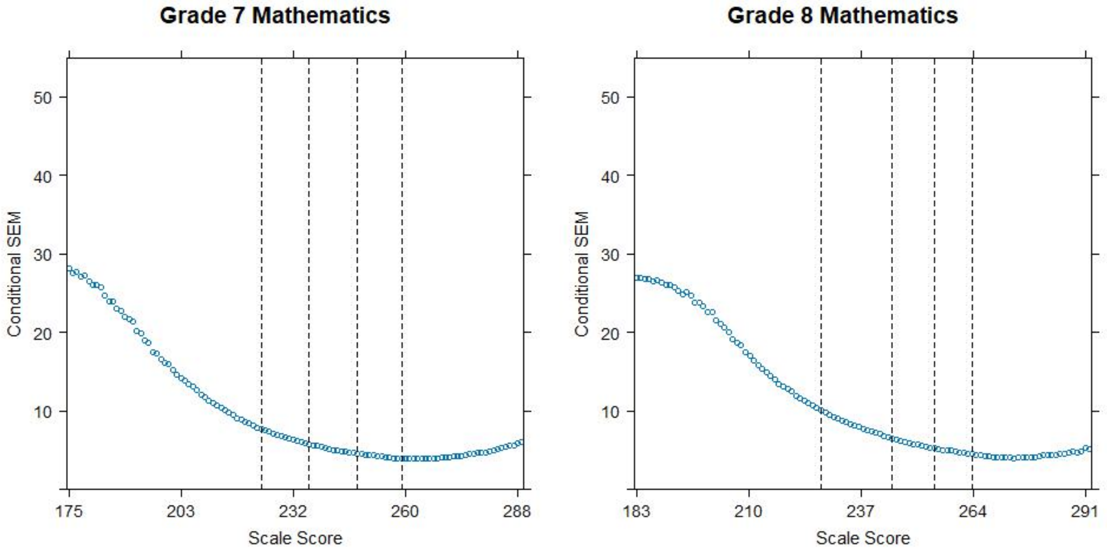
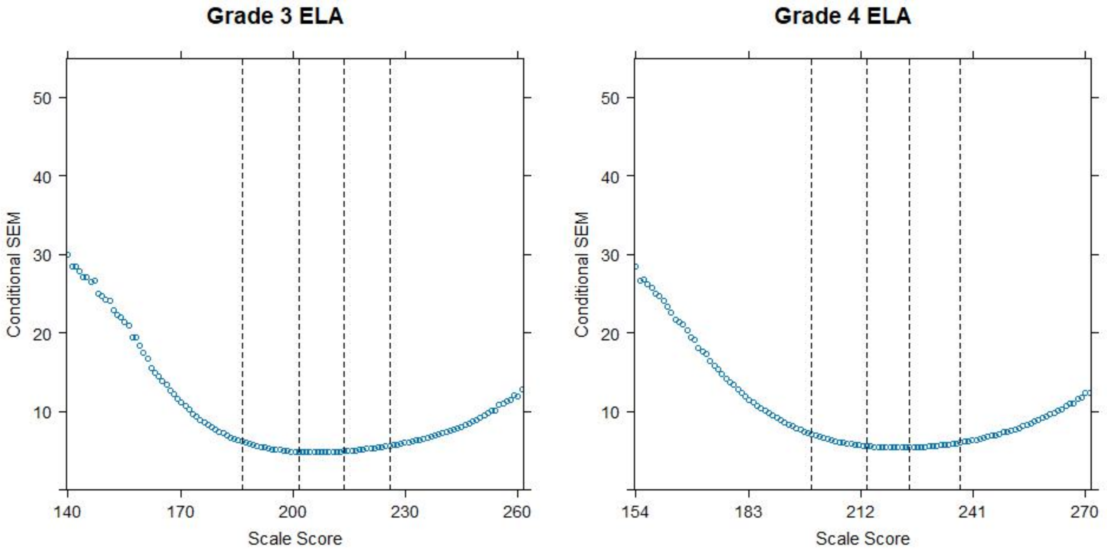
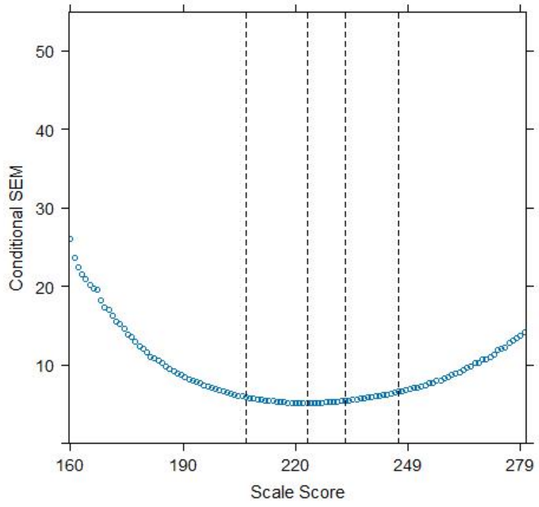


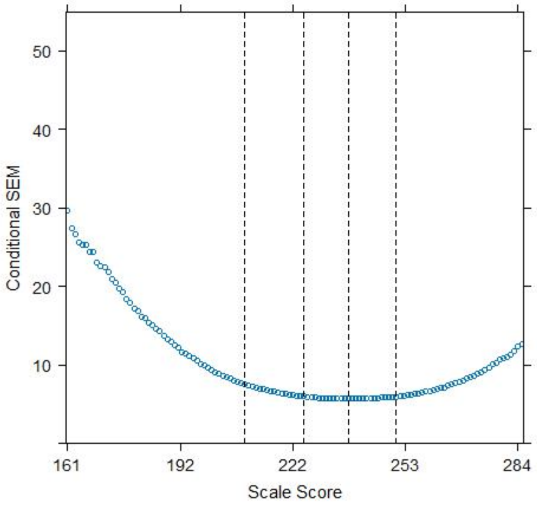
Figure 2: Conditional Standard Errors of Measurement (ELA)



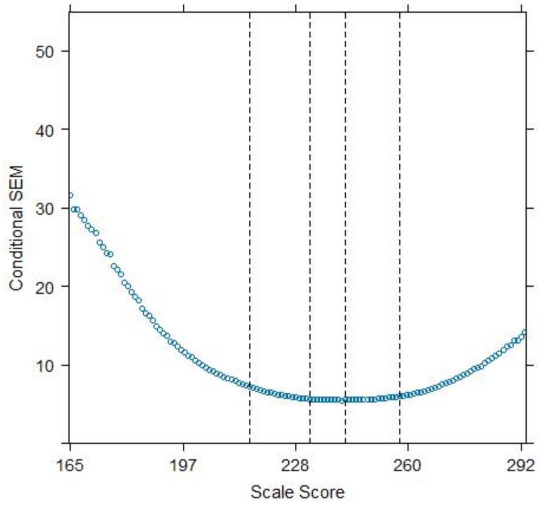
Grade 5 ELA



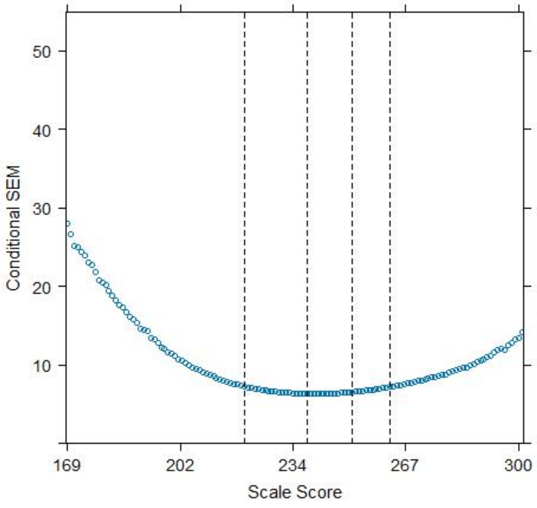
Grade 6 ELA



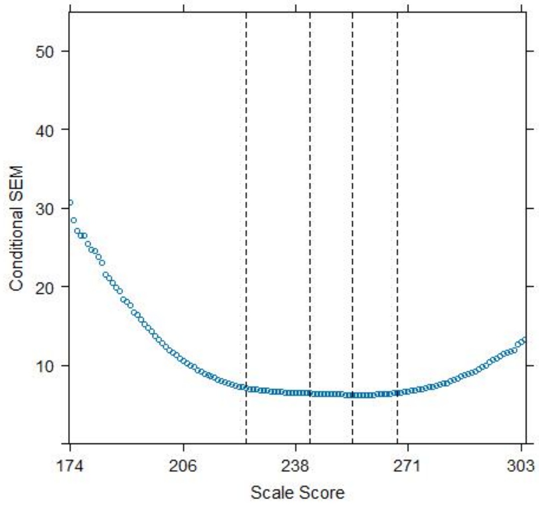
Grade 7 ELA



Grade 8 ELA



Grade 9 ELA



Grade 10 ELA

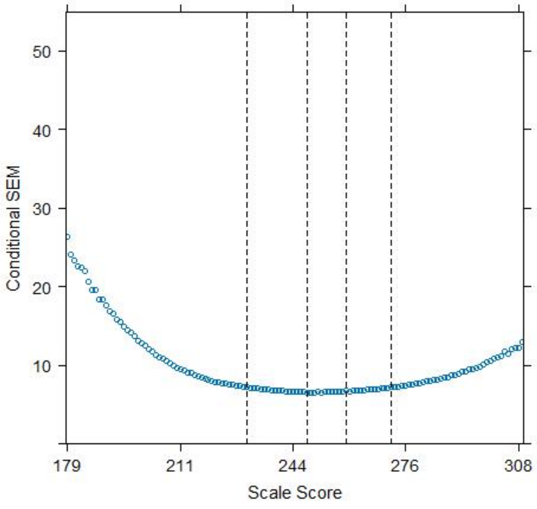


Figure 3: Conditional Standard Errors of Measurement (EOC)

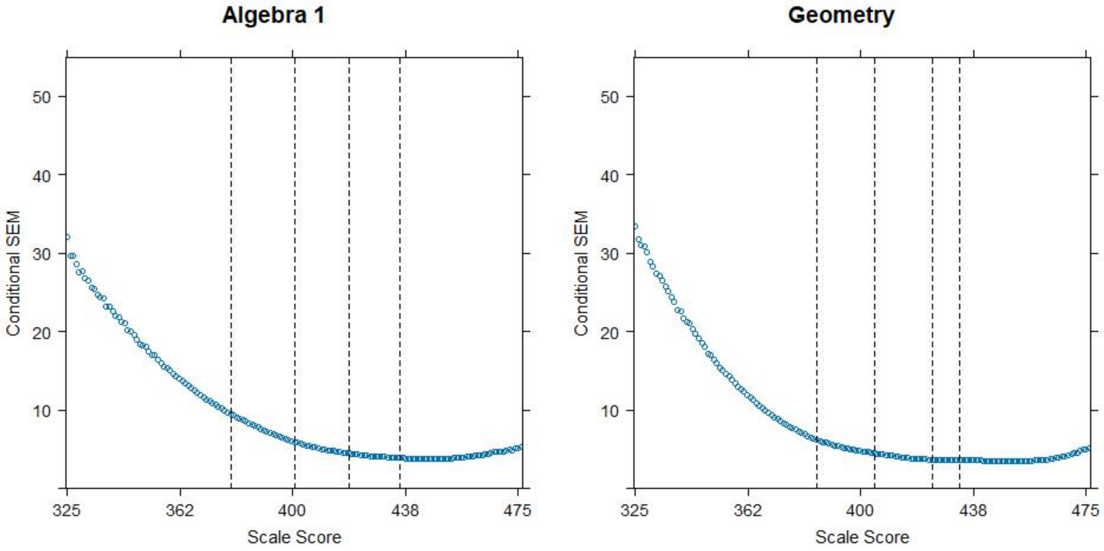
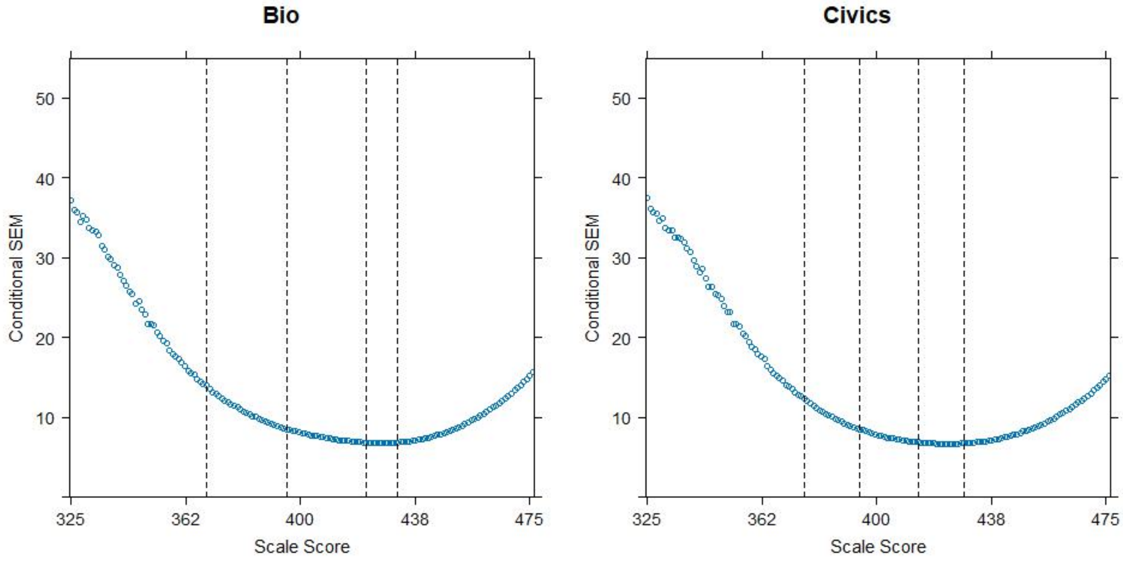
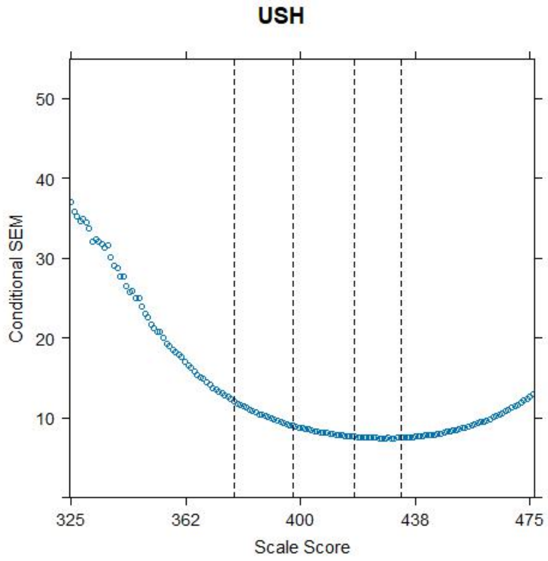
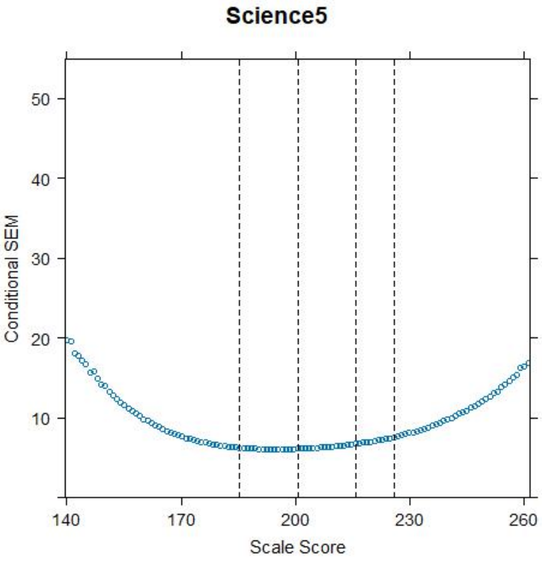
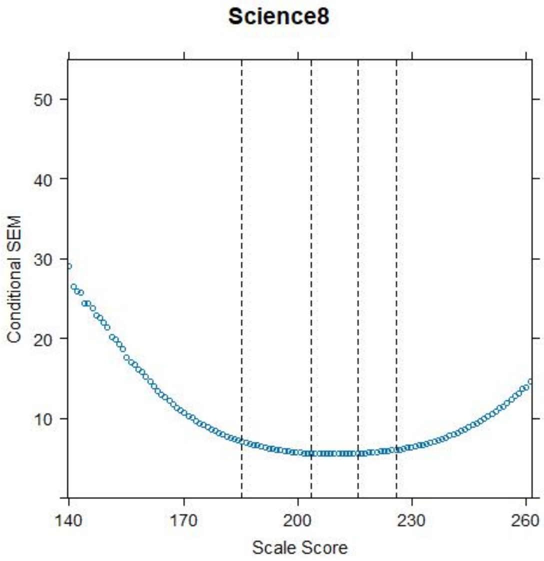


Figure 4: Conditional Standard Errors of Measurement (Science and Social Studies)









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 grades 7 and 8 mathematics and all EOC tests, the standard error curve is minimized at a higher point along the score scale. This suggests the items within these tests are somewhat challenging relative to the tested population. For grades 7 and 8 mathematics, this is in part because the population has lost its upper tail (higher performers) to Algebra or Geometry tests due to their accelerated course progression. With this shift in population, there is a need for developing easier items for these banks, which is a work in progress for the CAI and FDOE psychometric and content teams.

Appendix B, Conditional Standard Error of Measurement, includes scale score by scale score CSEM and corresponding achievement levels for each scale score. It also contains the curves for the mean CSEM across years. CSEM is used by establishing a confidence interval around a student’s observed scale score. This interval indicates where a student would have scored if he or she would have taken the same test again (with no new learning or no memory of questions taking place between test administrations). Reliability coefficients and CSEM for each reporting category are also presented in Appendix A, Reliability Coefficients.

3.3 Reliability of Achievement Classification

When students complete Florida's 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 Florida's standard-setting process.

During test construction, techniques are implemented to minimize misclassification of students, which can occur on any assessment. In particular, the CSEM curves can be constructed to ensure that smaller CSEMs are expected near important cut scores of the test or where most students are scoring. However, it is not possible to tailor the test for the entire ability spectrum, which is the problem that adaptive testing aims to solve.

3.3.1 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 Levels 2 and 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. We used students from the spring 2024 population data files with the status of reported scores.

Table 7 and Table 8 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 MLEs obtained from Cambium Assessment, Inc.'s scoring engine.

Table 7: Descriptive Statistics from Population Data (ELA Reading, Mathematics, and EOC)

ELA Reading				Mathematics			
Grade	N	Average Theta	SD of Theta	Grade	N	Average Theta	SD of Theta
3	215,574	0.05	1.12	3	214,927	0.08	1.08
4	212,165	-0.03	1.18	4	207,096	-0.02	1.10
5	203,412	0.1	1.10	5	197,192	0.07	1.07
6	205,054	0.07	1.15	6	194,855	0.12	1.08
7	214,938	0.01	1.18	7	144,768	-0.06	1.20
8	209,835	0.02	1.15	8	114,710	-0.13	1.27
9	216,621	0.06	1.12	Alg1	228,344	0.01	1.17
10	215,657	0.09	1.11	Geo	213,902	0.10	1.10

* Alg1: Algebra; Geo: Geometry

Table 8: Descriptive Statistics from Population Data (Science & Social Studies)

Science & Social Studies			
Subjects	N	Average Theta	SD of Theta
Biology 1	199,788	0.27	1.17
Civics	188,377	0.33	1.23
U.S. History	183,226	0.38	1.21
Grade 5 Science	174,486	0.15	1.18
Grade 8 Science	178,331	-0.01	1.18

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:

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

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 ith student with an 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 ith 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 classification accuracy indices (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 makes use of student-level item response data from the spring test administration. Drawing on Guo (2006) and Mislevy et al. (1992) 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 9 to Table 12 provide the overall CaI and the CaI for the individual cuts (CAIC) for the tests based on the observed score approach. Here, the overall classification accuracy of the test ranges from 0.675 to 0.779 for mathematics, 0.695 to 0.736 for ELA, 0.790 to 0.823 for EOC, and 0.705 to 0.730 for science and social studies.

Table 9: 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	0.767	0.955	0.931	0.927	0.952
4	0.770	0.942	0.934	0.934	0.958
5	0.769	0.943	0.931	0.938	0.956
6	0.779	0.942	0.930	0.941	0.965
7	0.716	0.897	0.901	0.938	0.969
8	0.675	0.889	0.884	0.920	0.962

Table 10: Classification Accuracy Index (ELA Reading)

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	0.736	0.933	0.926	0.926	0.946
4	0.710	0.922	0.915	0.920	0.946
5	0.724	0.945	0.918	0.914	0.943
6	0.711	0.930	0.912	0.917	0.946
7	0.731	0.930	0.920	0.923	0.951
8	0.712	0.935	0.912	0.916	0.942
9	0.712	0.936	0.911	0.913	0.946
10	0.695	0.938	0.903	0.906	0.939

Table 11: 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	0.790	0.931	0.931	0.950	0.976
Geometry	0.823	0.948	0.942	0.960	0.972

Table 12: Classification Accuracy Index (Science and Social Studies)

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
Biology 1	0.730	0.945	0.921	0.920	0.934
Civics	0.725	0.939	0.923	0.921	0.932
U.S. History	0.705	0.936	0.920	0.914	0.927
Grade 5 Science	0.709	0.942	0.914	0.913	0.931
Grade 8 Science	0.725	0.927	0.917	0.928	0.948

Table 13 to Table 16 provide the FPR and FNR from the IRT-based approach. The FNR and FPR rates for the Level 2/3 cut are around 3%–6% for mathematics and ELA, 3% for EOC, and 3%–5% for science and social studies.

Table 13 to Table 16 also provide the overall accuracy rates after accounting for both false positive and false negative rates. For example, the overall accuracy rate of 0.93 for the Level 2/3 cut in grade 3 mathematics suggests 93% 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 13: 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	0.025	0.021	0.955	0.033	0.037	0.930	0.033	0.041	0.926	0.019	0.029	0.952
4	0.031	0.027	0.942	0.032	0.034	0.934	0.030	0.036	0.934	0.017	0.025	0.958
5	0.030	0.026	0.944	0.033	0.037	0.930	0.028	0.035	0.937	0.018	0.026	0.956
6	0.031	0.027	0.942	0.032	0.039	0.929	0.026	0.033	0.940	0.013	0.022	0.965
7	0.055	0.052	0.894	0.043	0.052	0.905	0.024	0.034	0.942	0.011	0.017	0.972
8	0.072	0.054	0.874	0.044	0.068	0.888	0.027	0.044	0.929	0.012	0.020	0.968

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

	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Grade	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
3	0.034	0.029	0.937	0.032	0.040	0.927	0.031	0.045	0.925	0.021	0.033	0.946
4	0.043	0.035	0.922	0.037	0.047	0.916	0.032	0.049	0.919	0.020	0.034	0.946
5	0.029	0.026	0.945	0.036	0.047	0.917	0.036	0.052	0.912	0.022	0.035	0.943
6	0.037	0.033	0.930	0.038	0.051	0.911	0.033	0.052	0.915	0.019	0.033	0.947
7	0.036	0.032	0.932	0.034	0.046	0.920	0.031	0.047	0.922	0.018	0.030	0.952
8	0.035	0.030	0.936	0.039	0.051	0.910	0.035	0.051	0.914	0.022	0.037	0.941
9	0.034	0.028	0.937	0.038	0.055	0.907	0.035	0.054	0.911	0.020	0.033	0.947
10	0.032	0.031	0.937	0.042	0.059	0.899	0.039	0.059	0.902	0.022	0.038	0.940

Table 15: 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	0.040	0.031	0.929	0.030	0.037	0.934	0.019	0.029	0.951	0.009	0.014	0.977
Geometry	0.026	0.024	0.950	0.026	0.031	0.943	0.017	0.021	0.962	0.012	0.015	0.972

Table 16: False Classification Rates and Overall Accuracy Rates (Science and Social Studies)

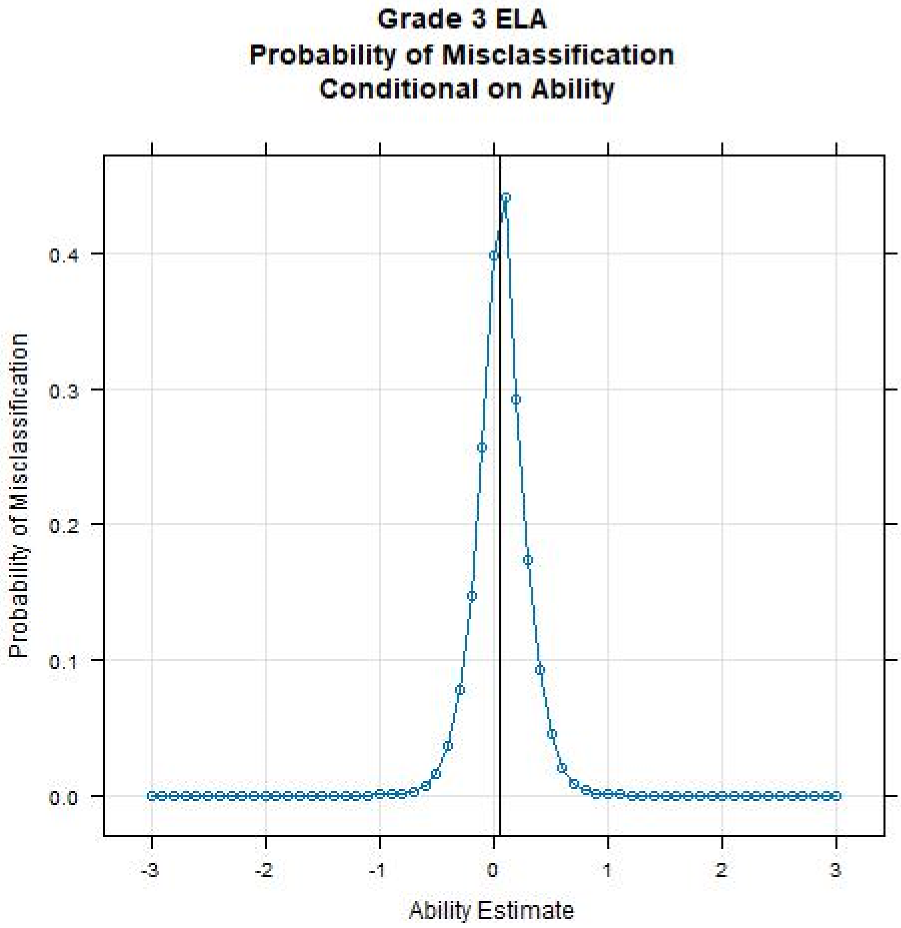
	1/2 cut			2/3 cut			3/4 cut			4/5 cut		
Subject	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
Biology 1	0.039	0.022	0.939	0.035	0.046	0.920	0.028	0.054	0.918	0.022	0.044	0.934
Civics	0.039	0.028	0.933	0.033	0.043	0.924	0.029	0.052	0.919	0.023	0.047	0.930
U.S. History	0.040	0.029	0.931	0.034	0.048	0.918	0.030	0.061	0.910	0.022	0.055	0.923
Grade 5 Science	0.030	0.028	0.942	0.038	0.050	0.911	0.033	0.059	0.908	0.022	0.051	0.926
Grade 8 Science	0.039	0.033	0.928	0.035	0.049	0.916	0.028	0.046	0.926	0.018	0.035	0.947

Figure 5 shows an example plot exhibiting the probability of misclassification for grade 3 ELA. The plot shows that students with scores below –0.308 on the theta scale, which corresponds to a scale score of 194, and students with scores above 0.325, corresponding to a scale score of 208, 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 subjects as well as by subgroups (ELLs and Students with Disabilities [SWD]). 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., on grade level). A properly functioning test yields increased misclassification probabilities approaching the cut, as the density of the posterior probability distribution is symmetric, and approximately half of its mass will fall on either side of the proficiency level cut as $\theta \rightarrow c$.

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

Figure 5: Probability of Misclassification Conditional on Ability



These results demonstrate that classification reliabilities are generally high, with some lower rates affecting tests known to be particularly challenging. We can compare Florida's classification accuracy rates to those of the State of New York, which is comparable in population size (New York State Education Department, 2022). Although New York administers a different testing program, estimated accuracy rates there range from 73%–79% in ELA and from 79%–83% in mathematics (2022). The individual cut accuracy was relatively similar between New York and Florida. For the Level 2/3 cut, Florida showed from 90%–93% in mathematics, from 90%–93% in ELA, and from 93%–94% in EOC. New York showed from 90%–92% in ELA and from 93%–95% in mathematics for the proficiency cut.

3.3.2 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, & 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, item parameters, and 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 17 to Table 24. In the cut 1 and cut 2, cut 2 and cut 3, and cut 3 and cut 4 results, all accuracy values are close to or higher than 0.90, and the 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 17: Classification Accuracy and Consistency (Cut 1 and Cut 2)

Grade	ELA		Grade/ Subject	Mathematics	
	Accuracy	Consistency		Accuracy	Consistency
3	0.933	0.911	3	0.955	0.937
4	0.922	0.89	4	0.942	0.919
5	0.945	0.923	5	0.943	0.921
6	0.930	0.901	6	0.942	0.918
7	0.930	0.903	7	0.897	0.851
8	0.935	0.909	8	0.889	0.827
9	0.936	0.911	Algebra 1	0.931	0.904
10	0.938	0.911	Geometry	0.948	0.932

Table 18: Classification Accuracy and Consistency (Cut 2 and Cut 3)

Grade	ELA		Grade/ Subject	Mathematics	
	Accuracy	Consistency		Accuracy	Consistency
3	0.926	0.898	3	0.931	0.902
4	0.915	0.882	4	0.934	0.906
5	0.918	0.883	5	0.931	0.902
6	0.912	0.875	6	0.930	0.901
7	0.920	0.888	7	0.901	0.868
8	0.912	0.873	8	0.884	0.848
9	0.911	0.87	Algebra 1	0.931	0.91
10	0.903	0.859	Geometry	0.942	0.924

Table 19: Classification Accuracy and Consistency (Cut 3 and Cut 4)

Grade	ELA		Grade/ Subject	Mathematics	
	Accuracy	Consistency		Accuracy	Consistency
3	0.926	0.895	3	0.927	0.896
4	0.920	0.888	4	0.934	0.907
5	0.914	0.877	5	0.938	0.912
6	0.917	0.884	6	0.941	0.916
7	0.923	0.892	7	0.938	0.92
8	0.916	0.882	8	0.920	0.906
9	0.913	0.879	Algebra 1	0.950	0.936
10	0.906	0.866	Geometry	0.960	0.948

Table 20: Classification Accuracy and Consistency (Cut 4 and Cut 5)

Grade	ELA		Grade/ Subject	Mathematics	
	Accuracy	Consistency		Accuracy	Consistency
3	0.946	0.928	3	0.952	0.936
4	0.946	0.93	4	0.958	0.943
5	0.943	0.925	5	0.956	0.939
6	0.946	0.931	6	0.965	0.952
7	0.951	0.937	7	0.969	0.962
8	0.942	0.923	8	0.962	0.958
9	0.946	0.93	Algebra 1	0.976	0.97
10	0.939	0.922	Geometry	0.972	0.963

Table 21: Classification Accuracy and Consistency (Cut 1 and Cut 2)

Subject	Science and Social Studies	
	Accuracy	Consistency
Biology 1	0.945	0.922
U.S. History	0.936	0.905
Civics	0.939	0.909
Grade 5 Science	0.942	0.919
Grade 8 Science	0.927	0.898

Table 22: Classification Accuracy and Consistency (Cut 2 and Cut 3)

Subject	Science and Social Studies	
	Accuracy	Consistency
Biology 1	0.921	0.886
U.S. History	0.920	0.884
Civics	0.923	0.893
Grade 5 Science	0.914	0.875
Grade 8 Science	0.917	0.883

Table 23: Classification Accuracy and Consistency (Cut 3 and Cut 4)

Subject	Science and Social Studies	
	Accuracy	Consistency
Biology 1	0.920	0.889
U.S. History	0.914	0.878
Civics	0.921	0.889
Grade 5 Science	0.913	0.875
Grade 8 Science	0.928	0.899

Table 24: Classification Accuracy and Consistency (Cut 4 and Cut 5)

Subject	Science and Social Studies	
	Accuracy	Consistency
Biology 1	0.934	0.913
U.S. History	0.927	0.901
Civics	0.932	0.907
Grade 5 Science	0.931	0.906
Grade 8 Science	0.948	0.931

3.4 Precision at Cut Scores

Table 25 to Table 28 present the mean CSEM at each achievement level by grade and subject. These tables also include achievement level cut scores and associated CSEM.

Table 25: Achievement Levels and Associated Conditional Standard Errors of Measurement (Mathematics)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	8.083		
	2	4.757	183	5.073
	3	4.661	198	4.639
	4	4.898	209	4.715
	5	6.728	225	5.321
4	1	8.143		
	2	4.600	200	4.983
	3	4.259	211	4.321
	4	4.523	221	4.263
	5	7.025	238	5.175
5	1	9.390		
	2	4.988	207	5.518
	3	4.576	222	4.658
	4	4.603	234	4.544
	5	5.917	246	4.760
6	1	9.425		
	2	4.926	213	5.631
	3	4.266	229	4.417
	4	4.061	239	4.116
	5	4.806	254	4.137
7	1	14.769		
	2	6.722	223	7.744
	3	5.213	235	5.805
	4	4.328	247	4.601
	5	4.175	258	3.988
8	1	18.286		
	2	8.148	227	10.104
	3	5.960	244	6.582
	4	4.990	254	5.287
	5	4.311	263	4.554

Table 26: Achievement Levels and Associated Conditional Standard Errors of Measurement (ELA Reading)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	13.193		
	2	5.370	186	6.194
	3	4.856	201	4.878
	4	5.233	213	4.960
	5	6.658	225	5.688
4	1	13.716		
	2	6.224	199	7.098
	3	5.497	213	5.634
	4	5.627	224	5.431
	5	7.098	237	6.051
5	1	10.285		
	2	5.414	206	5.941
	3	5.277	222	5.196
	4	5.895	232	5.459
	5	7.914	246	6.622
6	1	14.193		
	2	6.655	209	7.555
	3	5.824	225	6.024
	4	5.806	237	5.717
	5	6.844	250	6.001
7	1	14.387		
	2	6.287	215	7.249
	3	5.613	232	5.700
	4	5.692	242	5.580
	5	7.220	257	6.042
8	1	12.594		
	2	6.678	220	7.313
	3	6.418	238	6.380
	4	6.824	251	6.582
	5	8.261	262	7.243
9	1	12.585		
	2	6.694	224	7.188
	3	6.373	242	6.461

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
	4	6.332	254	6.280
	5	7.405	267	6.549
10	1	11.126		
	2	6.847	230	7.272
	3	6.643	247	6.610
	4	6.908	258	6.768
	5	8.021	271	7.226

Table 27: Achievement Levels and Associated Conditional Standard Errors of Measurement (EOC)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Algebra 1	1	16.621		
	2	7.572	379	9.574
	3	5.187	400	5.999
	4	4.223	418	4.514
	5	3.942	435	3.933
Geometry	1	12.643		
	2	5.239	385	6.290
	3	4.101	404	4.560
	4	3.699	423	3.721
	5	3.638	432	3.686

Table 28: Achievement Levels and Associated Conditional Standard Errors of Measurement (Science and Social Studies)

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Biology 1	1	23.051		
	2	10.591	369	13.977
	3	7.530	395	8.563
	4	6.854	421	6.886
	5	8.338	431	6.897
Civics	1	22.026		
	2	10.137	376	12.347
	3	7.572	394	8.536
	4	6.794	413	6.941
	5	8.371	428	6.785

Grade	Achievement Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
U.S. History	1	19.927		
	2	10.358	378	12.113
	3	8.245	397	9.068
	4	7.551	417	7.677
	5	8.494	432	7.532
Grade 5 Science	1	8.965		
	2	6.162	185	6.103
	3	6.395	200	6.218
	4	7.101	215	6.832
	5	9.850	225	8.048
Grade 8 Science	1	13.117		
	2	6.187	185	7.141
	3	5.596	203	5.649
	4	5.787	215	5.629
	5	7.541	225	6.099

3.5 WRITING HYBRID AUTOMATED AND HUMAN SCORING

During spring 2023, the writing assessments were decoupled from ELA and administered as stand-alone field tests based on a representative sample of schools. Volume 1, Section 4.3, Field Testing, details how the representative sample was derived. CAI and Data Recognition Corporation (DRC) conducted hybrid automated/human scoring of B.E.S.T. writing items in grades 4–10 in operational administration for spring 2024. The full report is found in Appendix I.

The hybrid scoring method has multiple steps. First, CAI's autoscoring system, Autoscore, is used to train models on scores and responses from the stand-alone field test administration conducted in spring 2023. Results from the field test can be found in Appendix J. Once deployed for operational scoring, all responses receive scores from Autoscore. Responses are routed for one of four reasons: 1) as a random read for monitoring purposes, similar to a reliability read in a fully handscored approach; 2) due to the assignment of certain condition codes that warrant human review; 3) due to a low-confidence designation that indicates the engine score is not likely to match the score of a trained human rater; and, 4) for responses receiving a score of 2 in the Development domain (grades 4 and 5 only).

Approximately 40% of responses were routed in grades 6–10 for human scoring. Approximately 75% of responses in grades 4 and 5 were routed for human scoring. When routed for human scoring, the human score is the final reported score. Responses routed for human scoring do not receive second reads from other human raters; raters are monitored for quality using backreads and validity sets. During the test administration, the performance of Autoscore and of human scoring on each item and domain was monitored daily.

We begin with a description of Autoscore and how it is trained and evaluated, then provide performance of Autoscore relative to human scoring on the sample used to validate the engine prior to use in B.E.S.T. operational scoring and on the random sample during operational testing. We end the section by comparing the human and Autoscore means and standard deviations to the hybrid scores for the full sample. Note that our analysis includes all responses routed to Autoscore.

Autoscore

Autoscore uses features associated with writing quality and features associated with response meaning to model human rater scoring behavior. Writing quality features include measures of syntax, grammatical/mechanical correctness, spelling correctness, text complexity, paragraphing quality, and sentence variation and quality. Measures of response meaning include the use of latent semantic analysis to identify key topics associated with patterns of words in a response (LSA; Deerwester et al., 1990) and 'deep learning' methods, which consist of a richer representation of language that includes a contextual representation of all the words in the response and how they are used relative to one another in language. Deep learning methods leverage models of language patterns learned from large bodies of text (Vaswani et al., 2017).

In Autoscore, for each item and domain, we train two models in parallel and combine the outputs of these models to predict the score for a response. More specifically, one model uses latent semantic analysis and writing quality features to model human raters. The other model is a deep learning model, ELECTRA (Clark et al., 2020). The logit or probabilistic outputs for each score from these two models are then used to estimate the parameters of a logistic regression to produce a final score. Combining the results, or ensembling the results, generally produces better performance than the use of a single model (Zhou et al., 2002).

Autoscore also assigns condition codes and confidence values. Condition codes assigned by Autoscore appear in Table 29, along with whether responses receiving the condition code are routed for human scoring. Autoscore condition codes may not perfectly align with each of the Florida B.E.S.T. human-assigned condition codes, in language and in function. The purpose of the Autoscore condition codes is to identify responses not meeting rubric requirements to achieve the minimal score (1, in the B.E.S.T. rubric) or to identify responses that are unusual in some way that should be reviewed by human raters. Note that any response routed due to a condition code is then assigned the condition code, or score, using the B.E.S.T rubrics. The choice to use the condition code, the threshold for the code (if applicable), and the routing status were made with FDOE using both the B.E.S.T. rubrics and the handscored data from the field test.

Table 29: Autoscore Condition Codes and Whether Routed for Human Scoring

Autoscore Condition Code	Routed for Human Verification?	Description	Thresholds
No Response	No	No non-blank characters are detected in the response.	n.a.
Common Refusal	No	Response only contains words associated with a refusal such as 'I don't know' or contains only non-alphanumeric characters.	n.a.
Not Enough Data	Yes	Student response is less than the minimum number of words configured in the rubric.	45

Autoscore Condition Code	Routed for Human Verification?	Description	Thresholds
Duplicate Text	Yes	Student response consists primarily of repeated text.	.40
Prompt Copy Match	Yes	Student response is primarily copied from the passage or item prompt. Percentage of characters in the response that appear in the passage.	80%
Non-Scorable Language	Yes	Response is longer than 30 characters and is written primarily in Spanish.	n.a.
Out-of-Vocabulary	Yes	The ratio of the sum of the lengths of words in a response that are in the engine training sample over the sum of length of all words in the response.	n.a.
Non-Specific	Yes	Essay scoring engine predicts the assignment of a human-based condition code using a statistical procedure.	n.a.
Unusual Scores	Yes	Identifies responses with Autoscore scores that are unusual in some way, including: Any response receiving non-adjacent domain scores (e.g., if the engine assigns a score in Development of 2 and Purpose/Structure of 4) Any response with greater than 45 words and fewer than 61 words and receiving a score of 2 or higher in Development or Purpose/Structure. Any response receiving a 3 or 4 in Development that do not contain evidence/citation according to grade-level criteria.	

For essays not receiving a condition code, Autoscore produces a confidence index. This index reflects the degree of confidence the engine has that the score it predicted matches the score a well-trained and experienced human scorer would assign. The held-out validation data are used to estimate the confidence model. The confidence value is based upon logistic regression output which uses the patterns of the model score probabilities to predict whether the engine score matches the human score. Inputs to the model are the individual model logits or probabilities for each score point and the ensembled logistic regression max probability; the dependent variable is where the engine score matches the final human resolved score (1 = match; 0 = non-match). A model is trained for each domain and then the domain confidences are summed after centering to 0 and rescaling to have standard deviation 1. The logistic regression model outputs are then mapped to a percentile scale that ranges from 0 to 100. A low value indicates that the engine has low confidence in the score it has assigned; a high value indicates that the engine has high confidence in the score. The confidence percentiles are produced for each domain and the overall confidence value. The overall confidence percentile is what is used in routing. Each item has its own confidence model, with all items using the same threshold. Responses with a confidence score below the 25th percentile are flagged and routed for human review.

Autoscore Training

CAI trains models for each item and domain. Data used to train Autoscore models are from the spring 2023 stand-alone field test. These data were scored by two independent, trained human raters with resolution of any non-exact score within domain. Please see Appendix J for more information on the hand-scoring method and results for those data. Data are divided into training, ensemble, and held-out validation sets, with 70% of responses used to train the two models, 15% used to train the ensemble, and 15% used to evaluate the engine performance. Data are stratified on the sum of the three final, resolved domain scores to ensure that score point distributions are evenly represented in both sets. Human-assigned condition codes are removed prior to training the models and are added later in the process when applying the Autoscore condition codes.

Evaluation Metrics

Metrics used to examine engine performance are those commonly used in the assessment industry (Williamson, Xi, and Breyer, 2012). These include measures of agreement (Exact Agreement, Quadratic Weighed Kappa or QWK using Fleiss-Cohen weights) and a distributional measure (Standardized Mean Difference or SMD using pooled standard deviation).

CAI used the following thresholds to identify poorly performing items:

- Engine-Final, resolved score exact agreement lower than 5.25% of human-human exact agreement (PARCC, 2015)
- Engine-Final, resolved QWK lower than .1 of human-human QWK (Williamson et al., 2012)
- Engine-Final, resolved SMD magnitude greater than .15 (Williamson et al., 2012)

3.5.1 Autoscore Performance on the Held-out Validation Sample

The performance of Autoscore on the held-out validation sample showed that Autoscore met or exceeded performance criteria for all three metrics across all items and domains. Autoscore (HSAS) showed similar or higher levels of exact and QWK agreement relative to the two human scores (H1H2) (refer to Table 30). Note that all analyses in this section are conducted on the responses in which both Autoscore and human-assigned condition codes were removed. This approach was taken because the core focus is on the ability of the engine to reproduce rubric scores.

Table 30: Autoscore Performance Compared to Human-Human Agreement on Exact Agreement and Quadratic Weighted Kappa on the Held-out Validation Sample (Condition Codes Removed)

Grade	Item ID	N	Domain	Exact Agreement			QWK		
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.
4	37641	657	Lang.	68.8%	77.9%	9.1%	0.65	0.74	0.09
4	37641	657	Dev.	72.3%	77.6%	5.3%	0.67	0.71	0.04
4	37641	657	P/S	69.9%	79.5%	9.6%	0.67	0.77	0.09
5	37736	695	Lang.	71.1%	76.1%	5.0%	0.73	0.74	0.01
5	37736	695	Dev.	74.0%	78.1%	4.2%	0.75	0.76	0.01
5	37736	695	P/S	72.4%	81.3%	8.9%	0.74	0.81	0.07
6	38112	733	Lang.	71.1%	81.3%	10.2%	0.66	0.74	0.08
6	38112	733	Dev.	71.8%	77.9%	6.1%	0.69	0.71	0.02
6	38112	733	P/S	72.0%	80.5%	8.5%	0.69	0.75	0.06
7	37678	681	Lang.	69.5%	80.3%	10.9%	0.66	0.77	0.11

Grade	Item ID	N	Domain	Exact Agreement			QWK		
				H1H2	HSAS	Diff.	H1H2	HSAS	Diff.
7	37678	681	Dev.	75.6%	84.4%	8.8%	0.60	0.73	0.13
7	37678	681	P/S	71.4%	81.6%	10.3%	0.66	0.75	0.09
8	38033	704	Lang.	67.9%	77.7%	9.8%	0.68	0.74	0.06
8	38033	704	Dev.	72.3%	78.0%	5.7%	0.73	0.77	0.03
8	38033	704	P/S	71.0%	78.4%	7.4%	0.72	0.77	0.06
9	37737	687	Lang.	72.6%	79.5%	6.8%	0.68	0.74	0.06
9	37737	687	Dev.	73.9%	79.5%	5.5%	0.72	0.77	0.04
9	37737	687	P/S	74.4%	79.9%	5.5%	0.73	0.77	0.04
10	37613	576	Lang.	69.1%	85.1%	16.0%	0.68	0.82	0.15
10	37613	576	Dev.	72.2%	82.1%	9.9%	0.75	0.83	0.08
10	37613	576	P/S	72.9%	84.0%	11.1%	0.76	0.85	0.09

Note: Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10; for essays. P/S refers to the Purpose/Structure dimension, Dev. Refers to the Development dimension, and Lang. refers to the Language dimension.

Table 31 presents item domain results for mean, standard deviation, and standardized mean difference for the human score and engine score. The HSAS SMD values ranged from -0.02 to .08 across items and domains, within the threshold of +/- 0.15.

Table 31: Autoscore Performance Compared to Human-Human Agreement on Standardized Mean Difference (SMD) on the Held-out Validation Sample (Condition Codes Removed)

Grade	Item ID	N	Domain	HS		AS		SMD	
				Mean	SD	Mean	SD	H1H2	HSAS
4	37641	657	Lang.	2.03	0.65	2.02	0.65	0.00	-0.02
4	37641	657	Dev.	1.99	0.63	1.99	0.62	-0.01	0.00
4	37641	657	P/S	2.06	0.67	2.04	0.67	-0.00	-0.03
5	37736	695	Lang.	2.19	0.70	2.16	0.65	0.01	-0.03
5	37736	695	Dev.	2.16	0.70	2.14	0.65	0.01	-0.03
5	37736	695	P/S	2.23	0.72	2.25	0.68	0.02	0.03
6	38112	733	Lang.	2.21	0.62	2.23	0.58	-0.00	0.03
6	38112	733	Dev.	2.18	0.66	2.17	0.57	-0.02	-0.01
6	38112	733	P/S	2.23	0.66	2.25	0.58	-0.03	0.02
7	37678	681	Lang.	2.07	0.66	2.05	0.63	0.00	-0.04
7	37678	681	Dev.	2.01	0.55	2.00	0.52	-0.02	-0.03
7	37678	681	P/S	2.11	0.62	2.11	0.59	-0.02	-0.01
8	38033	704	Lang.	2.28	0.70	2.23	0.63	0.03	-0.08
8	38033	704	Dev.	2.23	0.70	2.20	0.67	-0.02	-0.04
8	38033	704	P/S	2.26	0.71	2.23	0.67	-0.02	-0.04
9	37737	687	Lang.	2.26	0.65	2.24	0.62	0.03	-0.04
9	37737	687	Dev.	2.22	0.68	2.17	0.64	0.01	-0.07
9	37737	687	P/S	2.24	0.69	2.20	0.66	0.01	-0.05
10	37613	576	Lang.	2.24	0.67	2.22	0.63	-0.04	-0.02

Grade	Item ID	N	Domain	HS		AS		SMD	
				Mean	SD	Mean	SD	H1H2	HSAS
10	37613	576	Dev.	2.14	0.74	2.14	0.72	-0.03	0.00
10	37613	576	P/S	2.17	0.74	2.18	0.71	-0.04	0.02

Note: Target performance for SMD is within +/- 0.15. P/S refers to the Purpose/Structure dimension, Dev. refers to the Development dimension, and Lang. refers to the Language dimension.

Table 32 shows the inter-domain correlations for the final, resolved human scores and for Autoscore. We expect these correlations to be very similar. In general, the Autoscore domain correlations are slightly lower than the human domain correlations but are similar in magnitude.

Table 32: Correlations Between Domains Across Human Score and Autoscore

Item ID	N	Human Score			Autoscore		
		Dev. - Lang.	Dev. - P/S	Lang. - P/S	Dev. - Lang.	Dev. - P/S	Lang. - P/S
37613	623	0.91	0.97	0.94	0.89	0.96	0.91
37641	747	0.91	0.95	0.94	0.92	0.92	0.92
37678	763	0.82	0.91	0.90	0.83	0.87	0.88
37736	765	0.95	0.94	0.95	0.88	0.90	0.87
37737	751	0.92	0.98	0.93	0.90	0.95	0.91
38033	765	0.91	0.97	0.93	0.89	0.93	0.90
38112	810	0.87	0.94	0.89	0.86	0.89	0.88

Operational Routing Percentages

The number and percentages of responses routed for hand-scoring under the four routing conditions, as well as the total routed, appear in Table 33. Percentages in the table are shaded in grey when they are not within 5% of the target value. The target estimates for condition codes and low-confidence routing are expected to vary, especially for condition codes which can vary across samples. The low-confidence value for the grade 4 items is lower than expected.

Table 33: Number and Percentage of Responses Routed for Human Scoring by Routing Condition

Grade	Item ID	Total Tested	Random Selection		Condition Code		Low Confidence		Custom		Total Routed	
			Target=5%		Target=10%		Target=25%		Target=45%		Target=40%-85%	
			N	%	N	%	N	%	N	%	N	%
4	37641	210,101	10,421	5.0%	23,314	11.1%	41,383	19.7%	91,039	43.3%	166,157	79.1%
5	37736	201,541	10,136	5.0%	16,262	8.1%	45,185	22.4%	79,122	39.3%	150,705	74.8%
6	38112	202,992	10,153	5.0%	16,977	8.4%	49,963	24.6%			77,093	38.0%
7	37678	212,943	10,650	5.0%	18,962	8.9%	50,236	23.6%			79,848	37.5%
8	38033	207,912	10,575	5.1%	13,961	6.7%	47,420	22.8%			71,956	34.6%
9	37737	212,732	10,677	5.0%	16,291	7.7%	51,192	24.1%			78,160	36.7%
10	37613	210,527	10,461	5.0%	12,683	6.0%	52,463	24.9%			75,607	35.9%
		1,458,748	73,073	5.0%	118,450	8.1%	337,842	23.2%	170,161		699,526	48.0%

3.5.1 Hand-Scoring of Routed Responses

When responses are routed for hand-scoring, the scores arising out of the hand-scoring process were the score of record. Hand-scorers could assign scores in each domain on the rubric or condition codes (refer to Table 41). Responses routed under the Random Selection, Low-Confidence, and Custom rationales were routed to the general rater pool. Responses routed under the Condition Code rationale were routed to expert raters. Training and scoring are remotely conducted.

Raters undergo training using the anchor and training sets defined during range finding. The annotated anchor sets have approximately 16 responses, with 3–4 responses at each score point, considering also variation across the domains. There are three training sets, consisting of a total of 25 papers total. Training is conducted during live, synchronous sessions. Once training is completed, raters take two qualifying sets, each with 10 papers. Raters must achieve a 70% exact agreement rate in each domain in at least one of the sets in order to be qualified to score. Raters also undergo training for condition codes, with specific materials focused on the amount and type of copied text from the passage and prompt. Training on a prompt takes three to four days.

Once scoring begins, qualified raters assign scores independently of the Autoscore-assigned score. Raters are monitored using validity responses and daily calibrations; these are used to ensure that rater scoring remains true to the range finding decisions. Approximately six validity responses are assigned to each rater each data day. The exact agreement rates are calculated between the set of all raters and the ‘true’ score and the validity responses. Raters are expected to achieve 70% exact agreement in each domain. Score point distributions are computed, as well. Validity responses are approved by DRC and FDOE. Any condition code assigned by the raters are routed to scoring directors and expert scorers for final verification. Daily calibrations consist of 1–3 calibration responses, in which raters assign scores and discuss results with training leads. Calibration sets are also approved by DRC and FDOE. Note that there is no random second read for routed responses, and so no human rater reliability metrics can be computed. Table 34 presents the results on the validity responses and scores, aggregated across the administration.

Table 34: Validity Response Exact Agreements, as a Percentage

Grade	Purpose / Structure	Development	Language
4	89	90	89
5	87	86	86
6	86	86	84
7	82	81	81
8	90	90	89
9	92	92	92
10	91	91	90

3.5.2 Operational Performance in the Aggregate

The performance of Autoscore on the held-out validation sample showed that Autoscore met or exceeded performance criteria for all three metrics across all items and domains. Table 35 presents the Exact Agreement and Quadratic Weighted Kappa (QWK) of human-human agreement (H1H2), human-machine agreement (HSAS), and the difference between the two, for each item and domain on the responses routed randomly for human scoring. The H1H2 statistics reflect agreement from the held-out validation essays from the field-tested data and are used for comparison purposes to assess whether the agreement from operationally scored essays, referred to as the HSAS statistics, fall within an acceptable range.

Table 35: QWK and Exact Agreement of Autoscore Compared to Human-Human Agreement on the Random Routed Sample (Condition Codes Removed)

					Exact Agreement			QWK		
Grade	Item ID	N FT	N OP	Domain	H1H2 FT	HSAS OP	diff	H1H2 FT	HSAS OP	diff
4	37641	657	9,163	Lang.	68.8%	72.7%	3.9%	0.65	0.71	0.05
4	37641	657	9,163	Dev.	72.3%	74.2%	1.9%	0.67	0.69	0.02
4	37641	657	9,163	P/S	69.9%	71.8%	2.0%	0.67	0.70	0.03
5	37736	695	9,217	Lang.	71.1%	71.9%	0.8%	0.73	0.70	-0.03
5	37736	695	9,217	Dev.	74.0%	75.0%	1.1%	0.75	0.73	-0.02
5	37736	695	9,217	P/S	72.4%	73.3%	1.0%	0.74	0.73	-0.01
6	38112	733	9,254	Lang.	71.1%	73.9%	2.8%	0.66	0.70	0.04
6	38112	733	9,254	Dev.	71.8%	75.0%	3.2%	0.69	0.72	0.03
6	38112	733	9,254	P/S	72.0%	74.7%	2.7%	0.69	0.73	0.03
7	37678	681	9,596	Lang.	69.5%	74.7%	5.2%	0.66	0.71	0.05
7	37678	681	9,596	Dev.	75.6%	75.8%	0.2%	0.60	0.66	0.06
7	37678	681	9,596	P/S	71.4%	75.4%	4.0%	0.66	0.71	0.04
8	38033	704	9,815	Lang.	67.9%	73.0%	5.1%	0.68	0.71	0.03
8	38033	704	9,815	Dev.	72.3%	74.2%	1.9%	0.73	0.73	0.00
8	38033	704	9,815	P/S	71.0%	74.4%	3.4%	0.72	0.74	0.03
9	37737	687	9,844	Lang.	72.6%	76.0%	3.4%	0.68	0.73	0.05
9	37737	687	9,844	Dev.	73.9%	77.0%	3.0%	0.72	0.75	0.03
9	37737	687	9,844	P/S	74.4%	76.7%	2.3%	0.73	0.76	0.03
10	37613	576	9,880	Lang.	69.1%	77.5%	8.4%	0.68	0.73	0.06
10	37613	576	9,880	Dev.	72.2%	77.7%	5.5%	0.75	0.76	0.00
10	37613	576	9,880	P/S	72.9%	78.5%	5.5%	0.76	0.76	0.01

Note: Target performance for Exact Agreement is a difference of less than 5.25%. Target performance for QWK is a difference of less than 0.10; for essays, P/S refers to the Purpose/Structure dimension, Dev. refers to the Development dimension, and Lang. refers to the Language dimension. N FT is the number of human-scored responses from the field-tested sample, whereas N OP is the number of essays that received both a machine and human score during the operational testing window.

Table 36 presents the means and standard deviations of scores assigned by both the human rater (HS) and Autoscore (AS) during the operational testing window. Additionally, this table presents the Standardized Mean Difference (SMD) of human-human agreement (H1H2) and human-machine agreement (HSAS), for each item. In this table, as in the previous table, the H1H2 statistics represent the SMD value from the held-out validation sample of scored essays from the field-tested data. While not directly used in the evaluation, the H1H2 SMDs provide a useful reference for how two humans agree on the standardized mean score. All HSAS SMDs were within the .15 magnitude threshold, with the largest value being .10 (Grade 10 item 37513, Purpose/Structure).

Table 36: Standardized Mean Difference (SMD) of Autoscore Compared to Human-Human SMD in the Random Sample

					HS		AS		SMD	
Grade	Item ID	N FT	N OP	Domain	Mean	SD	Mean	SD	H1H2 FT	HSAS
4	37641	657	9,163	Lang.	2.05	0.74	2.04	0.65	0.02	-0.01
4	37641	657	9,163	Dev.	1.99	0.69	2.00	0.62	-0.00	0.02
4	37641	657	9,163	P/S	2.07	0.73	2.07	0.66	0.03	-0.00
5	37736	695	9,217	Lang.	2.18	0.74	2.14	0.65	0.03	-0.06
5	37736	695	9,217	Dev.	2.15	0.72	2.12	0.65	0.03	-0.04
5	37736	695	9,217	P/S	2.19	0.74	2.22	0.68	-0.03	0.04
6	38112	733	9,254	Lang.	2.21	0.71	2.23	0.61	-0.03	0.04
6	38112	733	9,254	Dev.	2.17	0.72	2.18	0.62	0.01	0.00
6	38112	733	9,254	P/S	2.20	0.73	2.26	0.64	-0.02	0.08
7	37678	681	9,596	Lang.	2.03	0.68	2.04	0.65	0.04	0.01
7	37678	681	9,596	Dev.	1.99	0.66	1.99	0.57	0.03	-0.01
7	37678	681	9,596	P/S	2.06	0.67	2.11	0.64	0.01	0.07
8	38033	704	9,815	Lang.	2.20	0.71	2.23	0.67	0.08	0.03
8	38033	704	9,815	Dev.	2.18	0.72	2.19	0.69	0.04	0.02
8	38033	704	9,815	P/S	2.19	0.71	2.23	0.71	0.04	0.05
9	37737	687	9,844	Lang.	2.24	0.70	2.26	0.64	0.04	0.04
9	37737	687	9,844	Dev.	2.22	0.72	2.23	0.67	0.07	0.01
9	37737	687	9,844	P/S	2.23	0.72	2.25	0.69	0.05	0.03
10	37613	576	9,880	Lang.	2.23	0.65	2.28	0.64	0.02	0.08
10	37613	576	9,880	Dev.	2.15	0.65	2.21	0.70	-0.00	0.08
10	37613	576	9,880	P/S	2.17	0.66	2.24	0.69	-0.02	0.10

Note: Target performance for SMD is within +/- 0.15. P/S refers to the Purpose/Structure dimension, Dev. Refers to the Development dimension, and Lang. refers to the Language dimension. N FT is the number of human-scored responses from the field tested sample, whereas N OP is the number of essays that received both a machine and human score during the operational testing window.

3.5.3 Operational Performance by Student Group

It is important to ensure that Autoscore is performing well, not just overall, but for student groups. In Appendix K, we analyze Autoscore performance, disaggregated by student groups. For the autoscore evaluation, we restrict our analysis to student groups with 300 or more examinees. We provide analysis by gender, ethnicity, ELL status, and primary disability status. Specifically, we examine performance across female and male students, Black, Latino, and White students, ELL Status (Y/N), and two disability types (Specific Learning Disability and Gifted). For each analysis, we present the HS and AS means and standard deviations, as well as ASHS SMD, QWK, and exact agreement. We flag any SMD value that exceeds .15 in magnitude.

The sample does not have two human scores, so there is no relative metric to evaluate the engine-human QWK and exact agreements. We should also expect that, especially when human mean scores are similar between groups, that the SMD, QWK, and Exact Agreements are also similar between the groups. When the mean scores differ between the groups, these values may differ by group, as they reflect agreement levels at different locations in the rubric scale. Across the comparisons, presented in Appendix K, almost all SMD values are within the .l5 magnitude for every item and domain. Below are the six exceptions. In these cases, aside from item 37641, Autoscore tended to assign scores that were slightly higher on average than those of human raters.

- Primary Disability Status
  - Item 37613 Development for Gifted students, with SMD = .24
  - Item 37613 Purpose/Structure for Gifted students, with SMD = .20
  - Item 37641 Development for Gifted students, with SMD = -.16
  - Item 37641 Purpose/Structure for Gifted students, with SMD = -.19
  - Item 38112 Purpose/Structure for Specific Learning Disability students, with SMD = .17
- ELL Status
  - Item 37737 Language for ELL Status = Y with SMD = .19

Particularly for gender and ethnicity, we see similar SMDs, QWK, and EA values across the demographic types (Male/Female, Black/Hispanic/White) within an item and domain. For ELL status (Y/N) and Primary Disability Status (Gifted/Specific Learning Disability), we do see more variation within item and domain, likely due to the location in the rubric scoring between the two groups, as indicated by the mean scores. Further subgroup information can be found in Appendix I of this volume.

3.5.4 Hybrid Scoring Comparison to Human and Automated Scoring

Finally, we examine the means and standard deviations of the final score on the full sample and compare that to the means and standard deviations on the randomly routed sample for both the final score (human score) and the automated score (Table 37). Across items and domains, the means scores were very similar, as were the standard deviations. In general, the full sample standard deviations were slightly smaller than the either of the random sample standard deviations.

Table 37: Means and Standard Deviations of both Final Scores and Autoscore for Full and Random Samples

Item	Domain	N		Means			Standard Deviation		
		Full Sample	Random Sample	Full Sample FS	Random Sample HS	Random Sample AS	Full Sample FS	Random Sample HS	Random Sample AS
37641	Lang.	190,561	9,407	1.99	2.03	1.99	0.70	0.74	0.72
37641	Dev.	190,561	9,407	1.95	1.97	1.95	0.68	0.70	0.69
37641	P/S	190,561	9,407	2.00	2.05	2.01	0.71	0.74	0.73

37736	Lang.	188,165	9,373	2.16	2.17	2.10	0.71	0.74	0.70
37736	Dev.	188,165	9,373	2.14	2.15	2.09	0.70	0.72	0.70
37736	P/S	188,165	9,373	2.18	2.19	2.18	0.72	0.74	0.73
38112	Lang.	192,165	9,579	2.18	2.19	2.16	0.63	0.72	0.72
38112	Dev.	192,165	9,579	2.15	2.15	2.10	0.64	0.73	0.72
38112	P/S	192,165	9,579	2.18	2.18	2.18	0.64	0.74	0.75
37678	Lang.	195,858	9,762	2.00	2.02	2.00	0.63	0.69	0.70
37678	Dev.	195,858	9,762	1.99	1.99	1.96	0.60	0.66	0.62
37678	P/S	195,858	9,762	2.03	2.05	2.07	0.62	0.67	0.69
38033	Lang.	196,115	9,936	2.20	2.20	2.20	0.68	0.71	0.71
38033	Dev.	196,115	9,936	2.18	2.17	2.16	0.68	0.72	0.73
38033	P/S	196,115	9,936	2.20	2.19	2.20	0.69	0.72	0.74
37737	Lang.	199,911	10,074	2.22	2.23	2.22	0.65	0.71	0.71
37737	Dev.	199,911	10,074	2.19	2.21	2.18	0.67	0.72	0.74
37737	P/S	199,911	10,074	2.20	2.22	2.20	0.67	0.72	0.76
37613	Lang.	202,850	10,103	2.24	2.22	2.23	0.64	0.66	0.71
37613	Dev.	202,850	10,103	2.18	2.14	2.16	0.67	0.66	0.76
37613	P/S	202,850	10,103	2.20	2.17	2.19	0.67	0.67	0.75

Note. Full Sample FS reflects final score rising out of the hybrid scoring process. Random HS represents the human score on the random sample. Random AS represents the Autoscore score on the random sample. Final-assigned condition codes are removed from the analysis, for the full sample, and for the random sample. The random sample contains Autoscore condition codes.

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, several 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 Appropriate Score Uses and Cautions for Score Use sections) 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 Florida's 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). 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 Florida's statewide assessments are designed to measure student achievement on Florida 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 Florida's 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 Florida's 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 Florida's 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 the 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 Florida's 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 the educator committees to ensure Florida's 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 Florida's 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. *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) 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 *Standards* (AERA, APA, & 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 to 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, & 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 definitions emphasize evidence and theory to support inferences and interpretations of test scores. *Standards* (AERA, APA, & 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 Florida's statewide assessment 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 provide citizens with 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 and of 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 this technical report to see the specific required uses and citations for Florida's statewide assessments.

For Florida's assessment 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's statewide assessments are represented in the following statements 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 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 following sections 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's 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 is not updated annually. Volume 3 can be found as part of *The Benchmarks for Excellent Student Thinking 2022–2023 Technical Report* and provides evidence on the validity of the process and the results of setting performance standards for Mathematics, English Language Arts (ELA), Algebra 1, and Geometry. For science and social studies, this information can be found in Chapter 5: Performance Standards from the Florida Statewide Science and EOC Assessments 2019 Technical Report.

Table 38 provides a 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 38 and in the following sections.

Table 38: Comprehensive Summary of Validity Evidence

Inferences	Claims	Evidence	Location
Scoring: Students are scored accurately and consistently.	**Model Fit.** The underlying assumptions of the item response theory (IRT) models are met. The assessments are essentially unidimensional.	o Item fit o Local independence o Confirmatory factor analysis o Ability estimate correlational analysis	o Volume 1, Sections 6.5.1 and 6.5.2 o Volume 4, Sections 4.2.2 and 4.2.3
	Scoring of Performance Tasks. The inter-rater reliability is reasonably high.	o Validity responses are provided 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 Volume 4, Section 3.5.1, Writing Handscoring Specifications* o Volume 4, Section 3.5, Autscoring Reports (Appendix I)
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, and test development o Alignment study plan o Detailed blueprints for each content level by each grade and subject	o Volume 2, Test Development o Volume 2, Section 3.7 and Appendix E Alignment Study Plan o Volume 2, Section 2.1.1, Target Blueprints, and Volume 4, Section 4.1.2
	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 study plan	o Volume 2, Section 2.1.1, Target Blueprints and Volume 4, Section 4.1.2, Section 4.1.3 o Volume 4, Section 4.3.1 o Volume 4, Appendix L, Cognitive Lab Study
	Validity Based on Relations to Other Variables. The State has documented adequate validity evidence that the State's	o Comparisons of reporting category correlations within and across subjects	o Volume 4, Section 4.3

Inferences	Claims	Evidence	Location
	assessment scores are related as expected with other variables.		
	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 Conditional standard error of measurement (CSEM) plots o Marginal reliability	o Volume 4, Section 3.2 CSEM o Volume 4, Section 3.1 Marginal Reliability
	Different Student Populations. Scores represent students in schools throughout Florida including participation from Home Education Program students, students with disabilities, English language learner (ELL) students, McKay Scholarship Program students, etc.	o Testing accommodation o Subgroup reliability	o Volume 5, Section 1.2 o Volume 4, Appendix A Reliability Coefficients, Appendix K
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 Students with Disabilities (SWD) under the Individuals with Disabilities Education Act (IDEA), students covered by Section 504, and ELLs.	o List of available accommodations o Description of accommodated form construction o Accommodated form statistics	o Volume 5, Section 1.2, Testing Accommodations and Appendix EE o Volume 1, Sections 2.2 and 6.4 o Volume 1, Appendix C o Volume 2, Section 4.4 o Volume 4, Appendix H
	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 ELLs.	o Description of ELL students and SWD o Description of available testing accommodations and practice activities	o Volume 5, Section 1.1, Eligible Students o Volume 5, Section 1.2, Testing Accommodations
	Fairness and Accessibility. Assessments are accessible to all students and are 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, Section 6.1, Fairness in Content and Section 6.2, Statistical Fairness in Item Statistics o Volume 4, Sections 5.1, 5.3, 5.4, and 5.5 o Volume 2, Section 3.4

Inferences	Claims	Evidence	Location
	Device Comparability. There are no meaningful differences in the scores for students when the tests are 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, Sections 4.3.2, 5.3, 5.4, and 5.5 o Appendix F of the *2017–2018 FSA Technical Report*: Device Comparability (Appendix D of this volume)
	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 7, Scoring and Chapter 6, Calibration Scaling o Volume 6, Section 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 confirmatory factor analysis (CFA) model	o Volume 4, Section 4.2.2, Evidence Based on Internal Structure
	Convergent 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 end-of-course (EOC)	o Volume 4, Section 4.3 Convergent and Discriminant Validity
Implication: The evidence supports the proposed use of test scores.	**Interpretation of Performance Standards.** The State uses technically sound and well-documented processes to develop scoring interpretations and performance standards.	o Standard-setting report o Achievement-Level Descriptors o Classification accuracy and consistency	o Volume 3 o Volume 6, Section 1.3, Achievement- Level Descriptors o Volume 4, Section 3.3, Classification Accuracy and Consistency
	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 Regarding the computation of the score: - A description of maximum likelihood estimation - Scale score transformation o Score interpretation guide	o Volume 1, Chapter 7, Scoring o Volume 6, Section 1.1, Overview of Florida's Score Report o Volume 6, Chapter 4, Appropriate Score Use and Chapter 5, Cautions for Score Use

**Confidential document*

4.2.2 Evidence Based on Internal Structure

Determining whether the test measures the intended construct is central to evaluating the validity of test score interpretations, and such an evaluation requires a clear definition of the measurement construct. Florida's statewide assessments represent a structural model of student achievement in grade-level and course-specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., reading prose and poetry, reading informational text, and reading across genres and vocabulary). Content strands within each subject area are, in turn, indicators of achievement in the subject area.

The assessments reported test scores as an overall performance measure in each subject area. 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 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 using these scoring and reporting methods. This section provides evidence that the methods for reporting the Florida statewide assessments strand scores align with the underlying structure of the test and provide evidence for appropriateness of the selected IRT models.

Model Fit and Scaling

IRT models provide a basis for Florida's 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. Therefore, any item displaying misfit is scrutinized before a decision is made to place the item into the item bank. Yen's (1981) Q1 and item fit plots are used to evaluate the degree to which the observed data fit the item response model. This is detailed in Volume 1, Section 6.5.1 Item Fit. Also, CAI conducts classical item analysis on field-test items to ensure that the items function as intended with respect to the underlying scales. In addition to model fit, key statistical analyses included item discrimination, distractor analysis, item difficulty analysis, and content review of items flagged by these statistical analyses by content experts. Most items in Florida's assessments display good model fit. Appendix B lists the number of field-test items by grade and subject flagged by Q1.

The validity of the application of IRT depends greatly on meeting the underlying assumptions of the models. One 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 several 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:

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

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

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

where y_{jk} is the weight for response category *k*, *m* is the number of response categories, and $P_{jk}(\hat{\theta}_i)$ is the probability of response category *k* to item *j* by 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 39 to Table 42 present average correlations of item scores between item pairs and summaries of the distributions of the Q_3 statistics—minimum, 5th percentile, median, 95th percentile, and maximum values from each grade and subject. We used the item responses from the 2022–2023 FAST and B.E.S.T. assessments and the 2023–2024 science and social studies assessments. Unlike a fixed-form test that administers the same items to all test takers, these assessments were adaptively conducted or administered randomly within the blueprint constraints.

To calculate Q_3 statistics, each item requires a paired set with every other item, so some items with a small sample size were excluded from the analysis to provide valid analysis results. We included items with a sample size of at least 1,500 and a paired item count of 150. For this reason, we do not update the Q3 analysis each year. As the operational pool size increases for adaptive tests, it will be increasingly difficult to achieve sufficient sample size. The assumption is the analysis is generalizable from year to year, because items will be based on the same test standards and blueprints each year.

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). The current Q3 statistic provides information for detecting local dependencies, but the results should be used with caution. Although the mathematics and EOC assessments administered adaptive tests, the Q3 statistic provided in this technical report did not take into account the item selection order and process applied by adaptive tests. Also, note that the Q3 statistics from the adaptive test condition have larger confidence intervals compared to traditional fixed-form tests (Mislevy et al., 2012). More careful interpretation is required.

Table 39: Mathematics Q_3 Statistic

Grade	Average Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
3	0.372	–0.244	–0.080	–0.023	0.024	0.593
4	0.415	–0.180	–0.077	–0.024	0.026	0.659
5	0.397	–0.195	–0.075	–0.025	0.025	0.550
6	0.262	–0.310	–0.090	–0.025	0.044	0.346
7	0.285	–0.293	–0.106	–0.020	0.057	0.564
8	0.247	–0.291	–0.095	–0.020	0.056	0.517

Table 40: ELA Q_3 Statistic

Grade	Average Correlation	Q3 Distribution					Within Passage Q_3*	
		Minimum	5th Percentile	Median	95th Percentile	Maximum	Minimum	Maximum
3	0.286	–0.164	–0.061	–0.024	0.008	0.229	–0.103	0.100
4	0.270	–0.146	–0.054	–0.021	0.011	0.135	–0.083	0.135
5	0.277	–0.138	–0.059	–0.022	0.010	0.092	–0.075	0.088
6	0.247	–0.187	–0.060	–0.021	0.012	0.122	–0.066	0.090
7	0.276	–0.198	–0.061	–0.023	0.008	0.137	–0.094	0.137
8	0.307	–0.172	–0.063	–0.023	0.011	0.135	–0.073	0.135
9	0.237	–0.153	–0.057	–0.022	0.005	0.080	–0.059	0.080
10	0.219	–0.226	–0.063	–0.020	0.018	0.168	–0.083	0.153

* Within Passage Q3 values are computed for each item pair within a passage.

Table 41: EOC Q_3 Statistic

Course	Average Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Algebra 1	0.289	–0.294	–0.078	–0.016	0.050	0.665
Geometry	0.269	–0.247	–0.071	–0.016	0.048	0.784

Table 42: Science and Social Studies Q_3 Statistic

Course	Average Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Biology 1	0.261	-0.326	-0.109	-0.020	0.070	0.398

Course	Average Correlation	Q_3 Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Civics	0.190	-0.202	-0.076	-0.019	0.036	0.423
U.S. History	0.237	-0.200	-0.076	-0.019	0.040	0.524
Grade 5	0.300	-0.338	-0.104	-0.021	0.063	0.474
Grade 8	0.261	-0.334	-0.095	-0.018	0.058	0.304

Confirmatory Factor Analysis

To assess the fit of the structural model to student response data from Florida's statewide assessments, a series of CFAs were conducted for each grade and subject assessment using the statistical program Mplus [version 8] (Muthén & Muthén, 2012). Mplus is commonly used for collecting validity evidence on the internal structure of assessments. Weighted least square mean 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 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 Florida's assessments 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.

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:

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

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 three categories for ELA and EOC, three to four reporting categories for mathematics, three to four for science, and three to four for social studies (see Table 76 to Table 79 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 model is illustrated in Figure 6. This figure is generally representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across grades.

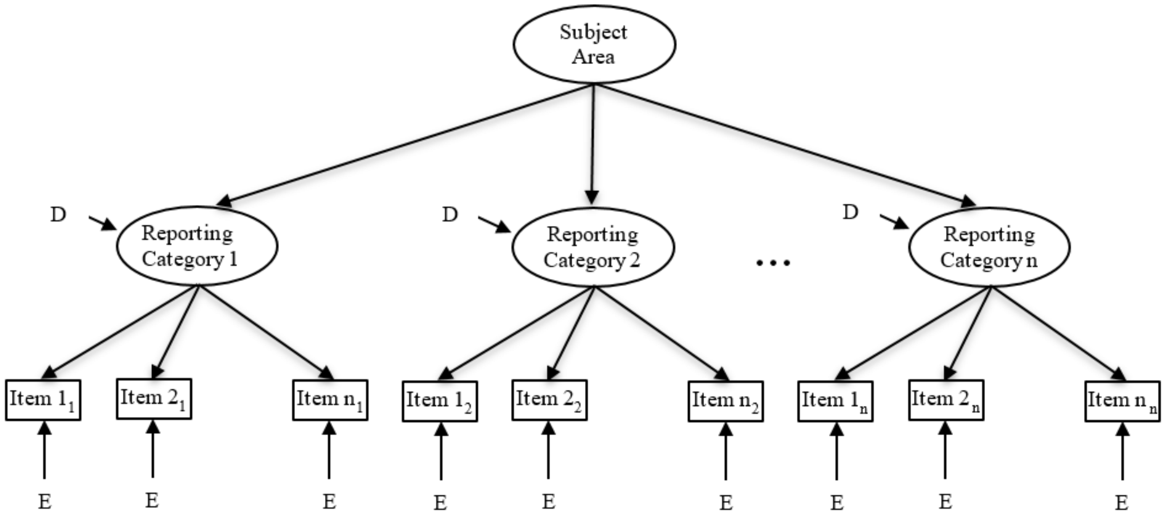
The purpose of conducting confirmatory factor analysis for Florida's assessments was to provide evidence that each individual assessment implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

The data for this analysis were taken from the 2022–2023 ELA and mathematics assessments and the 2023–2024 science and social studies assessments. They were adaptively administered for mathematics and administered randomly within the blueprint constraints for ELA, science, and social studies. In the absence of a common test form for all students, we attempted to construct a single representative form for each grade and subject comprising highly administered items that met content standard blueprint specifications. Because the ELA and mathematics assessments were administered with different test designs, we selected the representative forms of two subject areas differently. For ELA tests with four passages per student under content constraints, the set of passages with the largest number of students (containing four passages) was selected. The test score distribution of the sample was compared to the population to ensure that the sample was adequately representative of the population. For mathematics tests administered adaptively, a list of items was selected that meet the blueprints and have sufficient sample size between paired items. This ensured a well-conditioned covariance matrix comprising a sample of items representing the full breadth of the content domain specified by the blueprint. The numbers of items selected varied across tests: 43–52 items across ELA assessments, 35–36 items across mathematics assessments, and 45 items across B.E.S.T. assessments.

The same method of selecting the most highly administered items representing the blueprint was also applied to the science and social studies assessments. However, as these item banks are much larger (ranging from 548 to 838 items), the sample size between paired items was not sufficient to produce analyses that could converge for the larger banks. The two smallest banks, Civics and U.S. History, produced converged results but with warning flags in Mplus that suggest multiple problems with the solutions. For this reason, the analysis was not repeated for the most recent administration for ELA and mathematics – as the banks grow larger, the sample size requirement for item pairs becomes increasingly difficult to meet.

Evidence for the structural model for multiple fixed-form versions of Florida’s science and social studies assessments can be found in previous years’ versions of the Florida’s technical reports, dating back to 2015, with the latest being the *Florida Statewide Assessments Science and Social Studies 2022–2023 Technical Report: Volume 4*. In all scenarios, the empirical results suggested the implied model fits the data well. These results indicated that reporting an overall score in addition to separate scores for the individual reporting categories was reasonable. These previous fixed-form versions share the same content standards, blueprints, and item bank as the current adaptive assessments. Thus, they also provide evidence for the structural model for the current adaptive assessments. In addition, data for the fixed-form TTS 2023–2024 science and social studies assessments were analyzed.

Figure 6: Second-Order Factor Model



Several goodness-of-fit statistics from each of the analyses are presented in the following tables. 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 above 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 above 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.

We began by evaluating the fit of the first-order, general achievement model in which all items are indicators of a common subject area factor. This model evaluates the assumption of unidimensionality of the subject-area assessments and provides a baseline for evaluating the improvement of fit for the more differentiated second order (i.e., strand) model. The goodness-of-fit statistics for the first-order, general achievement models are shown in Table 43 to Table 45 All the statistics indicate that the general achievement factor model fits the data well across all subject areas and grades. The CFI and TLI values were all greater than 0.95, except for grades 6 and 8 mathematics, which had slightly lower values of 0.91 and 0.84 for CFI and 0.91 and 0.83 for TLI. The RMSEA values were at or below 0.02, indicating reasonable fit for the base model. The goodness-of-fit statistics for the hypothesized second-order models are also shown in Table 43 to Table 45. All the statistics indicate that the second-order models posited by the assessments fit the data well. The CFI and TLI values for the second-order models were all equal to or greater than 0.95, except for grades 6 and 8 mathematics, which had slightly lower values of 0.92 and 0.88 for CFI and 0.91 and 0.88 for TLI. The RMSEA values well below the 0.02 threshold used indicated good fit.

In addition to testing the goodness-of-fit of the first and second-order models, we examined the degree to which the second-order model improved fit over the more general one-factor model (i.e., first-order model) of academic achievement in each subject area. The one-factor, general achievement model was nested within the second-order model. A simple likelihood ratio test was used to determine whether the added information provided by the structure of the assessments' frameworks improved model fit over a general achievement model. The results of the comparison between the second-order model and the more general achievement model are presented in Table 43 to Table 45 We note that model fit for first-order models of general achievement are reasonably high and provide evidence for the unidimensionality of the subject-area assessments. The purpose of these analyses is to determine whether the posited second-order reporting model adds information beyond that provided by the first-order model. The chi-square difference test shows that across all subject areas and grade levels, the strand-based, second-order model showed significantly better fit than the general achievement first-order model. The χ2 *p*-value for the difference test was less than 0.001 across all grade levels and 0.003 for grade 10 ELA. Results indicating improved model fit for the second-order factor model provide support for the interpretation of learning standard performance at the strand level above that provided by the overall subject-area score. Given the sensitivity of the χ2 difference tests to sample size, some caution is needed. The CFI, TLI and RMSEA values for both models are extremely similar (less than 0.01 difference). So, while there is evidence from the χ2 difference that the second order adds more information, based on other statistics, there is little difference between the first and second models. Nevertheless, both models fit the data well, meaning either model can be used in practice. Since the second order model provides greater interpretability of students' individual abilities (via the reporting categories), while a single score may obscure important differences between the constructs, the second order model is preferred.

Table 43: Goodness-of-Fit Second-Order CFA (ELA)

Grade/Course	**Goodness-of-Fit**						**Difference in Fit between First- and Second-Order Models**		
	First-Order Models			**Second-Order Models**					
	CFI	**TLI**	**RMSEA**	**CFI**	**TLI**	**RMSEA**	χ^2	***df***	***p* value**
Grade 3	0.985	0.984	0.013	0.986	0.985	0.013	54.039	2*	< 0.001
Grade 4	0.976	0.975	0.015	0.978	0.977	0.015	195.986	2*	< 0.001
Grade 5	0.985	0.985	0.016	0.986	0.985	0.016	230.295	3	< 0.001
Grade 6	0.984	0.983	0.016	0.984	0.983	0.016	26.189	2*	< 0.001
Grade 7	0.986	0.986	0.014	0.987	0.986	0.013	21.920	3	< 0.001
Grade 8	0.983	0.983	0.015	0.984	0.983	0.015	54.167	2*	< 0.001
Grade 9	0.987	0.987	0.012	0.988	0.987	0.011	33.398	2*	< 0.001
Grade 10	0.982	0.981	0.017	0.982	0.981	0.017	12.001	2*	0.003

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

Table 44: Goodness-of-Fit Second-Order CFA (Mathematics)

Grade/Course	Goodness-of-Fit						Difference in Fit between First- and Second-Order Models		
	First-Order Models			Second-Order Models					
	CFI	TLI	RMSEA	CFI	TLI	RMSEA	χ^2	*df*	*p* value
Grade 3	0.980	0.979	0.010	0.984	0.982	0.009	1873.19	3*	< 0.001
Grade 4	0.986	0.985	0.009	0.987	0.986	0.009	359.97	3	< 0.001
Grade 5	0.991	0.990	0.008	0.991	0.990	0.008	95.66	4	< 0.001
Grade 6	0.910	0.905	0.017	0.915	0.910	0.016	1920.08	2*	< 0.001
Grade 7	0.965	0.963	0.008	0.975	0.974	0.007	1505.77	4	< 0.001
Grade 8	0.840	0.831	0.012	0.883	0.875	0.011	2637.46	4	< 0.001
Algebra 1	0.965	0.964	0.014	0.966	0.964	0.014	980.27	3	< 0.001
Geometry	0.949	0.947	0.015	0.950	0.948	0.015	639.93	2*	< 0.001

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

Table 45: Goodness-of-Fit Second-Order CFA (Science and Social Studies TTS)

Grade/Course	Goodness-of-Fit						Difference in Fit between First- and Second-Order Models		
	First-Order Models			Second-Order Models					
	CFI	TLI	RMSEA	CFI	TLI	RMSEA	χ^2	*df*	*p* value
Biology 1	0.966	0.965	0.020	0.969	0.967	0.019	407.633	3	< 0.001
U.S. History	0.964	0.963	0.019	0.967	0.966	0.018	314.263	3	< 0.001
Civics	0.970	0.968	0.019	0.972	0.971	0.018	675.520	4	< 0.001
Grade 5 Science	0.983	0.982	0.016	0.984	0.983	0.015	249.858	4	< 0.001
Grade 8 Science	0.981	0.980	0.016	0.982	0.981	0.016	205.166	4	< 0.001

The second-order factor model converged for all tests (except science and social studies adaptive assessments). However, the residual variance for some factors fell slightly below the boundary of zero for grades 3, 4, 6, 8, 9, and 10 ELA, grades 3 and 6 mathematics, and Geometry 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.

Items were 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 three-parameter logistic (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 46 to Table 49, we provide the estimated factor correlations between the reporting categories from the second-order factor model. In all cases, these correlations are very high. However, the results provide empirical evidence that there is some higher than expected dimensionality among reporting categories in grade 8 mathematics, where the correlations go as low as 0.58. CAI and FDOE will continue to monitor this issue in future research studies. Correlations for other grades are in the expected range.

Table 46: Correlations Among Mathematics Factors

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4
3	Number Sense and Additive Reasoning (Cat1)	1.00			
	Number Sense and Multiplicative Reasoning (Cat2)	0.93	1.00		
	Fractional Reasoning (Cat3)	0.88	0.88	1.00	
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	0.97	0.96	0.91	1.00
4	Number Sense and Operations with Whole Numbers (Cat1)	1.00			
	Number Sense and Operations with Fractions and Decimals (Cat2)	0.94	1.00		
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat3)	0.97	0.93	1.00	
5	Number Sense and Operations with Whole Numbers (Cat1)	1.00			
	Number Sense and Operations with Fractions and Decimals (Cat2)	0.98	1.00		
	Algebraic Reasoning (Cat3)	0.95	0.97	1.00	
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	0.95	0.97	0.95	1.00
6	Number Sense and Operations (Cat1)	1.00			
	Algebraic Reasoning (Cat2)	0.97	1.00		
	Geometric Reasoning, Data Analysis, and Probability (Cat3)	0.89	0.86	1.00	
7	Number Sense and Operations and Algebraic Reasoning (Cat1)	1.00			

Grade	Reporting Category	Cat1	Cat2	Cat3	Cat4
	Proportional Reasoning and Relationships (Cat2)	0.95	1.00		
	Geometric Reasoning (Cat3)	0.94	0.92	1.00	
	Data Analysis and Probability (Cat4)	0.90	0.88	0.87	1.00
8	Number Sense and Operations and Probability (Cat1)	1.00			
	Algebraic Reasoning (Cat2)	0.65	1.00		
	Linear Relationships, Data Analysis, and Function (Cat3)	0.75	0.58	1.00	
	Geometric Reasoning (Cat4)	0.89	0.70	0.79	1.00

Table 47: Correlations Among ELA Factors

Grade	Reporting Category	Cat1	Cat2	Cat3
3	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.94	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.96	0.98	1.00
4	Reading Prose and Poetry (Cat1)	1.00	0.91	
	Reading Informational Text (Cat2)	0.91	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.95	0.95	1.00
5	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.95	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.96	0.99	1.00
6	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.97	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.99	0.99	1.00
7	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.95	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.98	0.96	1.00
8	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.95	1.00	
	Reading Across Genres and Vocabulary (Cat3)	0.99	0.96	1.00
9	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.97	1.00	

Grade	Reporting Category	Cat1	Cat2	Cat3
	Reading Across Genres and Vocabulary (Cat3)	1.00	0.97	1.00
10	Reading Prose and Poetry (Cat1)	1.00		
	Reading Informational Text (Cat2)	0.98	1.00	
	Reading Across Genres and Vocabulary (Cat3)	1.00	0.98	1.00

Table 48: Correlations Among EOC Factors

Course/Form	Reporting Category	Cat1	Cat2	Cat3
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)	1.00		
	Linear Relationships (Cat2)	0.95	1.00	
	Non-Linear Relationships (Cat3)	0.94	0.96	1.00
Geometry	Logic, Relationships, and Theorems (Cat1)	1.00		
	Congruence, Similarity, and Constructions (Cat2)	0.97	1.00	
	Measurement and Coordinate Geometry (Cat3)	0.98	0.99	1.00

Table 49: Correlations Among Science and Social Studies (TTS) Factors

Grade/Course	Reporting Category	Cat1	Cat2	Cat3	Cat4
Biology 1	Molecular and Cellular Biology (Cat1)	1.00			
	Classification, Heredity, and Evolution (Cat2)	0.95	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	0.98	0.97	1.00	
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00			
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	1.01	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	0.98	1.00	1.00	
Civics	Origins and Purposes of Law and Government (Cat1)	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	0.93	1.00		

Grade/Course	Reporting Category	Cat1	Cat2	Cat3	Cat4
	Government Policies and Political Processes (Cat3)	0.98	0.95	1.00	
	Organization and Function of Government (Cat4)	0.97	0.94	0.99	1.00
Grade 5 Science	Earth and Space Sciences (Cat1)	1.00			
	Life Sciences (Cat2)	0.97	1.00		
	Nature of Science (Cat3)	0.93	0.95	1.00	
	Physical Sciences (Cat4)	0.96	0.98	0.94	1.00
Grade 8 Science	Earth and Space Sciences (Cat1)	1.00			
	Life Sciences (Cat2)	0.99	1.00		
	Nature of Science (Cat3)	0.99	0.97	1.00	
	Physical Sciences (Cat4)	0.98	0.97	0.97	1.00

Measurement Invariance Across Subgroups

This technical report provides the differential item functional analysis across demographic subgroups to identify potential bias at an item level (see Volume 1, Section 5.2 Differential Item Functioning Analyses). Furthermore, we conducted measurement invariance analysis in a more comprehensive way at the test level to ensure that the tests measure the same constructs across subgroups. Measurement invariance occurs when the likelihood of responding correctly conforms to the measurement model and is independent of group membership, and the parameters of a measurement model are statistically equivalent across groups. In general, measurement invariance testing can be conducted using a series of multiple-group CFA models, which impose identical parameters across groups. That is, the models that investigate the invariance of factor pattern (configural invariance), factor loadings (metric or weak invariance), latent intercepts/threshold (scalar or strong invariance), and unique or residual factor variances (strict invariance) are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups, and the latent variables can be used in structural models that hypothesize relationships among latent variables.

The full set of tables associated with these analyses is provided in Appendix G, Measurement Invariance Testing, for each of the assessments. The series A tables show a general approach of testing measurement invariance – evaluating model fit differences between less restricted and more restricted models, assuming continuous outcome variables. The series B tables treat the outcome variables as categorical variables. Since chi-square tests generally tend to be rejected when data are categorical, we reviewed the measurement invariance tests provided by Mplus software and further constructed a model fit analysis for the most restricted model (scalar) to provide evidence of the measurement invariance of the less restricted models.

The series A tables in Appendix G present the global model fit indices for the measurement invariance tests for each assessment. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using the $\chi2$

difference test and the examination of significant differences of the RMSEA (RMSEA, change in RMSEA ≤ 0.015; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across the following subgroups: gender (Model A), ethnicity (African American versus White and Hispanic versus White in Model B), Disability (Model C), and ELL status (Model D). Invariance tests of subgroups were investigated separately for each grade and subject-area test. There were several assessments that had subgroups for which the measurement invariance analysis did not converge, and this was mostly due to small sample sizes or sparse data.

The null hypothesis of the $\chi 2$ difference test is that the more restricted invariance model (e.g., metric) fits the data equally as well as the less restricted invariance model (e.g., configural). Given the sensitivity of the $\chi 2$ difference tests to sample size, we additionally examined significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retention of the more restricted invariance model (Chen, 2007). For all subject and grade assessments, the RMSEA change ranges were very small, with a maximum change of 0.002 in ELA and 0.004 in mathematics and EOC, and 0.001 in science TTS and Social Studies TTS.

Although the $\chi 2$ difference test should ideally be nonsignificant, all $\chi 2$ difference tests were significant or marginally significant at $\alpha = 0.05$ due to large sample sizes. Nevertheless, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0–0.002 for ELA, from 0–0.004 for mathematics, 0-0.001 for science TTS, and 0–0.001 for social studies TTS). Based on the similar magnitudes of the RMSEA (i.e., no material changed across all tested models; Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, the test scores have the same measurement structure across gender, ethnicity (classified as White, African American, or Hispanic), disability, and ELL status for each test.

In addition to evaluating the differences in model fit between less restricted and more restricted invariance models shown in series A, we further constructed a model fit analysis of a scalar invariance model. The scalar invariance model is the most restricted model that we constructed for evaluating the measurement invariance, so demonstrating a good model fit for the scalar invariance model is not limited to measurement invariance of the scalar model and confirms measurement invariance for less restricted invariance models. The series B tables in Appendix G show the model fit indices of scalar invariance models assuming the same factor pattern plus identical factor loadings plus identical latent intercept/threshold across subgroups. Global model fit indices included the CFI (Bentler, 1990) and RMSEA. CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested an acceptable fit to the data. For ELA, CFI values ranged from 0.97–0.99, and RMSEA values ranged from 0.009–0.017. For mathematics and EOC, CFI values ranged from 0.90–0.98, except for grade 8, and RMSEA values ranged from 0.007–0.017 for all grades. CFI values for grade 8 mathematics ranged from 0.81–0.83 across models, indicating unacceptable fit, although RMSEA values ranged from 0.012–0.013, indicating acceptable model fit. For science TTS, CFI values ranged from 0.913–0.953 and RMSEA ranged from 0.016–0.020. In social studies TTS, CFI ranged from 0.914–0.938 and RMSEA ranged from 0.018–0.020.

4.2.3 Correlations Among Reporting Category Scores

In this section, we explore the internal structure of Florida's statewide assessments using the scores provided at the reporting category level. It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional IRT model difficult, though reporting these separate scores could then easily be justified. On the contrary, if the reporting categories were perfectly correlated, a unidimensional model could be justified, but the reporting of separate scores could not.

One pathway to explore the internal structure of the test using subscale scores is to explore observed correlations between the subscores. Theta scores for each reporting category were used for this analysis. Again, the items in each reporting category were administered within the constraints of the blueprint, and the scores for each reporting category were based on the same scoring scale. 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.

The observed correlations among reporting category scores are presented in Table 50 to Table 53. In ELA, the observed correlations among the reporting categories range from 0.63–0.74. For mathematics, the observed correlations were between 0.37–0.80. For EOC mathematics, they were between 0.72–0.81. Finally, science and social studies ranged from 0.63-0.71.

Table 50: Observed Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Number Sense and Additive Reasoning (Cat1)	116	1.00				
	Number Sense and Multiplicative Reasoning (Cat2)	118	0.71	1.00			
	Fractional Reasoning (Cat3)	69	0.69	0.67	1.00		
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	141	0.74	0.71	0.68	1.00	
4	Number Sense and Operations with Whole Numbers (Cat1)	128	1.00				
	Number Sense and Operations with Fractions and Decimals (Cat2)	106	0.78	1.00			
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat3)	96	0.77	0.77	1.00		
5	Number Sense and Operations with Whole Numbers (Cat1)	92	1.00				
	Number Sense and Operations with Fractions and Decimals (Cat2)	128	0.72	1.00			
	Algebraic Reasoning (Cat3)	88	0.69	0.70	1.00		
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	142	0.71	0.73	0.71	1.00	
6	Number Sense and Operations (Cat1)	149	1.00				
	Algebraic Reasoning (Cat2)	155	0.80	1.00			
	Geometric Reasoning, Data Analysis, and Probability (Cat3)	146	0.75	0.73	1.00		
7	Number Sense and Operations and Algebraic Reasoning (Cat1)	94	1.00				
	Proportional Reasoning and Relationships (Cat2)	77	0.54	1.00			
	Geometric Reasoning (Cat3)	88	0.50	0.46	1.00		
	Data Analysis and Probability (Cat4)	108	0.53	0.53	0.45	1.00	
8	Number Sense and Operations and Probability (Cat1)	87	1.00				
	Algebraic Reasoning (Cat2)	60	0.46	1.00			
	Linear Relationships, Data Analysis, and Function (Cat3)	74	0.44	0.47	1.00		
	Geometric Reasoning (Cat4)	66	0.37	0.38	0.40	1.00	

Table 51: Observed Correlation Matrix Among Reporting Categories (ELA Reading)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Reading Prose and Poetry (Cat1)	100	1.00				
	Reading Informational Text (Cat2)	100	0.65	1.00			
	Reading Across Genres and Vocabulary (Cat3)	156	0.72	0.71	1.00		
4	Reading Prose and Poetry (Cat1)	141	1.00				
	Reading Informational Text (Cat2)	91	0.64	1.00			
	Reading across Genres and Vocabulary (Cat3)	200	0.69	0.68	1.00		
5	Reading Prose and Poetry (Cat1)	120	1.00				
	Reading Informational Text (Cat2)	121	0.68	1.00			
	Reading Across Genres and Vocabulary (Cat3)	201	0.73	0.74	1.00		
6	Reading Prose and Poetry (Cat1)	88	1.00				
	Reading Informational Text (Cat2)	103	0.63	1.00			
	Reading Across Genres and Vocabulary (Cat3)	166	0.68	0.70	1.00		
7	Reading Prose and Poetry (Cat1)	90	1.00				
	Reading Informational Text (Cat2)	105	0.66	1.00			
	Reading Across Genres and Vocabulary (Cat3)	179	0.73	0.69	1.00		
8	Reading Prose and Poetry (Cat1)	79	1.00				
	Reading Informational Text (Cat2)	100	0.65	1.00			
	Reading Across Genres and Vocabulary (Cat3)	152	0.71	0.72	1.00		
9	Reading Prose and Poetry (Cat1)	92	1.00				
	Reading Informational Text (Cat2)	132	0.65	1.00			
	Reading Across Genres and Vocabulary (Cat3)	165	0.70	0.72	1.00		
10	Reading Prose and Poetry (Cat1)	100	1.00				

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Reading Informational Text (Cat2)	113	0.63	1.00			
	Reading Across Genres and Vocabulary (Cat3)	178	0.69	0.71	1.00		

Table 52: Observed Correlation Matrix Among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)	99	1.00			
	Linear Relationships (Cat2)	108	0.78	1.00		
	Non-Linear Relationships (Cat3)	111	0.73	0.72	1.00	
Geometry	Logic, Relationships, and Theorems (Cat1)	118	1.00			
	Congruence, Similarity, and Constructions (Cat2)	114	0.81	1.00		
	Measurement and Coordinate Geometry (Cat3)	133	0.81	0.81	1.00	

Table 53: Observed Correlation Matrix Among Reporting Categories (Science and Social Studies)

Subject	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Grade 5 Science	Earth and Space Sciences (Cat1)	121	1.00			
	Nature of Science (Cat2)	240	0.63	1.00		
	Physical Sciences (Cat3)	232	0.65	0.70	1.00	
	Life Sciences (Cat4)	209	0.63	0.69	0.68	1.00
Grade 8 Science	Earth and Space Sciences (Cat1)	119	1.00			
	Nature of Science (Cat2)	204	0.65	1.00		
	Physical Sciences (Cat3)	170	0.65	0.68	1.00	
	Life Sciences (Cat4)	201	0.64	0.68	0.67	1.00
Biology 1	Molecular and Cellular Biology (Cat1)	288	1.00			
	Classification, Heredity, and Evolution (Cat2)	202	0.66	1.00		
	Organisms, Populations, and Ecosystems (Cat3)	348	0.71	0.70	1.00	
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	197	1.00			

Subject	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	212	0.70	1.00		
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	206	0.70	0.70	1.00	
Civics	Origins and Purposes of Law and Government (Cat1)	149	1.00			
	Roles, Rights, and Responsibilities of Citizens (Cat2)	150	0.68	1.00		
	Government Policies and Political Processes (Cat3)	111	0.67	0.65	1.00	
	Organization and Function of Government (Cat4)	136	0.65	0.62	0.61	1.00

The disattenuated correlations were between 0.92–1.00 for ELA, 0.60–1.00 for mathematics, 0.98–1.00 for EOC mathematics, and 0.90–1.00 for science and social studies, as presented in Table 54 to Table 57. The same tables are available for accommodated forms in Appendix H. 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. Per convention, all disattenuated correlations above 1.0 were capped at 1.0.

Table 54: Disattenuated Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Number Sense and Additive Reasoning (Cat1)	116	1.00				
	Number Sense and Multiplicative Reasoning (Cat2)	118	0.98	1.00			
	Fractional Reasoning (Cat3)	69	0.99	0.95	1.00		
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	141	1.00	0.98	0.98	1.00	
4	Number Sense and Operations with Whole Numbers (Cat1)	128	1.00				
	Number Sense and Operations with Fractions and Decimals (Cat2)	106	1.00	1.00			
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat3)	96	1.00	1.00	1.00		
5	Number Sense and Operations with Whole Numbers (Cat1)	92	1.00				

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Number Sense and Operations with Fractions and Decimals (Cat2)	128	1.00	1.00			
	Algebraic Reasoning (Cat3)	88	0.97	0.98	1.00		
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)	142	1.00	1.00	0.98	1.00	
6	Number Sense and Operations (Cat1)	149	1.00				
	Algebraic Reasoning (Cat2)	155	1.00	1.00			
	Geometric Reasoning, Data Analysis, and Probability (Cat3)	146	0.97	0.97	1.00		
7	Number Sense and Operations and Algebraic Reasoning (Cat1)	94	1.00				
	Proportional Reasoning and Relationships (Cat2)	77	0.83	1.00			
	Geometric Reasoning (Cat3)	88	0.73	0.73	1.00		
	Data Analysis and Probability (Cat4)	108	0.81	0.87	0.71	1.00	
8	Number Sense and Operations and Probability (Cat1)	87	1.00				
	Algebraic Reasoning (Cat2)	60	0.70	1.00			
	Linear Relationships, Data Analysis, and Function (Cat3)	74	0.74	0.82	1.00		
	Geometric Reasoning (Cat4)	66	0.60	0.65	0.74	1.00	

Table 55: Disattenuated Correlation Matrix Among Reporting Categories (ELA Reading)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
3	Reading Prose and Poetry (Cat1)	100	1.00				
	Reading Informational Text (Cat2)	100	0.94	1.00			
	Reading Across Genres and Vocabulary (Cat3)	156	0.98	0.98	1.00		
4	Reading Prose and Poetry (Cat1)	141	1.00				
	Reading Informational Text (Cat2)	91	0.95	1.00			
	Reading Across Genres and Vocabulary (Cat3)	200	1.00	1.00	1.00		
5	Reading Prose and Poetry (Cat1)	120	1.00				

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Reading Informational Text (Cat2)	121	0.99	1.00			
	Reading Across Genres and Vocabulary (Cat3)	201	1.00	1.00	1.00		
6	Reading Prose and Poetry (Cat1)	88	1.00				
	Reading Informational Text (Cat2)	103	0.92	1.00			
	Reading Across Genres and Vocabulary (Cat3)	166	0.96	1.00	1.00		
7	Reading Prose and Poetry (Cat1)	90	1.00				
	Reading Informational Text (Cat2)	105	0.96	1.00			
	Reading Across Genres and Vocabulary (Cat3)	179	1.00	0.98	1.00		
8	Reading Prose and Poetry (Cat1)	79	1.00				
	Reading Informational Text (Cat2)	100	0.96	1.00			
	Reading Across Genres and Vocabulary (Cat3)	152	1.00	1.00	1.00		
9	Reading Prose and Poetry (Cat1)	92	1.00				
	Reading Informational Text (Cat2)	132	0.97	1.00			
	Reading Across Genres and Vocabulary (Cat3)	165	1.00	1.00	1.00		
10	Reading Prose and Poetry (Cat1)	100	1.00				
	Reading Informational Text (Cat2)	113	0.96	1.00			
	Reading Across Genres and Vocabulary (Cat3)	178	1.00	1.00	1.00		

Table 56: Disattenuated Correlation Matrix Among Reporting Categories (EOC)

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)	99	1.00			
	Linear Relationships (Cat2)	108	1.00	1.00		
	Non-Linear Relationships (Cat3)	111	0.98	0.98	1.00	
Geometry	Logic, Relationships, and Theorems (Cat1)	118	1.00			

Course/Form	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Congruence, Similarity, and Constructions (Cat2)	114	1.00	1.00		
	Measurement and Coordinate Geometry (Cat3)	133	1.00	1.00	1.00	

Table 57: Disattenuated Correlation Matrix Among Reporting Categories (Science and Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
Grade 5 Science	Nature of Science (Cat1)	121	1.00				
	Earth and Space Sciences (Cat2)	240	1.00	1.00			
	Physical Sciences (Cat3)	232	1.00	1.00	1.00		
	Life Sciences (Cat4)	209	1.00	1.00	1.00	1.00	
Grade 8 Science	Nature of Science (Cat1)	119	1.00				
	Earth and Space Sciences (Cat2)	204	1.00	1.00			
	Physical Sciences (Cat3)	170	1.00	1.00	1.00		
	Life Sciences (Cat4)	201	1.00	1.00	1.00	1.00	
Biology 1	Molecular and Cellular Biology (Cat1)	288	1.00				
	Classification, Heredity, and Evolution (Cat2)	202	0.97	1.00			
	Organisms, Populations, and Ecosystems (Cat3)	348	0.99	1.00	1.00		
Civics	Origins and Purposes of Law and Government (Cat1)	149	1.00				
	Roles, Rights, and Responsibilities of Citizens (Cat2)	150	1.00	1.00			
	Government Policies and Political Processes (Cat3)	111	0.97	0.99	1.00		
	Organization and Function of Government (Cat4)	136	0.93	0.93	0.90	1.00	
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	197	1.00				
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)	212	0.99	1.00			
	The United States and the Defense of the International Peace, 1940–Present (Cat3)	206	0.99	1.00	1.00		

4.3 Convergent and Discriminant Validity

According to Standard 1.16 of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), evidence must be provided of convergent and discriminant validity, a part of validity evidence demonstrating that assessment scores are related as expected with criterion and other variables for all student groups. Convergent evidence supports the relationship between measures assessing the same construct, while discriminant evidence distinguishes the test from other measures assessing different constructs. However, a second, independent test measuring the same constructs as the statewide assessments in Florida during the same time period, which could easily permit for a cross-test set of correlations, was not available. Therefore, as an alternative, the correlations between the subscores (theta scores for each reporting category) within and across mathematics, English language arts (ELA), science, and social studies were examined. 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, ELA). These correlations are based on a small number of items; consequently, 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 correlations and the disattenuated correlations are provided. Generally, 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. Per convention, all disattenuated correlations above 1.0 were capped at 1.0.

The correlations among reporting category scores, both observed and corrected for attenuation, are presented in Table 58 to Table 71. The same analysis could not be repeated for accommodated forms due to the small number of students who take the forms, resulting in an even smaller overlap between those who take common subjects.

Table 58: Grade 3 Observed Score Correlations

Subject	Reporting Category	Mathematics				ELA Reading		
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3
Mathematics	Number Sense and Additive Reasoning (Cat1)	1.00	0.71	0.69	0.74	0.57	0.58	0.63
	Number Sense and Multiplicative Reasoning (Cat2)		1.00	0.67	0.71	0.53	0.53	0.57
	Fractional Reasoning (Cat3)			1.00	0.68	0.55	0.55	0.60
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)				1.00	0.57	0.58	0.62
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.65	0.72
	Reading Informational Text (Cat2)						1.00	0.71
	Reading Across Genres and Vocabulary (Cat3)							1.00

Table 59: Grade 3 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics				ELA Reading		
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3
Mathematics	Number Sense and Additive Reasoning (Cat1)	1.00	0.98	0.99	1.00	0.80	0.81	0.84
	Number Sense and Multiplicative Reasoning (Cat2)		1.00	0.95	0.98	0.74	0.74	0.76
	Fractional Reasoning (Cat3)			1.00	0.98	0.80	0.81	0.84
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)				1.00	0.80	0.82	0.84
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.93	0.98
	Reading Informational Text (Cat2)						1.00	0.98
	Reading Across Genres and Vocabulary (Cat3)							1.00

Table 60: Grade 4 Observed Score Correlations

Subject	Reporting Category	Mathematics			ELA Reading		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
Mathematics	Number Sense and Operations with Whole Numbers (Cat1)	1.00	0.78	0.77	0.57	0.57	0.62
	Number Sense and Operations with Fractions and Decimals (Cat2)		1.00	0.77	0.55	0.55	0.59
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat3)			1.00	0.56	0.57	0.61
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.64	0.69
	Reading Informational Text (Cat2)					1.00	0.68
	Reading Across Genres and Vocabulary (Cat3)						1.00

Table 61: Grade 4 Disattenuated Score Correlations

Subject	Reporting Category	Mathematics			ELA Reading		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
Math	Number Sense and Operations with Whole Numbers (Cat1)	1.00	1.00	1.00	0.79	0.80	0.85
	Number Sense and Operations with Fractions and Decimals (Cat2)		1.00	1.00	0.76	0.77	0.81
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat3)			1.00	0.79	0.81	0.85
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.95	1.00
	Reading Informational Text (Cat2)					1.00	0.99
	Reading Across Genres and Vocabulary (Cat3)						1.00

Table 62: Grade 5 Observed Score Correlations

Subject	Reporting Category	Science				ELA Reading			Mathematics			
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 4
Science	Nature of Science (Cat1)	1.00	0.63	0.65	0.63	0.59	0.61	0.65	0.55	0.56	0.57	0.59
	Earth and Space Sciences (Cat2)		1.00	0.70	0.69	0.58	0.61	0.65	0.59	0.59	0.59	0.62
	Physical Sciences (Cat3)			1.00	0.68	0.60	0.63	0.67	0.58	0.58	0.59	0.62
	Life Sciences (Cat4)				1.00	0.59	0.61	0.65	0.55	0.55	0.56	0.59
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.68	0.73	0.53	0.53	0.55	0.56
	Reading Informational Text (Cat2)						1.00	0.74	0.55	0.56	0.57	0.59
	Reading Across Genres and Vocabulary (Cat3)							1.00	0.58	0.58	0.60	0.62
Mathema tics	Number Sense and Additive Reasoning (Cat1)								1.00	0.72	0.69	0.71
	Number Sense and Multiplicative Reasoning (Cat2)									1.00	0.70	0.73
	Fractional Reasoning (Cat3)										1.00	0.71
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)											1.00

Table 63: Grade 5 Disattenuated Score Correlations

Subject	Reporting Category	Science				ELA Reading			Math			
		Rep 1	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**
Science	Nature of Science (Cat1)	1.00	0.82	0.84	0.82	0.76	0.79	0.80	0.70	0.71	0.72	0.74
	Earth and Space Sciences (Cat2)		1.00	0.88	0.88	0.73	0.76	0.79	0.73	0.72	0.72	0.76
	Physical Sciences (Cat3)			1.00	0.87	0.75	0.78	0.81	0.71	0.72	0.72	0.76
	Life Sciences (Cat4)				1.00	0.74	0.77	0.80	0.68	0.68	0.69	0.73
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.84	0.87	0.64	0.65	0.66	0.67
	Reading Informational Text (Cat2)						1.00	0.88	0.68	0.68	0.69	0.71
	Reading Across Genres and Vocabulary (Cat3)							1.00	0.69	0.69	0.71	0.72
Mathematics	Number Sense and Additive Reasoning (Cat1)								1.00	0.86	0.83	0.86
	Number Sense and Multiplicative Reasoning (Cat2)									1.00	0.84	0.87
	Fractional Reasoning (Cat3)										1.00	0.84
	Geometric Reasoning, Measurement, Data Analysis, and Probability (Cat4)											1.00

Table 64: Grade 7 All Subjects Observed Score Correlations

Subject	Reporting Category	Civics				ELA Reading			Math			
		Rep 1	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**
Civics	Origins and Purposes of Law and Government (Cat1)	1.00	0.68	0.67	0.65	0.60	0.60	0.65	0.47	0.45	0.41	0.47
	Roles, Rights, and Responsibilities of Citizens (Cat2)		1.00	0.65	0.62	0.60	0.59	0.64	0.45	0.44	0.39	0.47
	Government Policies and Political Processes (Cat3)			1.00	0.61	0.59	0.59	0.63	0.43	0.43	0.38	0.46
	Organization and Function of Government (Cat4)				1.00	0.54	0.54	0.58	0.41	0.41	0.37	0.43
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.66	0.73	0.47	0.44	0.39	0.47
	Reading Informational Text (Cat2)						1.00	0.69	0.45	0.42	0.37	0.46
	Reading Across Genres and Vocabulary (Cat3)							1.00	0.51	0.48	0.42	0.52
Mathematics	Number Sense and Operations and Algebraic Reasoning (Cat1)								1.00	0.54	0.50	0.53

	Proportional Reasoning and Relationships (Cat2)									1.00	0.46	0.53
	Geometric Reasoning (Cat3)										1.00	0.45
	Data Analysis and Probability (Cat4)											1.00

Table 65: Grade 7 All Subjects Disattenuated Score Correlations

Subject	Reporting Category	Civics				ELA Reading			Math			
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 4
Civics	Origins and Purposes of Law and Government (Cat1)	1.00	0.84	0.82	0.79	0.73	0.74	0.77	0.56	0.57	0.50	0.60
	Roles, Rights, and Responsibilities of Citizens (Cat2)		1.00	0.82	0.77	0.74	0.75	0.78	0.56	0.58	0.50	0.61
	Government Policies and Political Processes (Cat3)			1.00	0.76	0.72	0.73	0.76	0.53	0.55	0.47	0.59
	Organization and Function of Government (Cat4)				1.00	0.66	0.67	0.70	0.50	0.52	0.46	0.54
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.81	0.87	0.56	0.56	0.48	0.60
	Reading Informational Text (Cat2)						1.00	0.84	0.55	0.54	0.47	0.59
	Reading Across Genres and Vocabulary (Cat3)							1.00	0.60	0.59	0.51	0.64
Math	Number Sense and Operations and Algebraic Reasoning (Cat1)								1.00	0.69	0.61	0.67
	Proportional Reasoning and Relationships (Cat2)									1.00	0.60	0.70
	Geometric Reasoning (Cat3)										1.00	0.58
	Data Analysis and Probability (Cat4)											1.00

Table 66: Grade 8 All Subjects Observed Score Correlations

Subject	Reporting Category	Science				ELA Reading			Mathematics			
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 4
Science	Nature of Science (Cat1)	1.00	0.65	0.65	0.64	0.55	0.59	0.62	0.32	0.32	0.37	0.30
	Earth and Space Sciences (Cat2)		1.00	0.68	0.68	0.55	0.58	0.62	0.35	0.35	0.41	0.33
	Physical Sciences (Cat3)			1.00	0.67	0.55	0.58	0.62	0.36	0.36	0.41	0.33
	Life Sciences (Cat4)				1.00	0.55	0.59	0.63	0.34	0.34	0.39	0.31
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.65	0.71	0.32	0.32	0.36	0.27
	Reading Informational Text (Cat2)						1.00	0.72	0.34	0.34	0.38	0.29

	Reading Across Genres and Vocabulary (Cat3)							1.00	0.36	0.37	0.42	0.32
Mathematics	Number Sense and Operations and Probability (Cat1)								1.00	0.46	0.44	0.37
	Algebraic Reasoning (Cat2)									1.00	0.47	0.38
	Linear Relationships, Data Analysis, and Functions (Cat3)										1.00	0.40
	Geometric Reasoning (Cat4)											1.00

Table 67: Grade 8 All Subjects Disattenuated Score Correlations

Subject	Reporting Category	Science				ELA Reading			Mathematics			
		Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 4
Science	Nature of Science (Cat1)	1.00	0.83	0.82	0.82	0.70	0.74	0.77	0.40	0.42	0.51	0.40
	Earth and Space Sciences (Cat2)		1.00	0.85	0.85	0.69	0.72	0.76	0.43	0.45	0.55	0.43
	Physical Sciences (Cat3)			1.00	0.84	0.69	0.72	0.75	0.44	0.46	0.55	0.44
	Life Sciences (Cat4)				1.00	0.70	0.73	0.77	0.42	0.43	0.53	0.42
ELA Reading	Reading Prose and Poetry (Cat1)					1.00	0.81	0.87	0.39	0.41	0.49	0.36
	Reading Informational Text (Cat2)						1.00	0.87	0.41	0.42	0.50	0.38
	Reading Across Genres and Vocabulary (Cat3)							1.00	0.44	0.46	0.55	0.41
Mathematics	Number Sense and Operations and Probability (Cat1)								1.00	0.58	0.59	0.48
	Algebraic Reasoning (Cat2)									1.00	0.65	0.51
	Linear Relationships, Data Analysis, and Functions (Cat3)										1.00	0.57
	Geometric Reasoning (Cat4)											1.00

Table 68: Grade 9 Observed Score Correlations

Subject	Reporting Category	U.S. History			ELA Reading			Algebra 1			Geometry			Biology 1		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00	0.70	0.70	0.62	0.63	0.66	0.42	0.38	0.34	0.41	0.45	0.41	0.47	0.47	0.53
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)		1.00	0.70	0.61	0.59	0.64	0.41	0.37	0.32	0.39	0.43	0.39	0.46	0.46	0.51
	The United States and the Defense of the International Peace, 1940–Present (Cat3)			1.00	0.63	0.63	0.69	0.41	0.38	0.32	0.40	0.43	0.40	0.47	0.46	0.53
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.65	0.70	0.41	0.38	0.31	0.37	0.40	0.38	0.52	0.53	0.55
	Reading Informational Text (Cat2)					1.00	0.72	0.43	0.41	0.33	0.41	0.44	0.42	0.56	0.56	0.59
	Reading Across Genres and Vocabulary (Cat3)						1.00	0.46	0.43	0.34	0.43	0.47	0.44	0.59	0.59	0.63
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)							1.00	0.78	0.73	0.57	0.59	0.58	0.53	0.51	0.55
	Linear Relationships (Cat2)								1.00	0.72	0.55	0.56	0.55	0.51	0.50	0.53
	Non-Linear Relationships (Cat3)									1.00	0.49	0.48	0.48	0.45	0.43	0.45
Geometry	Logic, Relationships, and Theorems (Cat1)										1.00	0.81	0.81	0.60	0.58	0.61
	Congruence, Similarity, and Constructions (Cat2)											1.00	0.81	0.61	0.59	0.63
	Measurement and Coordinate Geometry (Cat3)												1.00	0.60	0.57	0.60
Biology 1	Molecular and Cellular Biology (Cat1)													1.00	0.66	0.71
	Classification, Heredity, and Evolution (Cat2)														1.00	0.70
	Organisms, Populations, and Ecosystems (Cat3)															1.00

Table 69: Grade 9 Disattenuated Score Correlations

Subject	Reporting Category	U.S. History			ELA Reading			Algebra 1			Geometry			Biology 1		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00	0.84	0.84	0.76	0.77	0.79	0.49	0.45	0.40	0.48	0.52	0.48	0.57	0.58	0.63
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)		1.00	0.85	0.76	0.73	0.76	0.48	0.44	0.38	0.46	0.50	0.47	0.56	0.57	0.61
	The United States and the Defense of the International Peace, 1940–Present (Cat3)			1.00	0.78	0.77	0.82	0.48	0.45	0.38	0.47	0.51	0.47	0.57	0.57	0.63
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.81	0.85	0.49	0.47	0.38	0.44	0.48	0.46	0.65	0.67	0.68
	Reading Informational Text (Cat2)					1.00	0.86	0.51	0.49	0.39	0.48	0.51	0.49	0.68	0.69	0.71
	Reading Across Genres and Vocabulary (Cat3)						1.00	0.54	0.51	0.40	0.50	0.54	0.51	0.71	0.72	0.74
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)							1.00	0.91	0.86	0.65	0.67	0.67	0.63	0.62	0.64
	Linear Relationships (Cat2)								1.00	0.85	0.63	0.65	0.63	0.61	0.60	0.62
	Non-Linear Relationships (Cat3)									1.00	0.56	0.56	0.56	0.54	0.52	0.54
Geometry	Logic, Relationships, and Theorems (Cat1)										1.00	0.92	0.92	0.70	0.69	0.70
	Congruence, Similarity, and Constructions (Cat2)											1.00	0.92	0.72	0.70	0.73
	Measurement and Coordinate Geometry (Cat3)												1.00	0.70	0.69	0.71
Biology 1	Molecular and Cellular Biology (Cat1)													1.00	0.82	0.85
	Classification, Heredity, and Evolution (Cat2)														1.00	0.85
	Organisms, Populations, and Ecosystems (Cat3)															1.00

Table 70: Grade 10 Observed Score Correlations

Subject	Reporting Category	U.S. History			ELA Reading			Algebra 1			Geometry			Biology 1		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00	0.70	0.70	0.54	0.57	0.61	0.42	0.38	0.34	0.41	0.45	0.41	0.47	0.47	0.53
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)		1.00	0.70	0.52	0.55	0.60	0.41	0.37	0.32	0.39	0.43	0.39	0.46	0.46	0.51
	The United States and the Defense of the International Peace, 1940–Present (Cat3)			1.00	0.52	0.56	0.60	0.41	0.38	0.32	0.40	0.43	0.40	0.47	0.46	0.53
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.63	0.69	0.32	0.29	0.21	0.40	0.42	0.38	0.43	0.45	0.50
	Reading Informational Text (Cat2)					1.00	0.71	0.36	0.33	0.25	0.43	0.46	0.41	0.46	0.48	0.53
	Reading Across Genres and Vocabulary (Cat3)						1.00	0.40	0.36	0.27	0.46	0.49	0.45	0.51	0.52	0.58
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)							1.00	0.78	0.73	0.57	0.59	0.58	0.53	0.51	0.55
	Linear Relationships (Cat2)								1.00	0.72	0.55	0.56	0.55	0.51	0.50	0.53
	Non-Linear Relationships (Cat3)									1.00	0.49	0.48	0.48	0.45	0.43	0.45
Geometry	Logic, Relationships, and Theorems (Cat1)										1.00	0.81	0.81	0.60	0.58	0.61
	Congruence, Similarity, and Constructions (Cat2)											1.00	0.81	0.61	0.59	0.63
	Measurement and Coordinate Geometry (Cat3)												1.00	0.60	0.57	0.60
Biology 1	Molecular and Cellular Biology (Cat1)													1.00	0.66	0.71
	Classification, Heredity, and Evolution (Cat2)														1.00	0.70
	Organisms, Populations, and Ecosystems (Cat3)															1.00

Table 71: Grade 10 Disattenuated Score Correlations

Subject	Reporting Category	U.S. History			ELA Reading			Algebra 1			Geometry			Biology 1		
		Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3	Rep 1	Rep 2	Rep 3
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 (Cat1)	1.00	0.84	0.84	0.67	0.69	0.73	0.49	0.45	0.40	0.48	0.52	0.48	0.57	0.58	0.63
	Global Military, Political, and Economic Challenges, 1890–1940 (Cat2)		1.00	0.85	0.65	0.68	0.71	0.48	0.44	0.38	0.46	0.50	0.47	0.56	0.57	0.61
	The United States and the Defense of the International Peace, 1940–Present (Cat3)			1.00	0.65	0.69	0.72	0.48	0.45	0.38	0.47	0.51	0.47	0.57	0.57	0.63
ELA Reading	Reading Prose and Poetry (Cat1)				1.00	0.80	0.84	0.39	0.36	0.26	0.47	0.51	0.46	0.54	0.56	0.61
	Reading Informational Text (Cat2)					1.00	0.85	0.43	0.40	0.30	0.51	0.54	0.49	0.57	0.60	0.65
	Reading Across Genres and Vocabulary (Cat3)						1.00	0.47	0.42	0.32	0.53	0.57	0.52	0.60	0.63	0.68
Algebra 1	Expressions, Functions, and Data Analysis (Cat1)							1.00	0.91	0.86	0.65	0.67	0.67	0.63	0.62	0.64
	Linear Relationships (Cat2)								1.00	0.85	0.63	0.65	0.63	0.61	0.60	0.62
	Non-Linear Relationships (Cat3)									1.00	0.56	0.56	0.56	0.54	0.52	0.54
Geometry	Logic, Relationships, and Theorems (Cat1)										1.00	0.92	0.92	0.70	0.69	0.70
	Congruence, Similarity, and Constructions (Cat2)											1.00	0.92	0.72	0.70	0.73
	Measurement and Coordinate Geometry (Cat3)												1.00	0.70	0.69	0.71
Biology 1	Molecular and Cellular Biology (Cat1)													1.00	0.82	0.85
	Classification, Heredity, and Evolution (Cat2)														1.00	0.85
	Organisms, Populations, and Ecosystems (Cat3)															1.00

Summative and Interim Correlations

Test takers who took PM1 and PM3 and those who took PM2 and PM3 were identified for conducting the cross-test set of correlations. Table 72 to Table 75 present the correlations between summative and interim assessments for ELA and mathematics. Observed correlations range from 0.60–0.87. Disattenuated correlations range from 0.88–0.98. The number (N) of students, mean, and standard deviation of scale score, and reliability coefficient reported in tables are based on students who took both the summative and interim assessments.

Table 72: Correlations, ELA, PM1 vs. PM3

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	PM1	185.15	21.74	0.68	0.70	0.93	205,425
	PM3	201.92	21.88	0.85			
4	PM1	200.19	20.10	0.80	0.77	0.94	200,461
	PM3	212.72	22.06	0.83			
5	PM1	210.59	20.37	0.86	0.80	0.92	193,997
	PM3	222.74	20.99	0.88			
6	PM1	218.84	21.59	0.83	0.79	0.95	193,699
	PM3	225.20	22.95	0.84			
7	PM1	221.43	23.68	0.81	0.78	0.95	202,912
	PM3	229.65	24.35	0.85			
8	PM1	226.96	23.54	0.80	0.77	0.92	197,825
	PM3	236.11	24.38	0.87			
9	PM1	232.59	23.74	0.81	0.77	0.92	200,906
	PM3	241.31	23.40	0.86			
10	PM1	237.42	24.09	0.82	0.75	0.88	199,693
	PM3	246.54	23.13	0.87			

Table 73: Correlations, Mathematics, PM1 vs. PM3

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	PM1	174.50	18.27	0.81	0.80	0.93	204,558
	PM3	202.49	21.13	0.92			
4	PM1	189.33	18.27	0.77	0.81	0.96	194,519
	PM3	214.56	20.97	0.91			
5	PM1	201.79	19.32	0.81	0.81	0.94	187,439
	PM3	224.27	21.95	0.91			

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
6	PM1	213.42	17.33	0.86	0.81	0.92	182,990
	PM3	231.00	20.84	0.91			
7	PM1	218.46	19.14	0.74	0.71	0.93	130,392
	PM3	231.17	22.35	0.79			
8	PM1	217.74	18.87	0.62	0.60	0.90	98,034
	PM3	235.22	22.68	0.71			

Table 74: Correlations, ELA, PM2 vs. PM3

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	PM2	193.36	22.34	0.81	0.78	0.95	209,711
	PM3	201.54	22.11	0.85			
4	PM2	205.54	22.57	0.78	0.79	0.98	206,148
	PM3	212.17	22.34	0.83			
5	PM2	216.33	21.58	0.85	0.83	0.96	197,780
	PM3	222.35	21.28	0.88			
6	PM2	220.47	23.15	0.83	0.81	0.97	197,511
	PM3	224.77	23.23	0.84			
7	PM2	224.23	24.77	0.81	0.81	0.98	206,580
	PM3	229.28	24.57	0.84			
8	PM2	230.20	24.66	0.84	0.80	0.94	201,276
	PM3	235.73	24.61	0.87			
9	PM2	235.28	24.46	0.84	0.80	0.94	205,272
	PM3	240.88	23.66	0.86			
10	PM2	239.20	24.68	0.86	0.78	0.90	203,464
	PM3	246.19	23.33	0.87			

Table 75: Correlations, Mathematics, PM2 vs. PM3

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	PM2	187.85	18.68	0.88	0.87	0.96	208,795
	PM3	202.16	21.31	0.92			
4	PM2	198.96	18.33	0.86	0.86	0.97	200,307
	PM3	214.13	21.18	0.92			

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
5	PM2	211.91	19.45	0.89	0.87	0.97	191,170
	PM3	223.93	22.13	0.91			
6	PM2	221.45	18.16	0.92	0.88	0.96	186,976
	PM3	230.68	21.01	0.91			
7	PM2	223.39	19.59	0.77	0.76	0.98	135,596
	PM3	231.05	22.46	0.79			
8	PM2	227.92	18.12	0.75	0.68	0.93	104,861
	PM3	235.58	22.75	0.72			

Discussion

The empirical results together from the Q3, confirmatory factor analysis, correlation analysis, and measurement invariance testing across subgroups suggest the implied model fits the data. 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.

Furthermore, the correlations among the separate reporting categories 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 reporting categories. The high correlations among the reporting categories suggest these alternative methods are unnecessary and that our current approach is in fact preferable.

Lastly, before items can be entered into the item bank, model fit is evaluated based on Q1 fit statistics, visual inspection of fit plots, and scrutiny of the item’s content by content experts for any items that display less than ideal fit. This ensures items are aligned to Florida’s standards and reporting categories.

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 for Educational and Psychological Testing* (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. 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. For science and social studies, a third-party, independent alignment study was conducted in February 2016. This report can be found in Volume 4, Evidence of Reliability and Validity, Appendix D, FSA Alignment Report, of the *Florida Standards*

Assessments 2015–2016 Technical Report. A new third-party, independent alignment study for the new B.E.S.T. Standards for ELA and mathematics is planned for 2025. To determine content representativeness, diverse panels of content experts will review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct. For details see this volume's Appendix E, ELA and Mathematics Alignment Study Proposal.

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. Examination of the correlational relationship among subscores is also used to evaluate content relevance. Results for this for the statewide assessments were 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 (TEIs) 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 TEIs developed by CAI, and the items are delivered by the same engine as is used for delivery of the Smarter Balanced Assessment. Hence, Florida's statewide assessments make 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 in Florida (see Volume 7 of the *Florida Standards Assessments 2014–2015 Technical Report*; Florida Department of Education, 2015).

In addition, Florida FAST, B.E.S.T., and science cognitive labs were conducted to examine the response processes of test takers for grades 3, 7, and 10 ELA, grades 3 and 7 mathematics, Algebra 1, grades 5 and 8 science, and Biology 1. These grades/courses were selected because they represent the item types, share similar blueprints (including the same content categories), and have the same test development procedures as the non-selected grades/courses. The assessments are all based on the same content standards and benchmarks, along with extensive content limits that define what is to be assessed. The studies were delayed due to the COVID-19 pandemic and school closings in 2020–2021. They were finally completed in 2024. In comparison to the intended cognitive complexity, it was found that the enacted cognitive complexity either met or exceeded the intended cognitive complexity in 58%–88% of the items. Evidence of linguistic complexity that was construct irrelevant was not found; however, students had significant difficultly reading algebra equations accurately, suggesting a focal point teachers should consider targeting during instruction. Study findings generalized across sampled grades. This study provides response process validity evidence that assessment items measure the intended cognitive processes represented in the State's academic content standards. The full findings can be found in the cognitive laboratory report in Appendix L.

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

4.3.1 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 following sections.

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 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

[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).

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 of 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 comprised 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 Florida's statewide assessments were representative of the content standards of the larger knowledge domain. We describe the content standards for Florida's assessments and discuss the test development process, mapping the assessments 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

Florida's statewide assessments are aligned to the Florida standards. Beginning with the 2022–2023 school year, Florida's statewide, standardized assessments in English language arts (ELA) reading, writing, and mathematics were aligned with the Benchmarks for Excellent Student Thinking (B.E.S.T.). Assessments for Science and Social Studies remain aligned to Florida's state academic standards that were adopted starting in 2008.

Table 76 to Table 79 present the reporting categories by grade and test, as well as the number of items measuring each category. For ELA, science, and social studies accommodated forms, 100% of these items are also available for form building. For mathematics, a small number within some reporting categories (less than 5%) are not able to be converted to some accommodated form formats such as paper.

Table 76: Number of Items for Each Mathematics Reporting Category

Grade*	Reporting Category	Number of Items
3	Number Sense and Additive Reasoning	116
	Number Sense and Multiplicative Reasoning	118
	Fractional Reasoning	69
	Geometric Reasoning, Measurement, Data Analysis, and Probability	141
4	Number Sense and Operations with Whole Numbers	128
	Number Sense and Operations with Fractions and Decimals	106
	Geometric Reasoning, Measurement, Data Analysis, and Probability	96
5	Number Sense and Operations with Whole Numbers	92
	Number Sense and Operations with Fractions and Decimals	128

Grade*	Reporting Category	Number of Items
	Algebraic Reasoning	88
	Geometric Reasoning, Measurement, Data Analysis, and Probability	142
6	Number Sense and Operations	149
	Algebraic Reasoning	155
	Geometric Reasoning, Data Analysis, and Probability	146
7	Number Sense and Operations and Algebraic Reasoning	94
	Proportional Reasoning and Relationships	77
	Geometric Reasoning	88
	Data Analysis and Probability	108
8	Number Sense and Operations and Probability	87
	Algebraic Reasoning	60
	Linear Relationships, Data Analysis, and Functions	74
	Geometric Reasoning	66

Table 77: Number of Items for Each ELA Reporting Category

Reporting Category	Grade*							
	3	4	5	6	7	8	9	10
Reading Prose and Poetry	100	141	120	88	90	79	92	100
Reading Informational Text	100	91	121	103	105	100	132	113
Reading Across Genres and Vocabulary	156	200	201	166	179	152	165	178

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

Table 78: Number of Items for Each EOC Reporting Category

Course	Reporting Category	Number of Items
Algebra 1	Expressions, Functions, and Data Analysis	99
	Linear Relationships	108
	Non-Linear Relationships	111
Geometry	Logic, Relationships, and Theorems	118
	Congruence, Similarity, and Constructions	114
	Measurement and Coordinate Geometry	133

Table 79: Number of Items by Reporting Category—Science and Social Studies

Grade	Reporting Category	Number of Items
Biology 1	Molecular and Cellular Biology	362
	Classification, Heredity, and Evolution	246
	Organisms, Populations, and Ecosystems	413
Civics	Origins and Purposes of Law and Government	250
	Roles, Rights, and Responsibilities of Citizens	195
	Government Policies and Political Processes	178
	Organization and Function of Government	243
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910	261
	Global Military, Political, and Economic Challenges, 1890–1940	316
	The United States and the Defense of the International Peace, 1940–Present	287
Grade 5 Science	Nature of Science	329
	Earth and Space Sciences	301
	Physical Sciences	194
	Life Sciences	315
Grade 8 Science	Nature of Science	300
	Earth and Space Sciences	256
	Physical Sciences	163
	Life Sciences	232

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 FAST and B.E.S.T. were comprised of test items that included traditional multiple-choice items, items that required students to type or write a response, and TEIs. TEIs are computer-delivered items that require students to interact with test content to select, construct, and support their answers. Science and social studies were comprised of multiple-choice items only. 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 passages in ELA and science. The minimum and maximum number of items by grade and subject and other details on the blueprint are presented in Appendices B1 to B5 of Volume 2.

Test Construction and CAT Algorithm

Test construction in Florida switched from building fixed-form tests to configuring the computer-adaptive test (CAT) system for the regular summative assessments in the 2022–2023 school year for ELA and mathematics, and in 2023–2024 for science and social studies. The accommodated tests remain fixed form. Details are provided in Volume 1, Section 3 Adaptive Testing Advantages, Algorithm and Simulation Studies Overview, and Volume 2, Section 4 Test Construction. The algorithm prioritizes blueprint match, followed by adapting to student ability and any other customizable item administration considerations and constraints deemed important for a particular test.

Before the testing window opens, the CAT configurations are evaluated to ensure that the forms every student receives will conform to the required test-specific specifications, using simulations. Simulation results are evaluated based on numerous checks. Typically, all forms generated by the simulations should (for operational and field-test items)

- match test blueprint (including overall minimum and maximum items);
- meet the minimum and maximum number of required passages;
- result in sufficient numbers for item calibration;
- result in satisfactory correlation between test difficulty and student estimated ability; and
- result in uniform item exposure across the bank.

Summary simulation outcome reports are in Volume 2, Appendix F.

Test Development

Florida's item pools grow 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 following factors were considered when embedding field-test items into the operational assessment for the spring administration:

- Ensured that field-test items did not cue or clue answers to other field-test items
- Ensured that 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
- Selected items in the field-test sets that reflected a range of difficulty levels and cognitive levels
- 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

Alignment of Item Bank to the Content Standards and Benchmarks

A third-party, independent alignment study for the new B.E.S.T. standards is planned for completion in 2025. For details, see this volume's Appendix E. The results from the previous alignment study for the Florida Standards Assessments (FSA) standards can be found in Volume 4, Appendix D, of the *2015–2016 Florida Standards Assessments Technical Report*.

The new study will be designed to yield evidence that pertains to fulfilling requirements as stated in federal statute related to the content alignment of statewide assessments with corresponding academic standards. Four main research questions will guide the work.

1. Framework Analysis: To what extent do the CAT algorithms, test blueprints, and other relevant test specifications and documentation reflect structure and design that support the capacity of alignment of test events with corresponding grade-level academic standards?

2. Aggregate Data Review: To what extent do the available aggregate data for test events administered in spring 2023 provide evidence that the algorithm and blueprints are yielding test forms as expected?

3. Validation of Internal Metadata: To what extent is independent coding of assessment targets reasonably consistent with the assessment targets identified within internal (vendor) item metadata?

4. Test Form–Level Alignment: What is the degree of alignment of actual test events, sampled from below satisfactory, on grade level and above satisfactory/mastery with corresponding Florida standards, based on agreed-on criteria and minimum cutoffs?

The study will yield multiple lines of evidence that will support a validity argument that would extend across all test events generated by a computer-adaptive assessment program. Beyond the content alignment evidence for individual test events, it is important to provide additional evidence that can help extend findings across all test events generated by a particular testing program. Because computer-adaptive test form assembly relies on internal metadata to meet blueprint specifications, validation of the internal metadata (based on independent item-level content analysis) allows for greater confidence that an assessment program has the capacity to generate test forms that include content consistent with blueprint intent and, therefore, that test form-level findings can be reasonably generalized across all test forms generated by the assessment program. By drawing on multiple lines of evidence, the overall study design allows for the potential to craft a logical argument for the capacity for alignment of all test events generated by the FAST and EOC assessment programs included in the study with the corresponding Florida B.E.S.T. Standards, as appropriate, based on results.

The resulting logic argument, stated in the positive, would be:

- If relevant test specifications and documentation reflect a structure and design to support the capacity of alignment of test events with corresponding grade-level academic standards;
- and if test events (sampled from below satisfactory, on grade level, and above satisfactory) meet minimum alignment criteria (based on agreed-on cutoffs for Categorical Concurrence, Depth of Knowledge [DOK] Consistency, Range of Knowledge Correspondence, and Balance of Representation);

- and if the test blueprints and algorithm are generating test events as intended (based on data from all administered test events);
- and if validation of internal metadata supports generalizability of alignment findings across all test forms generated by the assessment programs;
- then it is possible to make an argument for the capacity for alignment for all test events resulting from Florida FAST assessments for ELA grades 3–10, FAST assessments for mathematics grades 3–8, and B.E.S.T. EOC assessments for Algebra 1 and Geometry with corresponding Florida B.E.S.T. Standards.

For science and social studies, a third-party, independent alignment study was conducted in 2012 to evaluate the alignment between test items and benchmarks they intend to measure for grades 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 the Florida Statewide Assessments

Standards for Educational and Psychological Testing notes that “some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers” (AERA, APA, & NCME, 2014, 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 and 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.

Cognitive lab studies were conducted to examine the response processes of test takers for grades 3, 7, and 10 ELA, grades 3 and 7 mathematics, Algebra 1, science 5 and 8, and Biology 1. These grades were selected because they represent the item types, share similar blueprints (including the same content categories), and have the same test development procedures as the non-selected grades. The assessments are all based on the same content standards and benchmarks, along with extensive content limits that define what is to be assessed. For all grades, 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 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. Given these commonalities between the selected and non-selected grades, results from cognitive lab studies from the selected grades are generalizable to non-selected grades and non-selected item types. In the studies, students work through sample items. Eight students responded to each item, and their thinking processes were elicited through a combination of concurrent think-aloud (thinking out loud while reading and responding to an item) and focused probes that were tailored based on the anticipated solution path for a given item. The cognitive lab interviews used recorded audio, and the students’ responses to the test items were captured by the Test Delivery System (TDS). Following the cognitive lab, the interviewer reviewed all relevant information and filed a report that included, 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 were 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 they understood (at a general level) what the item was asking them to do, and 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 cognitive analysis followed Ferrara et al. (2003). In comparison to the intended cognitive complexity, it was found that the enacted cognitive complexity either met or exceeded the intended cognitive complexity in 58%–88% of the items. Evidence of linguistic complexity that was construct irrelevant was not found; however, students had significant difficulty reading Algebra equations accurately, suggesting a focal point teachers should consider targeting during instruction. Study findings generalized across sampled grades. This study provided response process validity evidence that assessment items measure the intended cognitive processes represented in the State's academic content standards.

The cognitive lab studies were delayed due to the COVID-19 pandemic and school closings in 2020–2021. The final cognitive laboratory report (including DOK distributions in the bank) can be found in Appendix G.

Evidence of Control of Measurement Error

Reliability and the CSEM are discussed in an earlier chapter of this volume. Tables reporting the CSEM and marginal reliability are also included. As discussed earlier, these measures show that Florida's assessment 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 arguments 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 Florida's statewide assessments are not more or less valid for use with one subpopulation of students relative to another. The 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 by the time PM3 is administered, and all tests are given to all students under standardized conditions.

Great care has been taken so that the items constituting Florida's 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 of this technical report: Appendix A, Operational Item Statistics, for operational items and Appendix B, Field-Test Item Statistics, for field-test items.

In addition, marginal 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 disabilities, and students with/without accommodations (see the reliability in 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.

4.3.2 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.

Analytical Evidence

Florida's standards and statewide assessments create a common foundation to be learned by all students and define the domain of interest. As documented in this report, the 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 ELLs 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 ELLs to demonstrate their mathematical ability on a test that limits the quantity and complexity of English language used in the items. Florida's 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 reliability measures for the ELL, disability, and accommodation groups (see the reliability and classification accuracy in 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's assessments in spring 2024 included grades 3–8 mathematics, grades 3–10 reading, grades 4–10 writing, all EOC assessments (Algebra 1 and Geometry), U.S. History, Civics, Biology 1, and science grades 5 and 8. 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 administrations 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. The 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 tablet (e.g., iPad) devices. 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 to further examine the score comparability between tablet and non-tablet devices.

Olsen (2014) compared the performance of grades 1–12 testing on tablets and computers. 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 ELA/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 were 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 a 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. 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 previously, finding an adequate criterion for a standards-based achievement test can be difficult.

According to *Standards (*AERA, APA, and NCME, 2014), 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 multi-method 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 were provided previously in this volume (see Table 50 to Table 57).

4.3.3 Implication Validity Evidence

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 Florida's 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 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. 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 the assessments, separate volumes are devoted to them. Volume 3 of the *Benchmarks for Excellent Student Thinking 2022–2023 Technical Report* discusses the details concerning performance standards, and Chapter 5: Performance Standards from the *Florida Statewide Science and EOC Assessments 2019 Technical Report* describes the standard for science and social studies. 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's assessment scores provide information reflecting what students know and can do in relation to 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's standards and 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 the assessments and standards. 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 the 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 this technical report narrates the details and cautions of score use.

This volume, as well as other volumes of this technical report, provide validity evidence supporting the appropriate inferences from Florida's Statewide Assessment scores. In general, the validity evidence provides supports to the primary claim that the 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. Validity arguments based on rationale and logic are strongly supported for Florida's 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 that the assessments are properly scored and scores can be generalized to the universe score.

5. EVIDENCE OF COMPARABILITY

Florida's Statewide Assessments are available to be administered in regular computer-adaptive test (CAT) mode as well as with accommodations in fixed-form format (see Volume 5, Section 1.2 Testing Accommodations). 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 or device of administration nor the type of test form.

To improve the accessibility of the statewide assessment, alternate assessments were provided to students whose Individual Educational Plan (IEP) or Section 504 Plan indicated such a need for the PM3 and spring summative assessments. The comparability of scores obtained via alternate means of administration must be established and evaluated. This section outlines the overall test development plans that ensured the comparability of CATs and accommodated tests across different devices.

5.1 MATCH-WITH-TEST BLUEPRINTS FOR BOTH CAT AND ACCOMMODATED TESTS

The accommodated versions of the tests were developed according to the same test specifications used for the CATs, including blueprints and content-level considerations. Specifically, the CAT algorithm was used directly in CAI's simulator to generate candidate forms for use as the spring 2024 accommodated forms in each grade for science and social studies. To create the spring 2024 accommodated forms for ELA and mathematics, CAI generated the forms for each grade in the automated form-building tool, which also uses the same underlying CAT algorithm. Thus, the blueprints for the accommodated forms matched the blueprint for the CAT tests—they were chosen directly from forms generated by the CAT. More information about accommodated form construction can be found in Volume 2, Section 4.4 Accommodation Form Construction.

5.2 COMPARABILITY OF TEST SCORES OVER TIME

The comparability of scores over time is ensured via two methods. First, during test development, both content and statistical requirements are implemented. All test items are aligned to the same standards and test blueprint specifications for each administration. In addition, for the accommodated forms, individual items and candidate forms are evaluated based on their statistics. The statistical criteria are consistent from year to year (an overview is included in Volume 1, Section 5 Item Analyses Overview and Section 6.2.2 Accommodated Forms). Second, in future years, drift analyses of the item response theory (IRT) item parameters will be conducted to ensure item parameters can be compared over time.

5.3 COMPARABILITY OF ONLINE AND ACCOMMODATED TESTS

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

Accommodated forms (in various modes) were offered as a special accommodation for students who qualified according to their IEP or Section 504 Plan. Various devices were used across Florida. In the following sections, 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 accommodated form is constructed, first and foremost, the accommodated version is constructed to the exact same blueprint and content-level specifications as the CAT. Items are drawn from exactly the same item bank. For English language arts (ELA), science, and social studies, 100% of items are available for use on accommodated forms. For mathematics, some technology-enhanced items are not able to be translated to paper versions, however, these items are less than 5% of the bank. From the psychometric point of view, the purpose of providing accommodations is to "increase the validity of inferences about students with special needs by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

Details for the rigorous process of translating items to different formats for accommodated forms can be found in Volume 2, Section 3.4 Item Translation to Braille Format and Section 4.4 Accommodation Form Construction. Details of available testing accommodations, their selection, appropriateness of use, appropriateness of implementation, and auditing are in multiple sections in Volume 5 of this technical report.

5.4 Comparability of Constructs

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.

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.

A device comparability study was conducted to provide evidence of the comparability of the Florida Statewide Assessments (FSA) across the most frequently used platforms. Score comparability across different devices was examined to assess whether student performance on the FSA 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 FSA on different devices. The details of the study can be found in Appendix F of *Florida Statewide Assessments 2021–2022 Technical Report* (Appendix D of this volume).

Although the study was conducted using the FSA (and not specifically the FAST/B.E.S.T. assessments), the results are still generalizable to the new ELA and mathematics assessments for reasons outlined in DePascale, Dadey, & Lyons (2016). That is, questions about score comparability across devices are distinct from other threats to score comparability, such as

- differences in test content;
- differences in the types of items and the format of items used on the assessment; and
- differences in scoring and/or the response that a student is expected to provide.

Instead, questions about score comparability across devices include concerns about differences among students in the manner in which content is presented, the manner in which students interact with the content presented, and the manner in which students respond to the content presented. That is, the issue of addressing device comparability is not assessment specific. Since no device effects were previously found in Florida's device comparability study, although the assessment standards and content have changed, the devices used in Florida and the way they are used have not changed. Thus, the study findings should still hold.

5.5 Comparability of Scores

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. The essence is that the accommodated items that are common with the CAT form use item parameters from the CAT calibrations. Since both CAT and accommodated forms are scored using the same IRT-calibrated item pool, the scores obtained from the accommodated form are comparable to those obtained from the CAT.

As for research on score comparability, a review of the literature by Arthur, Kapoor, and Steedle (2020) found most studies showed comparability between scores from paper and online testing but there were similar numbers of studies showing mode effects favoring paper and online testing. They also included meta-analyses that showed near-zero estimates of mode effects when combining results from numerous studies. Thus, any significant individual results showing differences are very likely due to specific circumstances, such as how forms are constructed, the items used, and how they are administered in a specific context. A corollary of this comparability can be achieved if care is taken to ensure comparability.

This is consistent with findings by DePascale, Dadey, & Lyons (2016) in their literature review. They found that (1) the majority of comparability studies have found their computer-based and paper-based tests to be comparable overall (e.g., Davis, Kong, & McBride, 2015; Davis, Orr, Kong, & Lin, 2015) and (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 and device features (e.g., screen size, input mechanism, keyboard). However, 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. The device comparability study mentioned in Section 5.4 is evidence that the State has been successful in avoiding threats to comparability due to devices. Furthermore, as described in Section 5.3 Comparability of Online and Accommodated Tests, numerous processes have been implemented in the design, development, and administration of Florida assessments that mirror best practices recommended by research to maximize comparability.

Empirical evidence is available in the observed data collected from the test administrations—test forms are reliable and students using the accommodated form also have a range of scores. This evidence indicates that high-performing students administered accommodated forms can still demonstrate high performance and are not impeded in any way by the nature of the form or its administration. An overall scale score summary (including mean score, standard deviation, mean conditional standard error of measurement, and marginal reliability) was presented in Table 2 and Table 3 in Section 3.1 (comparison with CATs can be found in Table 80 and Table 81), and by reporting category is presented for online and accommodated groups in Appendix A of this volume. Appendix H with correlations for accommodated scores shows a similar pattern to the CAT.

The marginal reliabilities for accommodated forms are generally lower. However, the sample size for accommodated forms is extremely small and the test-taking subgroup is restricted in terms of their ability distribution, which would contribute to the observed differences in reliabilities calculated from the sample data. In other words, these reliabilities are not estimated based on given theta values coming from the theoretical test information theta distribution. This mismatch between student abilities and the item difficulty distributions in the bank can be seen in Appendix F, especially text-to-speech (TTS) forms where the mismatch with the constructed form is very pronounced (even though the TTS sample size is larger than DEI accommodated forms), contributing to much lower marginal reliabilities.

Table 80: Marginal Reliability Coefficients for Accommodated vs. Regular Online Students

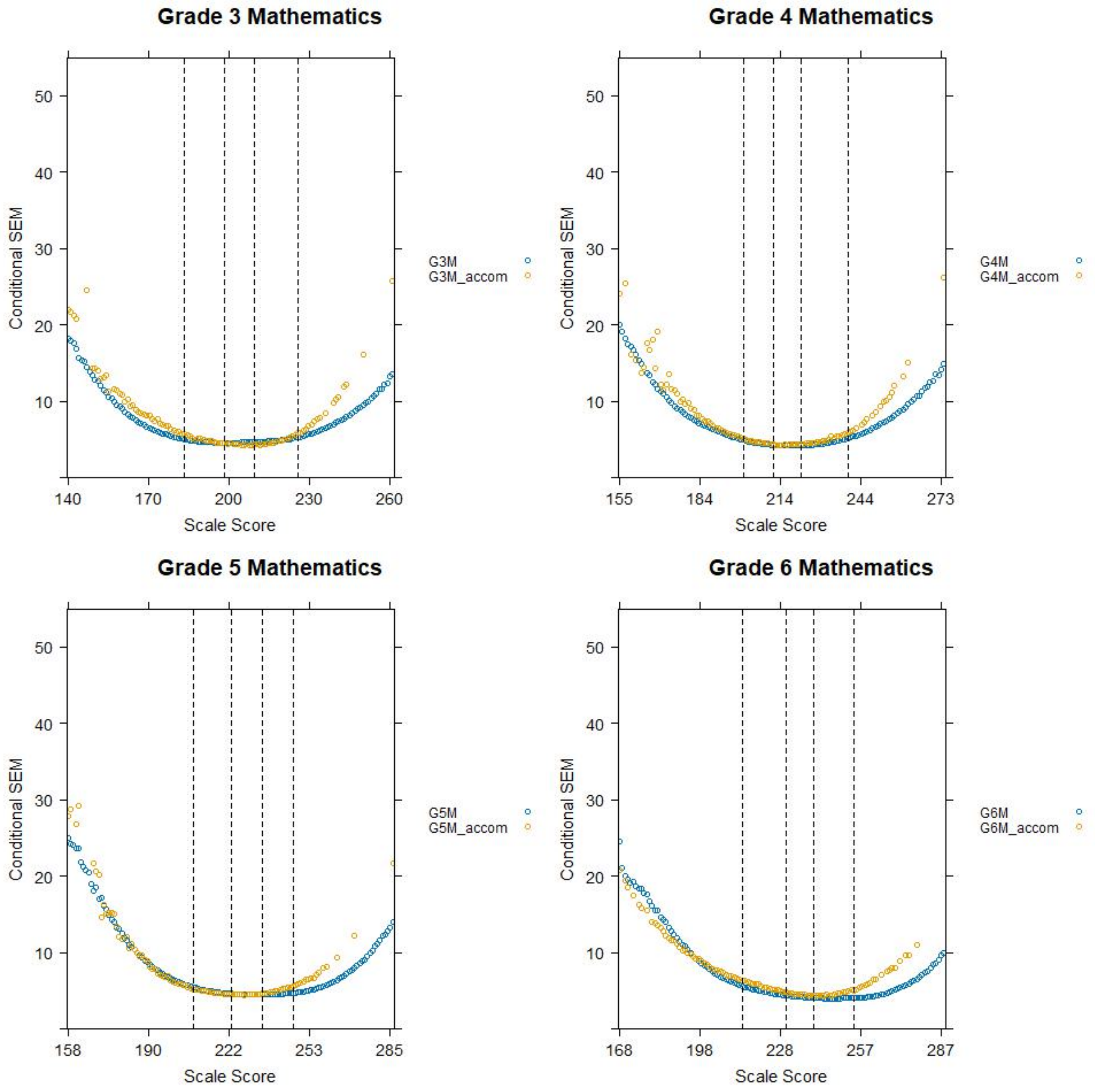
		Regular		Accommodated	
Subject	**Grade**	**N-Count**	**Reliability**	**N-Count**	**Reliability**
ELA Reading	3	215,574	0.85	895	0.77
	4	212,165	0.83	958	0.73
	5	203,412	0.88	800	0.76
	6	205,054	0.84	573	0.74
	7	214,938	0.84	330	0.75
	8	209,835	0.86	329	0.81
	9	216,621	0.86	385	0.80
	10	215,657	0.87	397	0.85
Mathematics	3	214,927	0.92	895	0.87
	4	207,096	0.91	940	0.86
	5	197,191	0.91	805	0.84
	6	194,855	0.91	564	0.87
	7	144,768	0.79	284	0.71
	8	114,710	0.72	243	0.63
Algebra		228,344	0.89	407	0.65
Geometry		213,902	0.91	365	0.65

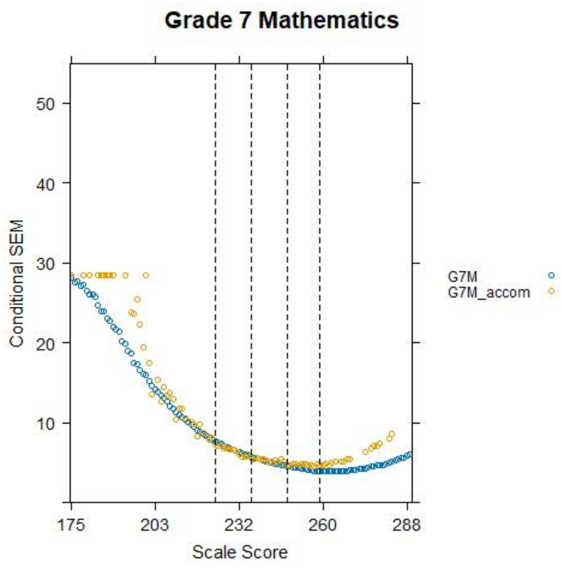
Table 81: Marginal Reliability Coefficients for Accommodated vs. Regular Online Students

	Regular		DEI		TTS	
Subject	**N-Count**	**Reliability**	**N-Count**	**Reliability**	**N-Count**	**Reliability**
Biology 1	199,788	0.85	335	0.81	15,494	0.75
Civics	188,377	0.85	316	0.83	26,176	0.75
U.S. History	183,226	0.85	361	0.84	9,359	0.74
Grade 5 Science	174,486	0.89	789	0.87	28,471	0.85
Grade 8 Science	178,331	0.85	305	0.83	23,173	0.73

Figure 7 to Figure 11 show comparison of mean conditional standard errors of measurement (CSEMs) for the accommodated tests with CAT forms (CSEM curves are the mean CSEM curves for all students). Mean CSEM means for each scale score, we take the average of all the CSEMs conditional on the scale score being equal to that. In general, the accommodated forms are very comparable to a typical CAT form with regards to the standard errors, except for at the lower and upper tails of the distribution, where the CAT forms may be superior in their capability to match student abilities (an established advantage of CAT over fixed-form assessments). This is very pronounced in EOC mathematics across most of the ability range. This is due to the accommodated form's mismatch with student ability (seen in Appendix F), as the form (based on FDOE policy) aims to maximize the information at the level 3 cut, while the student population for accommodated forms is clustered well below that.

Figure 7: Conditional Standard Errors of Measurement (Mathematics)




Grade 7 Mathematics
Conditional SEM
50
40
30
20
10
175
203
232
260
288
Scale Score
G7M
G7M_accom

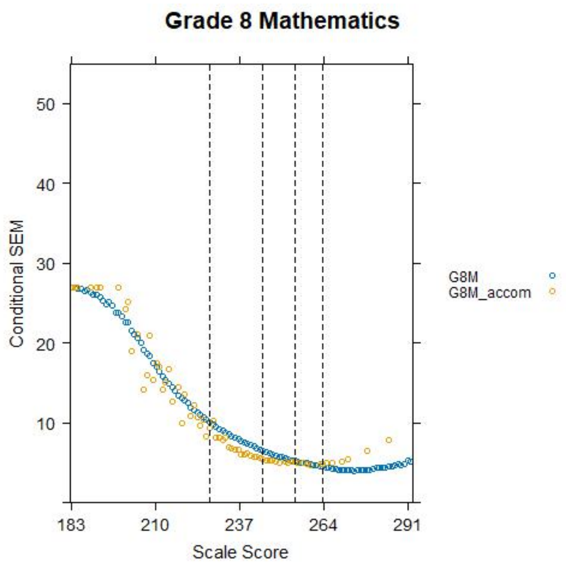
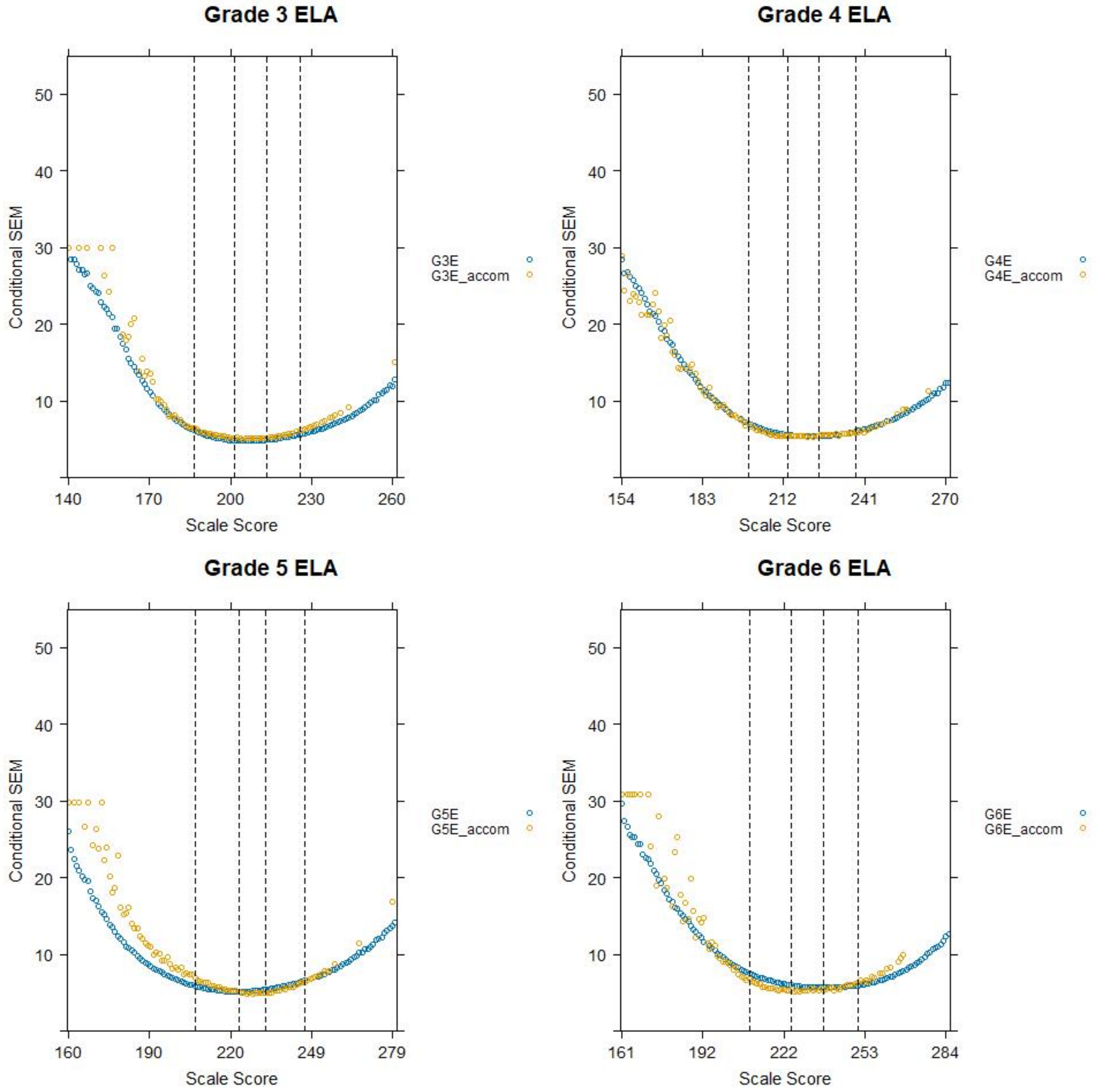
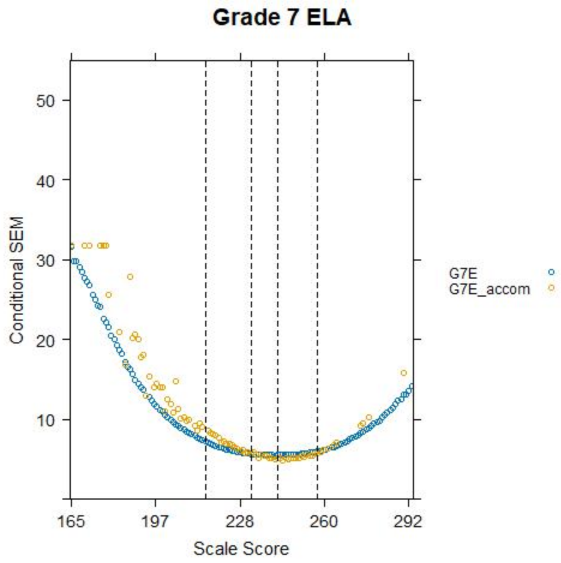
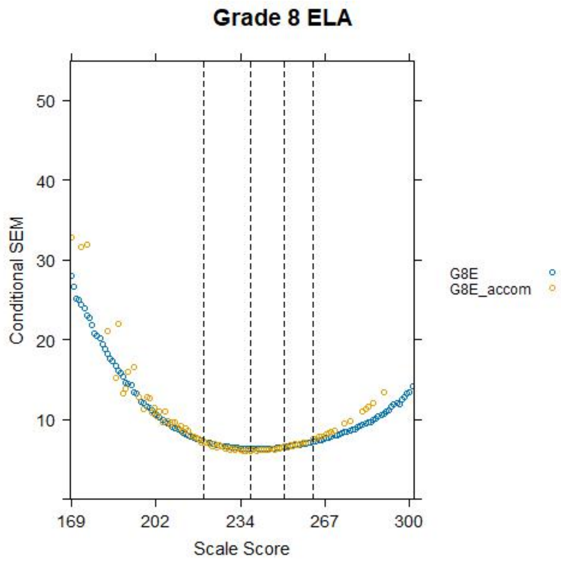
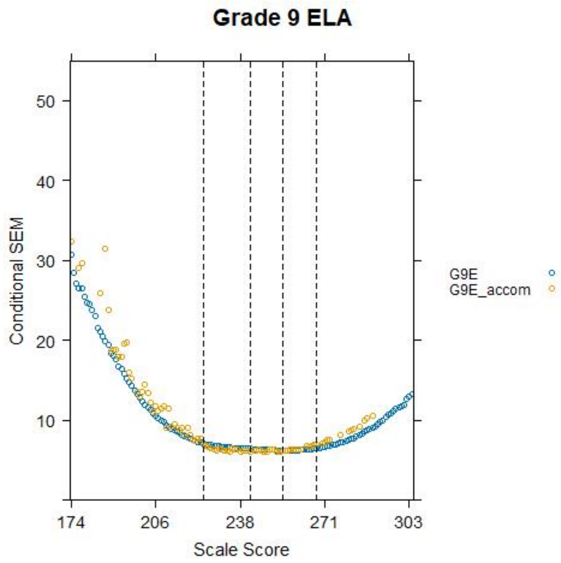

Grade 8 Mathematics
Conditional SEM
50
40
30
20
10
183
210
237
264
291
Scale Score
G8M
G8M_accom

Figure 8: Conditional Standard Errors of Measurement (ELA)




Grade 7 ELA
Conditional SEM
50
40
30
20
10
165
197
228
260
292
Scale Score
G7E
G7E_accom


Grade 8 ELA
Conditional SEM
50
40
30
20
10
169
202
234
267
300
Scale Score
G8E
G8E_accom


Grade 9 ELA
Conditional SEM
50
40
30
20
10
174
206
238
271
303
Scale Score
G9E
G9E_accom

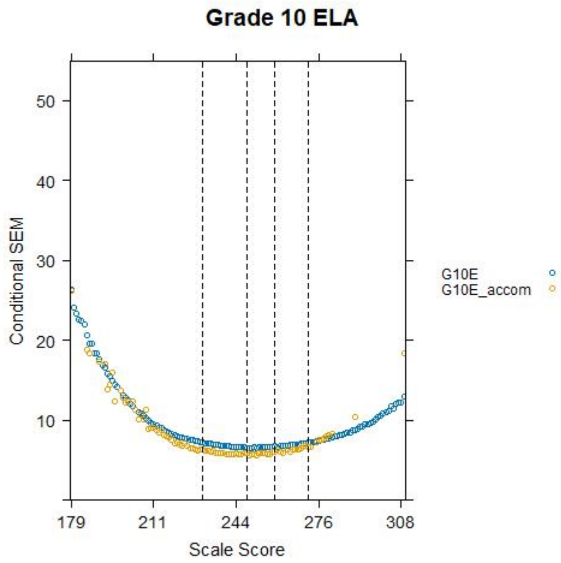

Grade 10 ELA
Conditional SEM
50
40
30
20
10
179
211
244
276
308
Scale Score
G10E
G10E_accom

Figure 9: Conditional Standard Errors of Measurement (EOC)

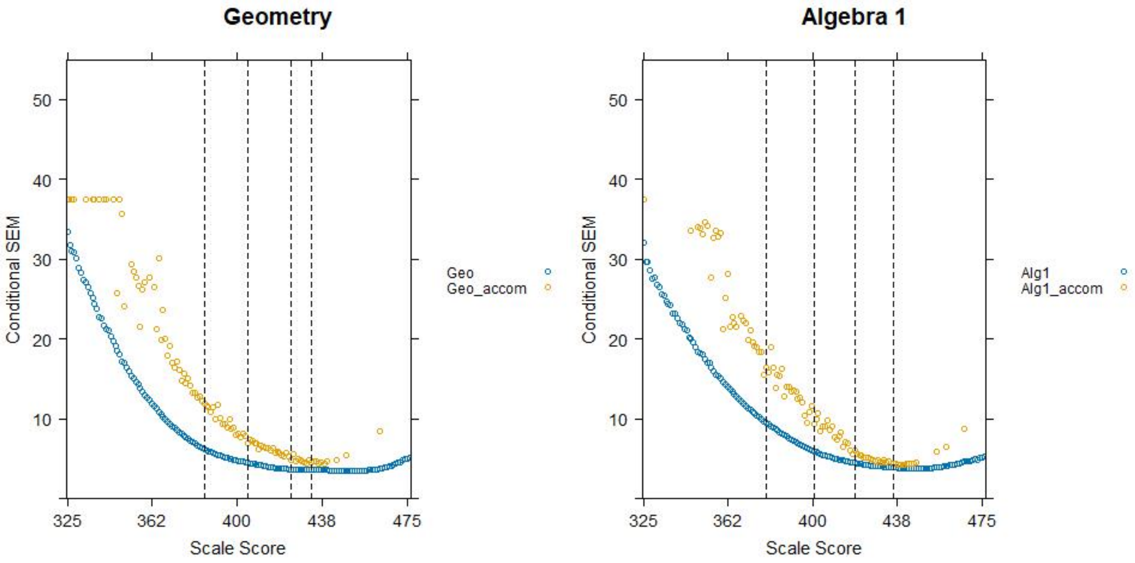


Figure 10: Conditional Standard Errors of Measurement (Science and Social Studies DEI)

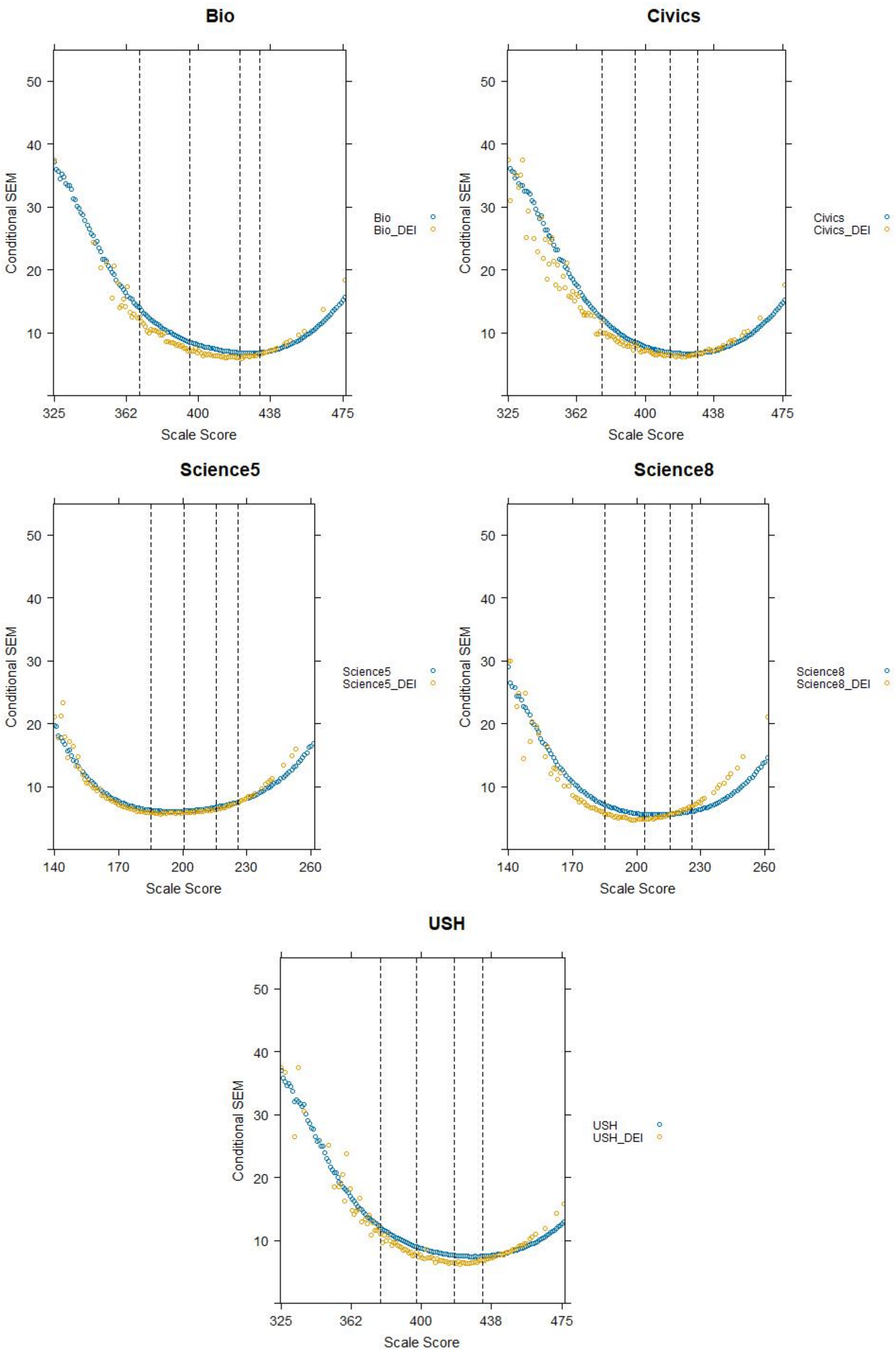


Figure 11: Conditional Standard Errors of Measurement (Science and Social Studies TTS)

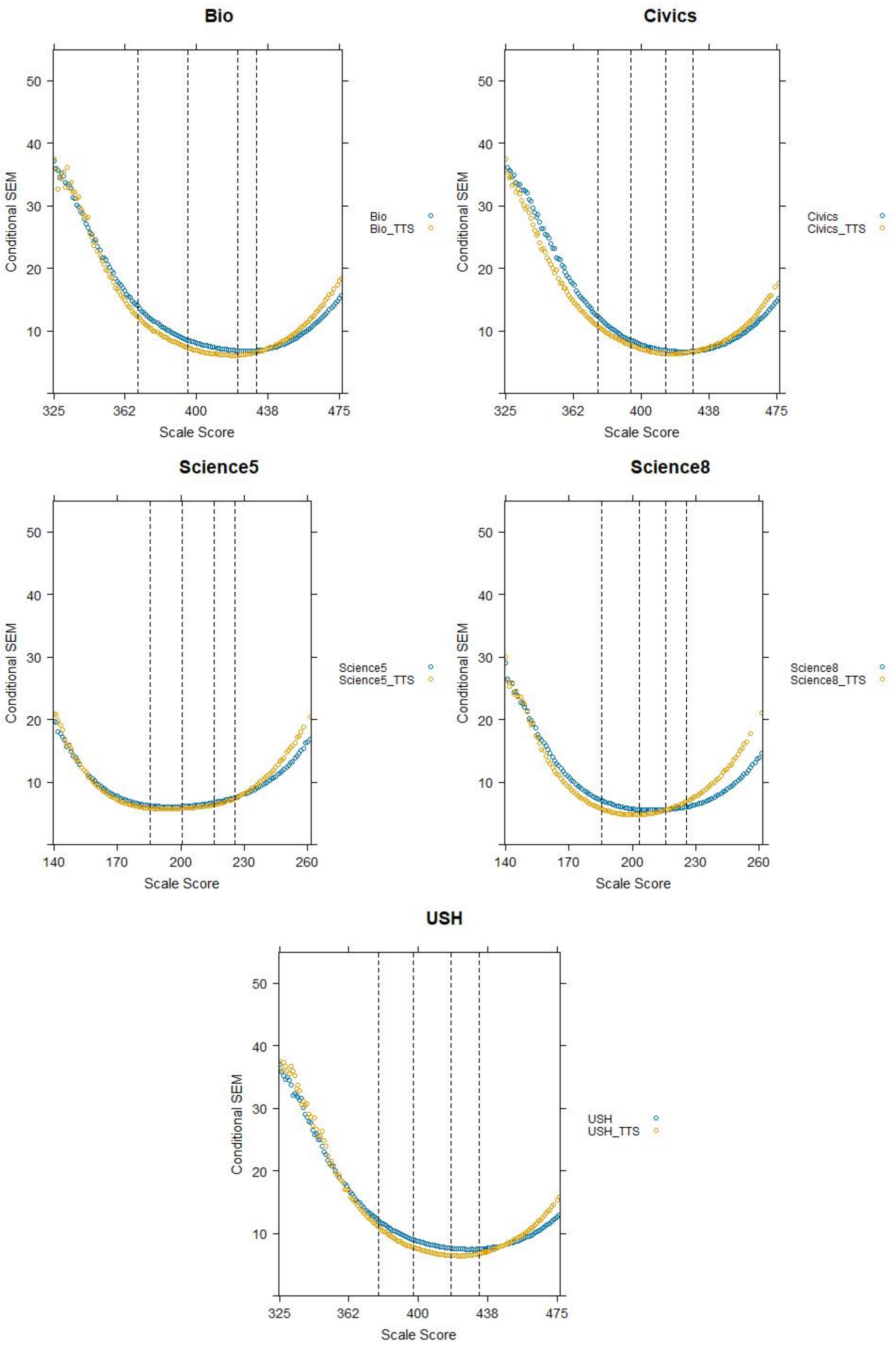


Figure 12 to Figure 15 show comparisons of test characteristic curves (TCCs) for an accommodated form against a typical form (chosen at random from those administered to students scoring at the on-grade cut). There is generally a good match.

Figure 12: Test Characteristic Curves (TCCs) Compared (Mathematics)

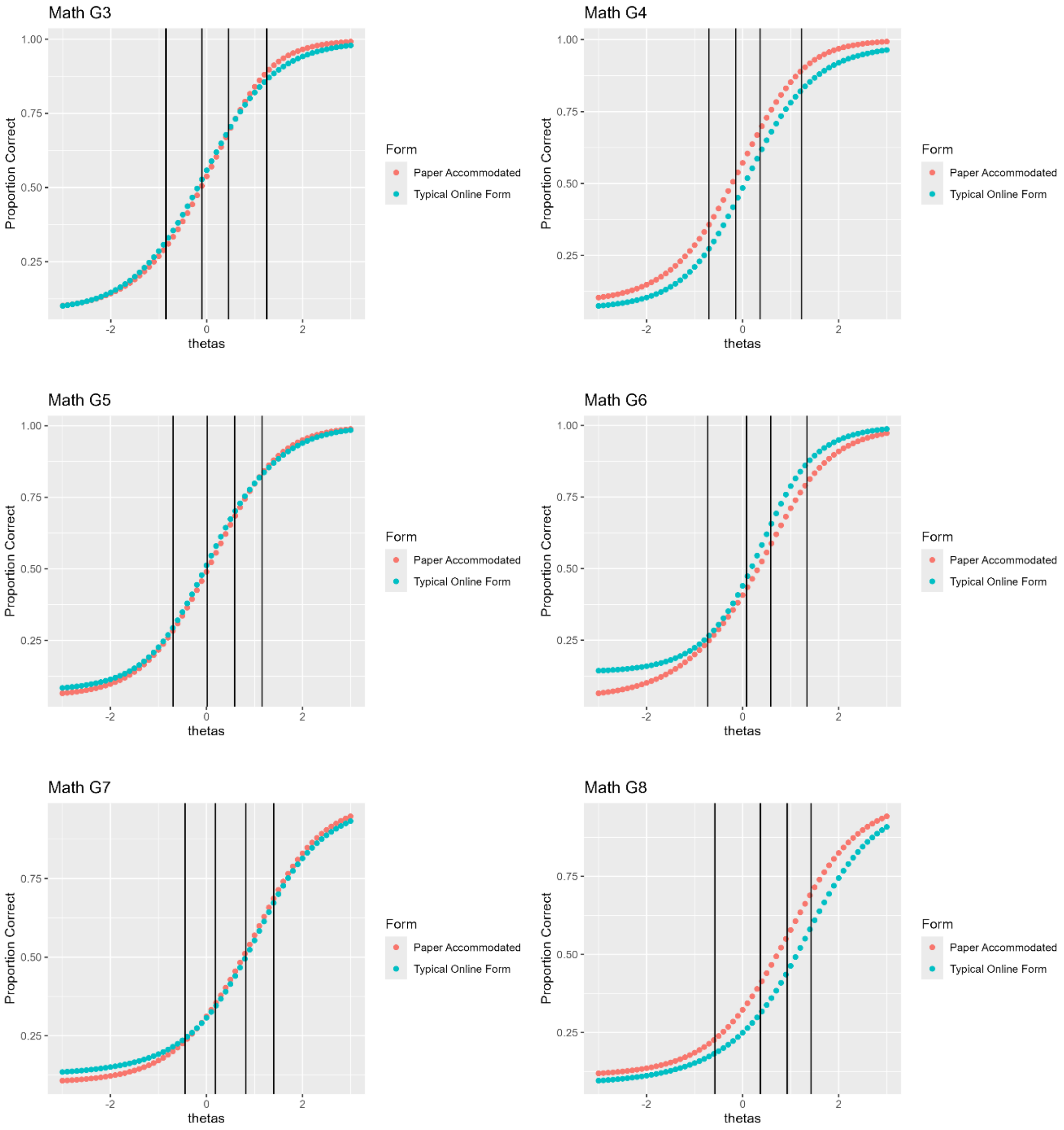


Figure 13: Test Characteristic Curves (TCCs) Compared (ELA)

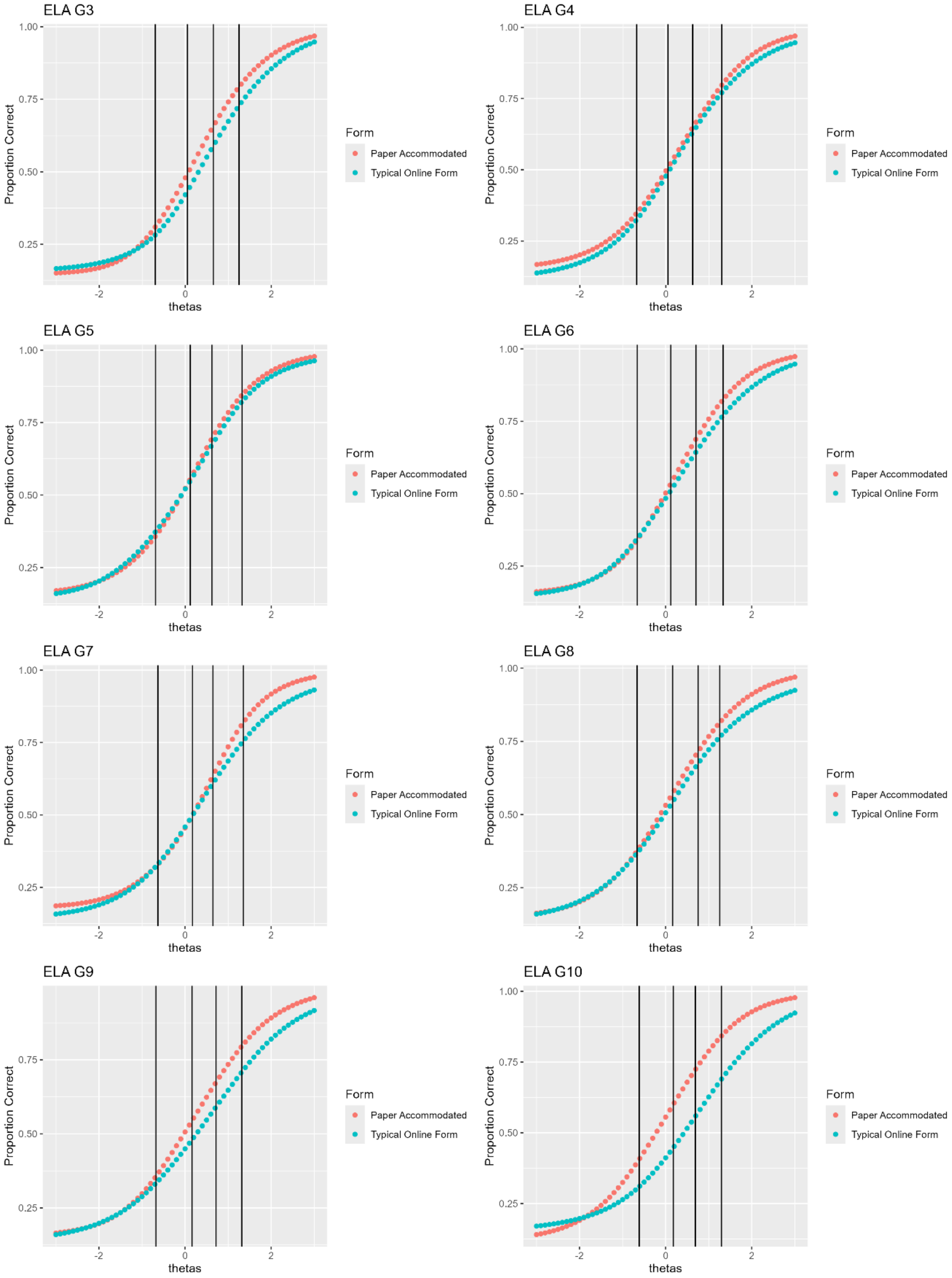


Figure 14: Test Characteristic Curves (TCCs) Compared (EOC)

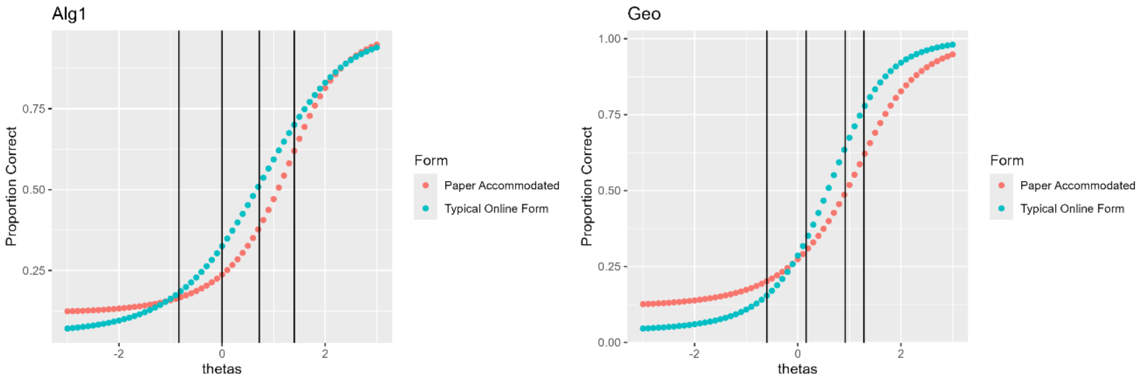
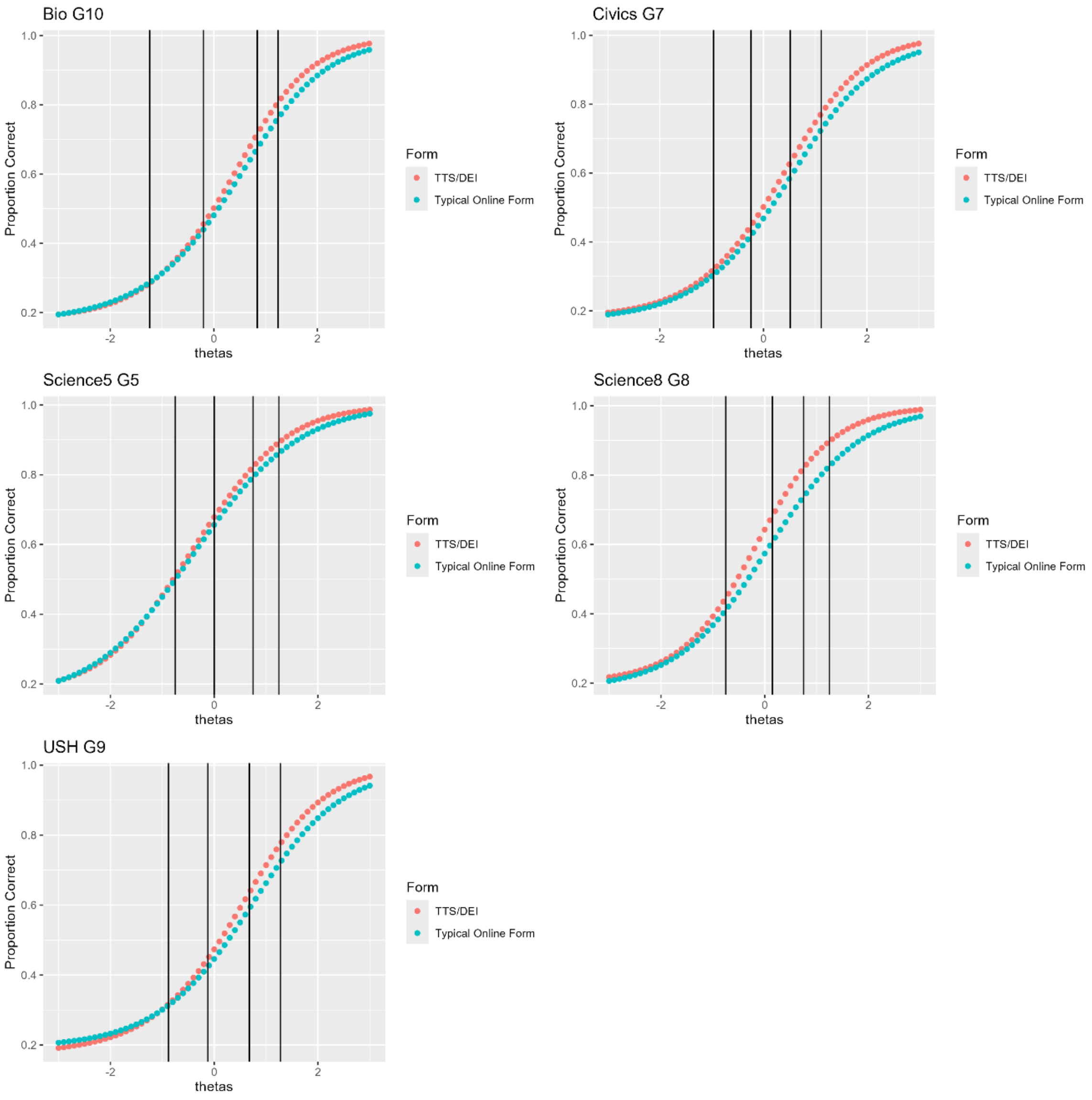


Figure 15: Test Characteristic Curves (TCCs) Compared (Science and Social Studies)



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 the 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 section, we provide evidence that the technical properties of scale scores are comparable between online and 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 the Florida 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 differential item functioning (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, and 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 field testing to evaluate the quality of items, one notable statistic that was used 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 this technical report. The DIF statistics for each test item are presented in the appendices of Volume 1.

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 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.** Appropriate 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 accommodated forms, on different devices.
- **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.*

Arthur, A., Kapoor, S., & Steedle, J. (2020, December). *Paper and online testing mode comparability: A review of research from 2010–2020*. ACT Research & Policy. https://www.act.org/content/dam/act/unsecured/documents/R1842-paper-online-testing-modes-2020-12.pdf

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, *87*(3), 513–524. https://doi.org/10.1037/0033-2909.87.3.513

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS Report No. RR-01-23). Educational Testing Service. https://www.researchgate.net/publication/248940593_Effects_of_Screen_Size_Screen_Resolution_and_Display_Rate_on_Computer-Based_Test_Performance

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

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). Sage.

Chen, F., Bollen, K. A., Paxton, P., Curran P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508. https://doi.org/10.1177/0049124101029004003

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

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.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://asset-pdf.scinapse.io/prod/2089871805/2089871805.pdf

Clark, K., Luong, M-T, Le, Q., & Manning, C. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators.* Paper presented at the Seventh International Conference on Learning Representations. arXiv. https://doi.org/10.48550/arXiv.2003.10555.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.), Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://conservancy.umn.edu/bitstream/handle/11299/184279/1_07_Cronbach.pdf?sequence

Davis, L., Morrison, K., Kong, X., & McBride, Y. (2017). Disaggregated effects of device on score comparability. *Educational Measurement: Issues and Practice*, *36*(3), 35–45. https://doi.org/10.1111/emip.12158

Davis, L. L., Kong, X., & McBride, M. (2015, April). *Device comparability of tablets and computers for assessment purposes* [Paper presentation]. National Council on Measurement in Education annual meeting, Chicago, IL, United States. https://docs.acara.edu.au/resources/20150409_NCME_DeviceComparabilityofTablesComputers.pdf

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. https://doi.org/10.1080/08957347.2016.1243538

Davis, L.L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment*, *20*(3), 180–198. https://doi.org/10.1080/10627197.2015.1061426

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* [White paper]. Pearson.

Deerwester, S., Dumais, S.T., & Landauer, T.K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6),* 391–407.

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED610777.pdf

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

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, *11*(6), 1–9. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1192&context=pare

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

http://expsylab.psych.uoa.gr/fileadmin/expsylab.psych.uoa.gr/uploads/papers/Hu_Bentler_1999.pdf

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). 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 presentation]. American Educational Research Association annual meeting, New Orleans, LA, United States.

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. https://doi.org/10.1177/014662102237797

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109–126). Guilford Press.

Mislevy, J. L., Rupp, A. A., & Harring, J. R. (2012). Detecting local item dependence in polytomous adaptive data. *Journal of Educational Measurement*, *49*(2), 127–147. http://www.jstor.org/stable/41653580

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement*, *29*(2), 133–161. https://www.jstor.org/stable/1434599

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https ://doi.org/10.1007/BF02294210

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*. Conditionally accepted for publication in *Psychometrika*. https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (2012). Mplus user’s guide, 7th Edition.

New York State Education Department (2022). *New York state testing program 2022: English language arts and mathematics grades 3–8.* https://files.eric.ed.gov/fulltext/ED591458.pdf

Olsen, J. B. (2014). *Score comparability for web and iPad delivered adaptive tests* [Paper presentation]. Council on Measurement in Education meeting, Philadelphia, PA, United States.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

PARCC (2015, March 9). *Research Results of PARCC Automated Scoring Proof of Concept Study.* Retrieved from: http://www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf.

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). https://doi.org/10.7275/an9m-2035

Rudner, L. M. (2005) Expected classification accuracy. *Practical Assessment, Research & Evaluation*, 10(13), 1–4. https://doi.org/10.7275/56a5-6b14

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. U.S. Government Printing Office. 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*. University of Kansas.

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *PARCC spring 2015 digital devices comparability research study*. 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*(2), 132–152. https://doi.org/10.1080/08957341003673773

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*(2), 225–243. https://doi.org/10.1007/BF02293865

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need.* Paper presented at the 31st Conference on Neural Information Processing Systems.

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). Routledge.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1),* 2-13.

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

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*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

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*(6), 325–338. https://doi.org/10.1177%2F0145482X1210600602

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, 239-263.