State of Florida

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

2024–2025

Volume 1
Annual Technical Report

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 FDOE (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from CAI: Dr. Myvan Bui, Dr. Sherry Li, Dr. Peter Diao, Matt Gordon, Zoe Dai, Oliver Brown, Daniel Lam, and Dr. Yuan Hong. The major contributors from FDOE are as follows: Susie Lee, Racquel Harrell, Sally Donnelly, Dr. Shakia Johnson, Kristina Lamb, Jenny Black, Catherine Altmaier, Dr. Esra Kocyigit, Dr. Salih Binici, Wenyi Li, Saeyan Yun, and Yiting Yao.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This technical report describes the Florida Assessment of Student Thinking (FAST) assessments for Grades 3–10 English Language Arts (ELA) and Grades 3–8 mathematics; the Benchmarks for Excellent Student Thinking (B.E.S.T.) assessments for writing, and End-of-Course (EOC) Algebra 1 and Geometry; and Florida's state academic standards assessments in Grades 5 and 8 and the EOC assessments for Biology 1, U.S. History, and Civics. The details of the Voluntary Prekindergarten (VPK) to Grade 2 assessments in reading and mathematics are provided in the *Renaissance Learning Star Assessments™ for Reading Technical Manual – Florida* and *Star Assessments™ for Math Technical Manual – Florida*.

Beginning with the 2022–2023 school year, Florida's statewide standardized assessments in reading, writing, and mathematics were aligned with B.E.S.T. standards. A portion of the FAST assessments were administered as progress monitoring (PM) assessments. They include VPK through Grade 10 ELA and VPK through Grade 8 mathematics assessments. B.E.S.T. assessments that are not part of the FAST PM program are Grades 4–10 writing and EOC assessments in Algebra 1 and Geometry.

The PM program does not include Florida's science and social studies assessments, which are aligned to respective state academic standards approved in 2008. Revised civics and government (CG) standards were adopted by the State Board of Education on July 14, 2021, after House Bill 807 (2019) required the Florida Department of Education (FDOE) to review the statewide civics education course standards.

The *Florida Statewide Assessments 2024–2025 Technical Report* documents all methods used in test construction, outlines psychometric properties of the tests, summarizes student results, and documents evidence and support for intended uses and interpretations of the test scores. The technical reports are written as separate, self-contained volumes. They consist of the following:

1. ***Annual Technical Report.*** Volume 1 is updated each year and provides a global overview of the tests administered to students.
2. ***Test Development.*** Volume 2 summarizes the adaptive algorithm and procedures used to construct test forms. It also provides summaries of the item development process.
3. ***Standard Setting.*** Volume 3 documents the methods and results of the B.E.S.T. standard-setting process for the ELA and mathematics assessments, as well as the state academic standards standard setting for science and social studies. This volume is not updated each year because standard setting was finalized between 2012 and 2015 for science and social studies and 2023 for ELA and mathematics.
4. ***Evidence of Reliability and Validity.*** Volume 4 provides technical summaries of the test quality and special studies to support the intended uses and interpretations of the test scores.
5. ***Summary of Test Administration Procedures.*** Volume 5 describes the methods used to administer all forms, security protocols, and modifications or accommodations available.
6. ***Score Interpretation Guide.*** Volume 6 describes the score types reported and the appropriate inferences that can be drawn from each score reported.

7. ***Special Studies.*** During the year, FDOE may request technical studies to investigate issues surrounding the test. This volume, labeled as Volume 7 when required, comprises a set of reports provided to FDOE in support of any requests to further investigate test quality, validity, or other issues as identified. There are no reports to include in this volume for 2024–2025.

These volumes are available for download at FDOE's public-facing website: K-12 Student Assessment Technical Reports. Appendices for technical reports are available upon request by contacting the Office of Assessment at Assessment@fldoe.org.

1.1 Purposes and Intended Uses of the Assessments

The primary purpose of Florida's K–12 assessment system is to measure students' achievement of Florida's education standards. The assessment process supports instruction and student learning, and test 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.

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. Assessment results are also used in teacher evaluations to measure how effectively teachers move student learning forward. Florida's assessment and accountability efforts have had a significant positive impact on student achievement over time.

The tests are constructed to meet rigorous technical criteria in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014) and to ensure that all students have access to the test content via the principles of universal design and appropriate accommodations. Information about the academic standards to which the assessments are aligned and the test blueprints is presented in Volume 2, *Test Development.* Additional verification of content validity is presented in Volume 4, *Evidence of Reliability and Validity.* The documentation about the comparability of online and accommodated tests can also be found in Volume 4.

Florida's assessments yield test scores that are useful for understanding whether individual students have a firm grasp of the Florida standards and whether student performance is improving over time. Additionally, scores can be aggregated to evaluate the performance of subgroups, and both individual and aggregated scores can be compared over time using program evaluation methods. The reliability of the test scores is presented in Volume 4.

The assessments are criterion-referenced tests intended to measure whether students have progressed on the B.E.S.T. standards in ELA and mathematics and Florida's state academic standards in science and social studies. The standards and test blueprints are discussed in Volume 2.

Table 1 outlines the required uses of Florida's assessments based on Florida Statute (F.S.) and Florida Administrative Code (F.A.C.).

Table 1: Required Uses and Citations for Florida's Assessments

Assessment	Assessment Citation	Required Use	Required Use Citation
Statewide Assessment Program	s. 1008.22, F.S. s. 1008.25, F.S. Rule 1.09422, F.A.C. Rule 1.0943, F.A.C Rule 6A 1.09430, F.A.C. Rule 1.09432, F.A.C.	Third Grade Retention; Student Progression; Remedial Instruction; Reporting Requirements; Progress Monitoring	s. 1008.25, F.S. Rule 6A-1.094221, F.A.C. Rule 6A-1.094222, F.A.C.
		Middle Grades Promotion	s. 1003.4156, F.S.
		VPK Education Program Accountability	s. 1002.68, F.S.
		High School Standard Diploma	s. 1003.4282, F.S.
		Standard High School Diploma Designations	s. 1003.4285, F.S.
		High School Graduation Requirements for Students with Disabilities	Rule 6A-1,09963, F.A.C,
		School Grades	s. 1008.34, F.S. Rule 6A-1.09981, F.A.C.
		School Improvement Rating	s. 1008.341, F.S. Rule 6A-1.099822, F.A.C. Rule 6A-1.099828, F.A.C.
		Personnel Evaluations	s. 1012.34 Rule 6A-5.0411, F.A.C.
		District Grades	s. 1008.34, F.S. Rule 6A-1.09981, F.A.C.
		Differentiated Accountability	s. 1008.33, F.S. Rule 6A-1.099811, F.A.C.
		Virtual Instructional Program	s. 1002.45, F.S.
		Opportunity Scholarship Florida Tax Credit Scholarship Family Empowerment Scholarship	s. 1002.38, F.S. s. 1002.395, F.S. s. 1002.394, F.S
		The New Worlds Reading Initiative	s. 1003.485, F.S. Rule 6A-6.0532, F.A.C.
		Implementation of State System of School Improvement and Education Accountability	s. 1008.345, F.S.
		The Reading Achievement Initiative for Scholastic Excellence (RAISE) Program	s. 1008.365 Rule 6A-6.0531, F.A.C.

Appendix F, Glossary of Terms, Abbreviations, and Acronyms, of this volume provides a glossary of terms, abbreviations, and acronyms used throughout the technical report.

1.2 BACKGROUND AND HISTORICAL CONTEXT OF TEST

During the 2022–2023 school year, for ELA and mathematics, FDOE began transitioning from the Florida Standards Assessment (FSA) to the FAST assessment, in accordance with changes to state statutes. Science and social studies remain aligned to Florida's state academic standards adopted by the Florida State Board of Education in 2008. These standards have been previously referred to as Next Generation Sunshine State Standards (NGSSS). Revised civics and government standards were adopted by the State Board of Education on July 14, 2021, after House Bill 807 (2019) required FDOE to complete a review of the statewide Civics education course standards.

In spring 2022, the first set of FAST items developed to align with the B.E.S.T. standards were field-tested. In summer 2022, the field-test items were calibrated and placed on the FSA scale. After the spring 2023 administration of FAST (i.e., PM3) and B.E.S.T. assessments, the items were calibrated to establish new on-grade scales for the FAST assessments and new scales for the B.E.S.T. assessments. The FAST assessments in ELA for Grades 3–10 and in mathematics for Grades 3–8 were placed on a common vertical scale via a linking design that allowed item response theory (IRT) calibrations at each grade to be linked to the adjacent grade scale. All calibration work was completed before the standard-setting workshops conducted on July 24–28, 2023. Standard setting was conducted for all grades in ELA reading, mathematics, B.E.S.T. writing, Algebra 1, and geometry. The newly set cut scores were presented to the State Board of Education for approval. In the 2023–2024 school year and beyond, FDOE reported scores on the new FAST scale.

FAST is administered as a PM assessment. Students participate three times per year: once at the beginning of the year (PM1, August to October), once in the middle of the year (PM2, December to January), and once at the end of the year (PM3, May to June).

- 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 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.

Grades 4–10 writing, which is currently not used in state accountability systems, and the mathematics EOC assessments in Algebra 1 and geometry were developed to assess the B.E.S.T. standards, but they are not part of the FAST PM program.

The ELA and mathematics assessments have been computer adaptive since 2022. Items become progressively harder as students successfully respond to items and easier if students answer more questions incorrectly, but in either scenario, the selected items measure the same knowledge and skills determined by the test blueprint. All assessments, including each PM event, cover the entire test blueprint for the full grade-level content.

Within the current statewide assessment program, students in Grade 3 must score at Level 2 or higher on the Grade 3 ELA assessment to be promoted to Grade 4. Grade 3 students who score at Level 1 may still be promoted through one of seven Good Cause Exemptions that are addressed in the statute and implemented at the district level. Students must score at Level 3 or above on the Grade 10 ELA and Algebra 1 EOC assessments to meet the assessment graduation requirements in the statute. Students who do not score at Level 3 or higher on these assessments can retake the assessments multiple times. They may also use concordant scores on the American College Test (ACT), Classic Learning Test (CLT), or Scholastic Aptitude Test (SAT) to meet the Grade 10 ELA requirement, or they may earn a comparative passing score on the Preliminary Scholastic Aptitude Test (PSAT), SAT, ACT, CLT, or the B.E.S.T. EOC Geometry or Algebra 1. Also, students' scores on the EOC assessments must count for 30% of their final course grade for those courses for which a statewide EOC test is administered.

Beginning in spring 2024, the summative Science assessment in Grades 5 and 8, as well as the Biology 1, Civics, and U.S. History EOC assessments, were delivered in a computer-adaptive format that allows for immediate reporting. While the core content for these tests did not change, some administration details (e.g., reduced test length) and blueprint specifications (e.g., number of items each student will see) have been updated. The fall and winter 2023 administrations of the Science and Social Studies EOC assessments were computer-based, fixed-form tests, and results were available for all students after the testing window, as in previous years.

Recent changes to the assessments are highlighted in this section. This brief background should establish the legislative and curricular framework for the technical analyses described in the remaining sections of this volume and other volumes of the technical report.

Developments in 2012

The NGSSS statewide science assessments were administered on paper in Grades 5 and 8 beginning in spring 2012. Standard-setting meetings for science occurred with educators in September 2012. The online version of NGSSS Biology 1 was first administered to students in spring 2012, and the standard-setting meeting with educators took place in fall 2012.

Developments in 2013

The first online administration of NGSSS U.S. History occurred in spring 2013, and the standard-setting meeting with educators was in fall 2013.

Developments in 2014

The online administration of NGSSS Civics was first administered to students in spring 2014, and the standard-setting meeting with educators took place in fall 2014.

In response to Executive Order 13-276, the state of Florida issued an Invitation to Negotiate to solicit proposals for the development and administration of new assessments aligned to the Florida Standards in ELA and mathematics. After the required competitive bid process, a contract was awarded to Cambium Assessment, Inc. (CAI), previously the American Institutes for Research (AIR), to develop the new FSA. The new assessments reflect the expectations of the Florida Standards, in large part by increasing the emphasis on measuring analytical thinking. As part of this contract, Pearson was responsible for developing test content, building test forms, conducting

psychometric analyses, administering and scoring test forms, and reporting test results for the NGSSS assessments described in this report.

Psychometricians and content experts from CAI, FDOE, and the Department's Test Development Center (TDC) met in summer 2014 to build test forms for spring 2015. Because it was necessary to implement an operational test in the following school year, items from the state of Utah's Student Assessment of Growth and Excellence (SAGE) assessment were used to construct Florida's test forms for the 2014–2015 school year. Assessment experts from FDOE, the Department's TDC, and CAI reviewed each item and its associated statistics to determine their alignment to Florida's academic standards and to judge the suitability of the statistical qualities of each item. Only items deemed suitable from all three perspectives were considered for inclusion on Florida's assessments and for constructing Florida's vertical scale.

From 2014 until 2022–2023, Florida used only post-equating each year. After the spring 2015 administration, all data used for evaluating student performance on the FSA were derived from the Florida population.

In addition to the operational test items, field-test items were embedded into test forms administered online to build the Florida-specific FSA item pool for future use. These items were placed on test forms using an embedded field-test design in the same fixed positions across all test forms within a grade. Many items were field-tested, as described in this volume, to build a substantial item bank and construct future FSA test forms.

It was also necessary to field-test a large pool of text-based writing prompts that could be used for future FSA ELA tests. This objective was accomplished via a stand-alone writing field test during winter 2014–2015. A scientific sample of approximately 25,000 students per grade was selected to participate in this field test, and each student responded to two writing prompts. Approximately 15 prompts were field-tested in each grade. Because only one prompt is used each year, this field test provided data on many prompts for the state. These prompts have been used since spring 2016.

Developments in 2015

The first operational test administration of the FSA occurred in spring 2015. Grades 3 and 4 ELA and mathematics assessments were administered entirely on paper, and all other grades and subjects were administered primarily online. The only exceptions were Grades 4–7 text-based writing and a small percentage of students in each grade and subject who required paper-based tests as accommodations in accordance with an Individualized Education Program (IEP) or Section 504 Plan.

Until new performance standards for this test were in place, statutory requirements called for linking 2015 student performance on Grade 3 ELA, Grade 10 ELA, and Algebra 1 to 2014 student performance on Grades 3 and 10 Florida Comprehensive Assessment Test (FCAT) 2.0 reading and NGSSS Algebra 1 EOC, respectively. This linking was required to determine student-level eligibility for promotion (Grade 3 ELA) and graduation (Grade 10 ELA and Algebra 1), which are also statutory requirements. Equipercentile linking for Grade 10 ELA and Algebra 1 was used to accomplish this. Further legislation enacted in spring 2015 changed the promotion requirement for Grade 3 ELA, instead requiring that student scores in the bottom quintile be identified for districts to use at their discretion in making promotion and retention decisions for that year only.

Existing legislation also prohibits students from being assessed on a grade-level statewide assessment if enrolled in an EOC in the same subject area. Due to this legislation, many students in Grade 8 participated in the Algebra 1 EOC but not in the Grade 8 mathematics assessment. This is detailed in other volumes of the technical report, especially in relation to the Grades 3–8 mathematics vertical scale.

During summer 2015, a new vertical scale for Grades 3–10 ELA and Grades 3–8 mathematics was established using statistics from the spring 2015 administration. Standard-setting meetings for Grades 3–10 ELA; Grades 3–8 mathematics; and EOC Algebra 1, Algebra 2, and geometry were conducted with educators in August and September 2015. The comprehensive process to set performance standards considered the feedback from more than 400 educators from across the state, as well as members of the community, businesses, and district-level education leaders. Additionally, the commissioner considered input from the public, who had the opportunity to submit comments at public workshops and via email, online comment forms, and traditional mail over approximately 12 weeks.

From 2015 until 2024, NGSSS Science (in Grades 5 and 8), Biology 1, U.S. History, and Civics EOC assessments were administered and managed by Pearson.

Developments in 2016

During spring 2016, the Grade 4 ELA reading portion transitioned to an online delivery. A paper form was made available to students whose IEPs or Section 504 Plans indicated such a need. Equating procedures were implemented to ensure comparability between scores in 2015 and 2016.

Developments in 2017

During spring 2017, the Grade 3 and Grade 4 mathematics assessments transitioned to online delivery. A paper form was made available to students whose IEPs or Section 504 Plans indicated such a need.

Developments in 2018

In spring 2018, Algebra 2 was not administered.

Developments in 2019

Per House Bill 7069, some grades and subjects were transitioned to a different mode of delivery beginning in spring 2019. Grades 4–6 reading and Grades 3–6 mathematics moved from online assessments back to paper assessments, and Grade 7 writing transitioned from paper assessments to online assessments in spring 2019.

Developments in 2020

As detailed in the *Special Note for 2019–2020 Annual Technical Report*, the cancellation of the spring 2020 assessments due to the COVID-19 pandemic affected test administration during the 2019–2020 school year. Specifically, as of the cancellation, only Grade 10 ELA writing and reading retakes and Algebra 1 EOC retake were completed, while the spring 2020 regular assessments were canceled. Because of the cancellation, no empirical data that depended on the spring 2020 regular assessments were available to populate the tables in the technical report. Therefore, results were reported based on the spring 2019 regular assessments for assessments that

were uncompleted before the cancellation, whereas results were reported based on spring 2020 for assessments that were completed before the cancellation.

Developments in 2021

Because of the cancellation of the spring 2020 regular assessments, FDOE could not field-test numerous newly developed items across all subjects in 2020 and could not replenish the item bank with statistics for these items. The number of field-test forms was increased in spring 2021 so that items developed in both 2020 and 2021 could be field-tested. This plan was feasible because Florida's large population of approximately 200,000 students per grade and subject helped in obtaining sufficient sample sizes for all field-test items. Statistics for the field-test items developed in both 2020 and 2021 are included in the *Florida Statewide Assessments 2020–2021 Technical Report*. FDOE reviewed all field-test items developed in 2020 to ensure that they were free from any bias or sensitivity issues due to the ongoing COVID-19 pandemic before they were field-tested in spring 2021.

Developments in 2022

Under the guidelines of Florida's new B.E.S.T. standards, new items were developed in Grade 3 reading, Grades 4–10 ELA, Grades 3–8 mathematics, and mathematics EOC tests (i.e., Algebra 1 and geometry). These items were field-tested in spring 2022. The B.E.S.T. items are used to develop the FAST assessments in Grades 4–10 reading and Grades 3–8 mathematics and the B.E.S.T. assessments for Algebra 1 and geometry EOC.

Developments in 2023

During the 2022–2023 school year, FDOE began transitioning from FSA to FAST. In spring 2022, the first set of FAST items developed to align with B.E.S.T. standards was field-tested.

Standard setting was conducted for all grades in ELA reading (K–10), mathematics (K–8), ELA writing (4–10), Algebra 1, and geometry. The State Board of Education presented the newly set cut scores for approval. In the 2023–2024 school year, FDOE began reporting scores on the new FAST scale.

The assessments transitioned from fixed-form tests to computer-adaptive testing for ELA and mathematics (including EOC Algebra 1 and geometry). For ELA Grades 3–10 and mathematics Grades 3–8, tests were administered over three PM periods: formative assessments in PM1 and PM2, culminating in a summative assessment in PM3. The writing assessments were decoupled from ELA and administered as an independent field test based on a representative sample of schools.

Developments in 2024

Beginning in spring 2024, the summative science assessment in Grades 5 and 8, as well as the Biology 1, Civics, and U.S. History EOC assessments, were delivered in a computer-adaptive format that allows for immediate reporting via CAI systems. These tests remain aligned to Florida's Statewide Standards adopted by the Florida State Board of Education in 2008. However, the entire bank was recalibrated to update the bank's item parameters. The new parameters were

placed back onto the existing scale by a linking design to maintain a connection to the standards and cut scores established in prior years.

B.E.S.T. Writing is administered once a year, with the first operational administration in spring 2024. The assessments are computer-based for all grade levels and consist of one text-based constructed-response item (i.e., students read a variety of texts and respond to a prompt). The rubrics used for the scoring of the Writing assessment are based on the B.E.S.T. ELA standards. While the FSA Writing assessment contributed to the overall ELA score (combined with FSA Reading), the new B.E.S.T. Writing assessment is a stand-alone raw score test that does not contribute to the FAST ELA Reading score.

1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF THE ASSESSMENTS

FDOE manages the FAST and B.E.S.T. programs with the assistance of multiple offices within FDOE, Florida educators, a Technical Advisory Committee (TAC), and vendors. FDOE fulfills the diverse requirements for implementing Florida's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014).

Florida Department of Education

Office of K–12 Student Assessment: The Office of K–12 Student Assessment oversees all aspects of Florida's statewide assessment program, including coordination with other FDOE offices, Florida public schools, and vendors.

Test Development Center: Funded by FDOE via a grant, the TDC works with Florida educators and vendors to develop test specifications and content and to build test forms.

Florida Educators

Florida educators participate in most aspects of the conceptualization and development of the Florida assessments. Educators help to develop the academic standards, clarify how these standards will be assessed, design the test, and review test items and passages.

Technical Advisory Committee

FDOE convenes a panel twice per year to discuss psychometric, test development, administrative, and policy issues relevant to current and future Florida testing. This committee is made up of several nationally recognized assessment experts and highly experienced practitioners from multiple Florida school districts.

Cambium Assessment, Inc.

CAI was the vendor selected through the state-mandated competitive procurement process. CAI is responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the assessments described in this report. All activities were conducted under the close direction of FDOE staff experts.

Human Resources Research Organization

The Human Resources Research Organization (HumRRO) has provided program evaluation to a wide variety of federal and state agencies, as well as corporate and nonprofit organizations and foundations. HumRRO conducts independent checks on the calibration, equating, and linking activities; reports its findings directly to FDOE; and provides consultative services to FDOE on psychometric matters.

Buros Institute of Mental Measurements

The Buros Institute of Mental Measurements (Buros) provides users of commercially published tests with professional assistance, expertise, and information. Buros provided independent operational checks on the psychometric services provided by CAI. Each year, Buros delivers reports on their observations, which are available on request.

Caveon Test Security

Caveon Test Security analyzes the assessment data using Caveon Data Forensics™ to identify highly unusual test results for two primary groups: (1) students with extremely similar test scores and (2) schools with improbable levels of similarity, gains, and/or erasures. Caveon also provides annual services related to on-site monitoring of test administration in samples of school districts.

1.4 Available Test Formats and Special Versions

Summative assessments for students in Grades 3–10 reading and Grades 3–8 mathematics, Algebra 1 and geometry EOC, Grades 5 and 8 science, Biology 1, U.S. History, and Civics are administered as online computer-adaptive tests during each spring. This includes PM3 for Grades 3–10 reading and Grades 3–8 mathematics. For all these assessments, accommodated versions are available to students whose IEPs or Section 504 Plans indicate such a need.

Administered tests contain operational items and embedded field-test (EFT) items randomly distributed throughout the test in field-test slots. Operational items are used to calculate student scores. EFT items are nonscored items and are used to populate the bank for future operational use.

1.5 Student Participation

By statute, all Florida public school students are required to participate in the statewide assessments. Students take mathematics, reading, writing, science, social studies, and EOC tests in the spring. Retake administrations for the EOC assessments occur in the summer, fall, and winter. Grade 10 ELA retake administrations occur in the fall, winter, spring, and summer.

Tables 2–6 show the number of students who were tested and the number of students who were reported in 2024–2025 by grade and subject area for online tests. The difference is due to the number of students who did not meet Florida's requirement of attempting at least six items to receive a reported score. Information for students who took accommodated forms is available in this volume, Section 2.2, Accommodations. The participation counts by subgroup, including gender, ethnicity, special education, and English language learner (ELL) status, are presented in this volume, Section 1.6, Demographics of Tested Population. Tables 7–10 present the percentages of students in each performance level for grades and subjects that were reported for the spring.

Appendix D, Distribution of Scale Scores and Standard Errors, presents descriptive statistics on the scale score distributions across all students and subgroups. Writing was reported with only raw scores (no scale scores or performance levels).

Table 2: Number of Students Participating in FAST and B.E.S.T. Assessments (PM3/Spring)

Mathematics			ELA Reading		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**
3	218,578	218,525	3	219,281	219,202
4	197,312	197,257	4	203,618	203,568
5	209,550	209,483	5	211,908	211,842
6	195,955	195,765	6	200,018	199,866
7	134,886	134,560	7	204,802	204,573
8	169,047	168,509	8	216,471	216,159
Algebra 1	271,981	270,651	9	212,507	212,057
Geometry	221,841	220,733	10	214,315	213,781

Table 3: Number of Students Participating in Science and Social Studies Assessments (Spring)

Science			Science & Social Studies EOC		
Grade	**Number Tested**	**Number Reported**	**Test**	**Number Tested**	**Number Reported**
5	178,081	178,048	Biology 1	204,698	204,102
8	178,997	178,710	Civics	184,933	184,532
			U.S. History	188,445	187,926

Table 4: Number of Students Participating in Writing Assessments (Spring)

B.E.S.T. Writing		
Grade	**Number Tested**	**Number Reported**
4	202,333	202,271
5	210,605	210,539
6	198,162	198,052
7	202,710	202,565
8	213,835	213,678
9	208,251	208,028
10	208,468	208,180

Table 5: Number of Students Participating in FAST Assessments (PM1)

Mathematics			ELA Reading		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**
3	221,084	220,965	3	222,851	222,675
4	194,310	194,187	4	200,893	200,657
5	210,601	210,525	5	211,939	211,874
6	196,484	196,236	6	200,147	199,978
7	137,347	136,509	7	205,597	205,427
8	163,587	163,348	8	216,807	216,592
			9	214,633	214,292
			10	216,869	216,520

Table 6: Number of Students Participating in FAST Assessments (PM2)

Mathematics			ELA Reading		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**
3	219,049	218,981	3	219,596	219,521
4	196,763	196,724	4	203,950	203,910
5	210,464	210,399	5	212,202	212,162
6	196,321	196,192	6	200,196	200,074
7	134,760	134,384	7	205,274	205,133
8	168,550	168,225	8	216,502	216,325
			9	212,959	212,652
			10	214,011	213,714

Table 7: Percentage of Students Across Performance Levels by Grade (PM3/Spring)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	16.1	20.8	21.6	27.3	14.2
	4	19.9	18.1	20.3	29.0	12.7
	5	20.6	22.5	22.4	18.7	15.8
	6	16.4	23.6	19.1	24.6	16.2
	7	28.1	21.5	22.7	15.4	12.3
	8	19.1	24.5	19.3	16.1	21.0
ELA Reading	3	21.8	21.0	22.7	20.6	13.9
	4	22.4	21.1	21.8	21.8	12.8
	5	20.0	24.3	20.3	23.2	12.1

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
	6	19.4	21.0	23.1	22.3	14.1
	7	21.8	21.5	18.5	25.2	13.0
	8	22.4	22.8	23.0	16.7	15.0
	9	21.1	23.5	22.3	20.3	12.7
	10	19.4	23.2	22.0	21.0	14.4
EOC	Algebra 1	22.5	23.4	23.0	19.0	11.9
	Geometry	22.3	23.4	27.5	10.8	15.8

Table 8: Percentage of Students Across Performance Levels by Grade (Science and Social Studies—Spring)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Biology 1	10	9.1	17.7	35.9	13.7	23.5
Civics	7	12.0	13.7	23.7	20.9	29.6
U.S. History	9	13.4	13.6	22.8	19.1	31.0
Grade 5 Science	5	17.9	21.6	28.2	14.7	17.5
Grade 8 Science	8	20.7	26.4	23.2	15.1	14.5

Table 9: Percentage of Students Across Performance Levels by Grade (PM1)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	63.6	28.2	6.4	1.6	0.2
	4	68.7	21.4	6.6	2.8	0.4
	5	55.6	29.8	10.3	3.4	0.9
	6	42.3	37.7	12.7	6.1	1.1
	7	47.2	27.5	17.7	5.7	1.7
	8	44.6	37.2	12.6	4.1	1.4
ELA Reading	3	48.6	26.9	15.1	7.2	2.1
	4	42.0	25.6	17.4	11.3	3.6
	5	38.5	28.5	16.0	12.8	4.2
	6	28.8	26.7	21.9	15.8	6.7
	7	32.6	27.2	16.9	17.1	6.2
	8	35.9	27.4	19.6	10.4	6.6
	9	33.7	26.9	18.9	13.6	6.8
	10	36.3	26.1	16.9	13.1	7.5

Table 10: Percentage of Students Across Performance Levels by Grade (PM2)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	36.0	35.2	18.7	8.5	1.6
	4	48.9	26.4	14.0	9.0	1.7
	5	36.5	32.3	19.4	9.0	2.8
	6	26.1	35.5	20.2	14.4	3.7
	7	40.0	27.3	20.3	8.6	3.8
	8	31.5	34.5	18.4	9.9	5.7
ELA Reading	3	34.9	24.8	20.4	13.6	6.2
	4	33.1	23.9	19.9	16.1	6.9
	5	29.1	27.4	18.9	17.5	7.1
	6	25.3	23.8	22.3	18.8	9.8
	7	28.8	24.2	17.5	20.7	8.8
	8	30.8	25.3	20.8	13.1	10.0
	9	28.9	26.0	20.4	15.8	8.9
	10	30.3	25.6	18.7	15.6	9.7

1.6 DEMOGRAPHICS OF TESTED POPULATION

Tables 11–20 present the distribution of students, in counts and percentages, who participated in each administration by grade and subject. The numbers are based on the reported status in the approved spring State Student Results (SSR) files and include only online test takers. Information for students who took accommodated tests is presented in Section 2.2, Accommodations. The subgroups reported are gender, ethnicity, students with disabilities (SWD), and ELL. Section 1.2, Testing Accommodations, Volume 5, Summary of Test Administration Procedures, of this technical report provides explicit definitions for the two major subgroups to which accommodations are available: ELL and SWD. Students who are offered accommodations may choose not to use the accommodations.

Table 11: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM3)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	218,525	106,721	111,804	44,304	84,444	72,545	28,365	40,247
	%	100	48.84	51.16	20.27	38.64	33.20	12.98	18.42
4	*N*	197,257	97,391	99,866	39,287	75,413	67,023	26,491	29,509
	%	100	49.37	50.63	19.92	38.23	33.98	13.43	14.96
5	*N*	209,483	102,755	106,728	42,262	80,920	70,370	29,836	28,382
	%	100	49.05	50.95	20.17	38.63	33.59	14.24	13.55
6	*N*	195,765	96,417	99,348	39,260	75,707	65,896	25,563	22,182
	%	100	49.25	50.75	20.05	38.67	33.66	13.06	11.33

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
7	*N*	134,560	66,811	67,749	31,618	54,239	40,251	22,473	19,557
	%	100	49.65	50.35	23.50	40.31	29.91	16.70	14.53
8	*N*	168,509	81,821	86,688	37,650	67,302	52,185	25,846	21,166
	%	100	48.56	51.44	22.34	39.94	30.97	15.34	12.56

Table 12: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM3)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,202	106,979	112,223	44,495	84,676	72,736	28,434	40,328
	%	100	48.80	51.20	20.30	38.63	33.18	12.97	18.40
4	*N*	203,568	100,150	103,418	39,977	77,508	69,850	26,720	30,062
	%	100	49.20	50.80	19.64	38.07	34.31	13.13	14.77
5	*N*	211,842	103,815	108,027	42,522	81,562	71,418	29,865	28,250
	%	100	49.01	50.99	20.07	38.50	33.71	14.10	13.34
6	*N*	199,866	98,157	101,709	39,914	76,525	67,685	25,633	22,087
	%	100	49.11	50.89	19.97	38.29	33.87	12.83	11.05
7	*N*	204,573	100,863	103,710	41,389	79,063	68,413	25,289	21,719
	%	100	49.30	50.70	20.23	38.65	33.44	12.36	10.62
8	*N*	216,159	105,558	110,601	45,152	85,013	70,208	28,483	22,229
	%	100	48.83	51.17	20.89	39.33	32.48	13.18	10.28
9	*N*	212,057	104,746	107,311	43,524	82,121	70,814	25,389	20,736
	%	100	49.40	50.60	20.52	38.73	33.39	11.97	9.78
10	*N*	213,781	105,930	107,851	44,428	82,491	71,160	24,356	19,153
	%	100	49.55	50.45	20.78	38.59	33.29	11.39	8.96

Table 13: Distribution of Demographic Characteristics of Tested Population, B.E.S.T. Writing

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
4	*N*	202,268	99,539	102,729	39,664	76,983	69,489	26,622	29,797
	%	100	49.21	50.79	19.61	38.06	34.35	13.16	14.73
5	*N*	210,533	103,195	107,338	42,182	81,077	71,032	29,968	28,114
	%	100	49.02	50.98	20.04	38.51	33.74	14.23	13.35
6	*N*	198,006	97,313	100,693	39,366	75,916	67,096	25,747	21,847
	%	100	49.15	50.85	19.88	38.34	33.89	13.00	11.03
7	*N*	202,526	99,845	102,681	40,753	78,372	67,837	25,406	21,506

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
	%	100	49.30	50.70	20.12	38.70	33.50	12.54	10.62
8	*N*	213,541	104,236	109,305	44,344	84,230	69,376	28,509	22,030
	%	100	48.81	51.19	20.77	39.44	32.49	13.35	10.32
9	*N*	207,978	102,769	105,209	42,255	80,565	69,797	25,158	20,275
	%	100	49.41	50.59	20.32	38.74	33.56	12.10	9.75
10	*N*	207,831	102,869	104,962	42,649	80,466	69,367	23,863	18,730
	%	100	49.50	50.50	20.52	38.72	33.38	11.48	9.01

Table 14: Distribution of Demographic Characteristics of Tested Population, Mathematics EOC

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Algebra 1	*N*	270,651	132,739	137,912	58,887	107,354	85,445	33,105	29,802
	%	100	49.04	50.96	21.76	39.67	31.57	12.23	11.01
Geometry	*N*	220,733	108,921	111,812	46,521	85,644	72,478	24,631	19,323
	%	100	49.35	50.65	21.08	38.80	32.84	11.16	8.75

Table 15: Distribution of Demographic Characteristics of Tested Population, Biology 1, U.S. History, Civics

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Biology 1	*N*	204,102	102,867	101,235	42,316	79,941	66,458	11,141	19,393
	%	100	50.40	49.60	20.73	39.17	32.56	5.46	9.50
U.S. History	*N*	187,926	94,666	93,260	38,712	72,799	62,825	12,616	16,310
	%	100	50.37	49.63	20.60	38.74	33.43	6.71	8.68
Civics	*N*	184,532	93,096	91,436	37,182	72,581	60,351	6,015	21,960
	%	100	50.45	49.55	20.15	39.33	32.70	3.26	11.90

Table 16: Distribution of Demographic Characteristics of Tested Population, Science (Grades 5 and 8)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
5	*N*	178,048	90,048	88,000	34,435	68,924	60,260	3,639	23,946
	%	100	50.58	49.42	19.34	38.71	33.84	2.04	13.45
8	*N*	178,710	89,514	89,196	36,280	69,838	59,138	6,427	20,243
	%	100	50.09	49.91	20.30	39.08	33.09	3.60	11.33

Table 17: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM1)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	220,965	107,647	113,318	45,194	84,908	73,578	25,495	40,720
	%	100	48.72	51.28	20.45	38.43	33.30	11.54	18.43
4	*N*	194,187	95,936	98,251	38,547	73,510	66,749	24,336	28,997
	%	100	49.40	50.60	19.85	37.86	34.37	12.53	14.93
5	*N*	210,525	103,237	107,288	42,510	80,634	71,424	28,871	29,114
	%	100	49.04	50.96	20.19	38.30	33.93	13.71	13.83
6	*N*	196,236	96,687	99,549	39,325	75,197	66,767	25,162	22,687
	%	100	49.27	50.73	20.04	38.32	34.02	12.82	11.56
7	*N*	136,509	67,765	68,744	31,855	53,704	42,214	22,625	18,981
	%	100	49.64	50.36	23.34	39.34	30.92	16.57	13.90
8	*N*	163,348	79,587	83,761	36,997	65,293	50,158	25,649	20,739
	%	100	48.72	51.28	22.65	39.97	30.71	15.70	12.70

Table 18: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM1)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	222,675	108,487	114,188	45,700	85,604	73,958	26,134	41,155
	%	100	48.72	51.28	20.52	38.44	33.21	11.74	18.48
4	*N*	200,657	98,674	101,983	39,062	75,508	70,000	24,415	29,337
	%	100	49.18	50.82	19.47	37.63	34.89	12.17	14.62
5	*N*	211,874	103,789	108,085	42,600	80,904	72,051	29,057	28,864
	%	100	48.99	51.01	20.11	38.18	34.01	13.71	13.62
6	*N*	199,978	98,197	101,781	39,821	75,870	68,577	25,384	22,560
	%	100	49.10	50.90	19.91	37.94	34.29	12.69	11.28
7	*N*	205,427	101,233	104,194	41,605	78,567	69,462	25,645	21,516
	%	100	49.28	50.72	20.25	38.25	33.81	12.48	10.47
8	*N*	216,592	105,647	110,945	45,200	84,588	70,999	28,954	22,105
	%	100	48.78	51.22	20.87	39.05	32.78	13.37	10.21
9	*N*	214,292	105,957	108,335	43,818	81,953	72,771	25,978	20,045
	%	100	49.45	50.55	20.45	38.24	33.96	12.12	9.35
10	*N*	216,520	106,889	109,631	44,733	83,250	72,735	24,984	19,343
	%	100	49.37	50.63	20.66	38.45	33.59	11.54	8.93

Table 19: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM2)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	218,981	106,814	112,167	44,455	84,354	72,929	26,261	40,347
	%	100	48.78	51.22	20.30	38.52	33.30	11.99	18.42
4	*N*	196,724	97,144	99,580	39,289	74,838	67,124	25,598	29,504
	%	100	49.38	50.62	19.97	38.04	34.12	13.01	15.00
5	*N*	210,399	103,169	107,230	42,417	81,044	70,938	29,388	28,700
	%	100	49.03	50.97	20.16	38.52	33.72	13.97	13.64
6	*N*	196,192	96,554	99,638	39,386	75,461	66,411	25,488	22,370
	%	100	49.21	50.79	20.08	38.46	33.85	12.99	11.40
7	*N*	134,384	66,790	67,594	31,822	53,736	40,389	22,611	19,525
	%	100	49.70	50.30	23.68	39.99	30.05	16.83	14.53
8	*N*	168,225	81,658	86,567	37,510	66,896	52,478	25,992	21,055
	%	100	48.54	51.46	22.30	39.77	31.20	15.45	12.52

Table 20: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM2)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,521	107,070	112,451	44,593	84,554	73,088	26,326	40,451
	%	100	48.77	51.23	20.31	38.52	33.29	11.99	18.43
4	*N*	203,910	100,305	103,605	40,101	77,313	70,268	25,834	30,140
	%	100	49.19	50.81	19.67	37.92	34.46	12.67	14.78
5	*N*	212,162	103,929	108,233	42,556	81,394	71,831	29,413	28,581
	%	100	48.99	51.01	20.06	38.36	33.86	13.86	13.47
6	*N*	200,074	98,187	101,887	39,970	76,301	68,078	25,575	22,339
	%	100	49.08	50.92	19.98	38.14	34.03	12.78	11.17
7	*N*	205,133	101,117	104,016	41,588	78,918	68,859	25,641	21,814
	%	100	49.29	50.71	20.27	38.47	33.57	12.50	10.63
8	*N*	216,325	105,566	110,759	45,124	84,930	70,548	28,732	22,255
	%	100	48.80	51.20	20.86	39.26	32.61	13.28	10.29
9	*N*	212,652	105,002	107,650	43,515	81,942	71,536	25,786	20,710
	%	100	49.38	50.62	20.46	38.53	33.64	12.13	9.74
10	*N*	213,714	105,573	108,141	44,326	82,516	71,273	24,562	19,388
	%	100	49.40	50.60	20.74	38.61	33.35	11.49	9.07

2. RECENT AND FORTHCOMING CHANGES TO THE TEST

This section highlights and documents any major issues affecting the test or test administration during the current year and any major changes to the test or test administration procedures over time.

In accordance with Section 1008.22(8), Florida Statutes (F.S.), effective June 30, 2021, the Florida Department of Education (FDOE) began releasing each statewide, standardized assessment, excluding assessment retakes, at least once on a triennial basis pursuant to a schedule determined by the commissioner of education. Senate Bill 1108, signed into law on June 22, 2021, changed the initial publication of assessments to June 30, 2024.

During the 2022–2023 school year, FDOE began transitioning from the Florida Standards Assessment (FSA) in mathematics, ELA reading, and writing to assessments aligned to the B.E.S.T. standards. Voluntary prekindergarten through Grade 8 Mathematics and voluntary prekindergarten through Grade 10 ELA reading are administered as progress monitoring (PM) assessments, which are called the FAST assessments. Writing and end-of-course (EOC) assessments are not part of FAST but are aligned to the B.E.S.T. standards.

During the 2022 legislative session, Senate Bill (SB) 1048 was passed and signed into law by Governor Ron DeSantis. Among other measures, the bill provides the following changes to the FAST assessments:

1. It adds Grades 9 and 10 to the ELA assessments administered as part of the PM system.
2. It identifies the third FAST administration in each school year as the statewide, standardized assessment for students in Grades 3–8 for mathematics and Grades 3–10 for ELA reading.
3. It requires the results for the FAST ELA Reading and Mathematics assessments to be available no later than May 31 each year, beginning with the 2023–2024 school year.

Per Section 9 of F.S. 1008.25, FAST assessments will be administered three times per year: the first (PM1) will occur within the first 30 days of school, the second (PM2) will occur in the middle of the school year, and the third (PM3) will occur at the end of the school year.

All FAST assessments are computer adaptive; thus, items may become progressively harder as students successfully respond to items and easier if students answer more items incorrectly. Each PM event is tied to a blueprint for the full grade-level content.

Each subject area test is administered in one day. It is recommended that each student take only one subject test a day. PM1 and PM2 are used for informational purposes only and not used for accountability. PM3 is a summative assessment used for accountability purposes. The baseline year for the new FAST/B.E.S.T. scale is considered 2022–2023. For 2023–2024 and beyond, new cut scores were applied. As with FSA, a Level 3 achievement level on the FAST assessments is considered passing. However, Senate Bill 1048 (2022) revised the definition of a Level 3 score from a "satisfactory performance" to a "grade-level performance."

B.E.S.T. Writing is administered once a year, with the first operational administration in spring 2024. In spring 2023, the B.E.S.T. Writing field test was administered to a representative sample

of Florida students in Grades 4–10. The assessments are computer-based for all grade levels and consist of one text-based constructed-response item (i.e., students read a variety of texts and respond to a prompt). The rubrics used for the scoring of the writing assessment are based on the B.E.S.T. ELA standards. While the FSA Writing assessment contributed to the overall ELA score (combined with FSA Reading), the new B.E.S.T. Writing assessment is a stand-alone test that does not contribute to the FAST ELA Reading score; it is also not used for accountability purposes.

The spring administration of the science and social studies assessments transitioned to computer-adaptive delivery in the 2023–2024 school year. The assessments were drawn from a common item bank that covered the full test blueprints and was reported on the existing scales. A new standard setting was not conducted, but quality assurance analyses were completed in summer 2024. Statewide science and social studies assessments will continue to be summative assessments; they are not part of the FAST PM system. To calibrate the complete item pools for computer-adaptive testing (CAT) in 2023–2024, an operational field-test (OPFT) design was implemented in spring 2024. Thus, item selection was configured to achieve a blueprint match for each test administration, but item selection was independent of item difficulty. Each item was therefore administered randomly to Florida students, supporting calibration of the complete item pools. Starting in spring 2025, item difficulty matching to student ability was turned on and the CAT algorithm was fully adaptive.

2.1 Spring Administration Procedures

Table 21 shows the schedule for the spring administration of the 2024–2025 assessments, broken down by testing window and subject area.

Table 21: Testing Windows by Subject Area

Assessment	Testing Window
Grades 3–10 FAST ELA Reading Grades 3–8 FAST Mathematics	**First Administration (PM1):** August 12–September 27, 2024 **Second Administration (PM2):** December 2, 2024–January 24, 2025 **Third Administration (PM3):** May 1–30, 2025
Grade 10 FAST ELA Reading Retake	September 9–October 4, 2024 December 2–20, 2024 May 1–30, 2025 July 14–25, 2025
Algebra 1 and Geometry	September 9–October 4, 2024 December 2–20, 2024 May 1–30, 2025 July 14–25, 2025
Writing	March 31–April 11, 2025
Grade 5 and 8 Science	May 1–30, 2025

Assessment	Testing Window
Biology 1, Civics, and U.S. History	September 9–October 4, 2024 December 2–20, 2024 May 1–30, 2025 July 14–25, 2025

According to state law, students were required to participate in the spring assessment, and all testing took place during the designated testing window. FAST, B.E.S.T., Science, and Social Studies assessments were administered in timed sessions, but students who did not finish within the session time could continue working up to the end of the school day, with the exception of B.E.S.T. Writing where students could work up to half the school day. Once a session began, a student was required to finish it before leaving the school's campus. A student could not return to a session once he or she left campus.

The key personnel involved with the administration included the district assessment coordinators (DACs), school assessment coordinators (SACs), and test administrators (TAs) who proctored the test. An online TA training course was available to TAs. More detailed information about the roles and responsibilities of the various testing staff is presented in Volume 5, Summary of Test Administration Procedures, of this technical report.

A secure browser developed by Cambium Assessment, Inc. (CAI) (CAI Secure Browser) was required to access the online assessments. The browser provided a secure environment for student testing by disabling the hotkeys, copy, and screen capture capabilities and by blocking access to desktop functionalities, such as the Internet and email. Other measures that protected the integrity and security of the online test are presented in Volume 5.

Students could participate in online tests via multiple platforms, such as Windows, Chrome, Mac, and iPad. Before test administration, a series of user acceptance testing (UAT) is performed on all platforms on which online tests are administered. This is conducted to ensure that the items are rendered as expected and have similar appearances across platforms to minimize potential device effects. In keeping with best practices recommended by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014, Standards 9.7 & 9.9), CAI conducted a device comparability study to provide evidence of comparability of the Florida Statewide Assessments scores across devices. Results of this study are presented in Volume 7, Special Studies, of the *Florida Standards Assessments 2019–2020 Technical Report.* A rigorous review is in place to ensure that the content of the items on accommodated tests matches the content of the items that are administered online (e.g., wording, graphics, paragraph breaks, and option order).

2.2 ACCOMMODATIONS

Florida assessments are designed to be inclusive for all students, which serves as evidence of test validity. To maximize the accessibility of the assessments, various accommodations were provided to students with special needs, as indicated by documentation such as Individualized Education Programs (IEPs) or Section 504 Plans. Such accommodations improve access to state assessments and help students with special needs demonstrate what they know and can do. 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 of available testing accommodations; their selection, use, and implementation; and the appropriateness of the accommodations are covered in Section 1.2, Testing Accommodations, of Volume 5 of this technical report. Also, Section 6.4, Accommodated Forms, of this volume provides details about the accommodated form construction, in addition to Appendix C, Test Characteristic Curves with SEMs, of this volume.

Observed data collected from the test administrations provide evidence that the test forms are equally as reliable and that students using the accommodated form also have a range of scores. This evidence indicates that high-performing students taking an accommodated form can still demonstrate high performance and are not impeded in any way by the nature of the form or its administration. A scale score summary (including mean score, standard deviation, mean conditional standard error of measurement, and marginal reliability) by reporting category is presented for online and accommodated groups in Appendix A, Reliability Coefficients, of Volume 4, *Evidence of Reliability and Validity,* of this technical report.

The TA and the school assessment coordinator were responsible for ensuring that arrangements for accommodations were made before the test administration dates. Various accommodations—such as large print, contracted braille, uncontracted braille, and displaying only one-item-per-page—were available for eligible students participating in accommodated assessments. Such accommodations as masking, text-to-speech (TTS), and regular- or large-print passage booklets were made available for eligible students participating in computer-based assessments. Students could use these accommodations only as dictated on their IEPs or Section 504 Plans. Additional accommodation guidelines are presented in Volume 5 of this technical report.

The number of students who took the accommodated versions of the 2024–2025 assessments is shown in Tables 22–24. When a paper-based version was provided as an accommodation, student responses from the paper-based tests were transcribed into the Data Entry Interface (DEI) to ensure timely results. In addition to accommodations that use DEI, students may receive an auditory presentation of text, referred to as TTS, including directions, prompts, items, and answer choices. TTS tests for science and social studies are adaptive but use a subset of items from the regular science and social studies item bank. These tests have separate TTS test IDs. Students who use these forms are analyzed separately and removed from the spring calibrations. TTS tests for ELA and mathematics (including EOC) are adaptive and use exactly the same bank of items and test IDs as the regular tests. Students who use these forms are analyzed together with the regular online tests and included in the spring calibrations.

Table 22: Counts of Accommodated Assessments by Grades and Subjects (DEI)

Subject	Grade	Spring 2025
Mathematics	3	683
	4	774
	5	858
	6	458
	7	304
	8	239
ELA Reading	3	691
	4	772
	5	866
	6	452
	7	375
	8	279
	9	290
	10	1033
Mathematics EOC	Algebra 1	394
	Geometry	310
Biology 1	10	312
U.S. History	9	243
Civics	7	349
Grade 5 Science	5	826
Grade 8 Science	8	235

Table 23: Counts of Accommodated Assessments by Grades and Subjects (Science and Social Studies [TTS])

Subject	Grade	Spring 2025
Biology 1	10	17,087
U.S. History	9	11,070
Civics	7	25,436
Grade 5 Science	5	32,874
Grade 8 Science	8	25,649

Table 24: Counts of Accommodated Assessments by Grades and Subject (B.E.S.T. Writing)

Subject	Grade	Spring 2025
Writing	4	706
	5	736
	6	441
	7	348
	8	237
	9	240
	10	312

Tables 25–30 present the distribution of accommodated students, in counts and percentages, who participated in the spring administration by grade and subject. The subgroups reported are gender, ethnicity, students with disabilities (SWD), and English language learners (ELLs).

Table 25: Distribution of Demographic Characteristics of Tested Accommodated Population, Mathematics

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	683	230	453	169	324	166	577	145
	%	100	33.67	66.33	24.74	47.44	24.30	84.48	21.23
4	*N*	774	287	487	200	351	193	649	140
	%	100	37.08	62.92	25.84	45.35	24.94	83.85	18.09
5	*N*	858	334	524	214	376	230	689	117
	%	100	38.93	61.07	24.94	43.82	26.81	80.30	13.64
6	*N*	458	194	264	111	196	129	356	37
	%	100	42.36	57.64	24.24	42.79	28.17	77.73	8.08
7	*N*	304	145	159	72	141	82	238	17
	%	100	47.70	52.30	23.68	46.38	26.97	78.29	5.59
8	*N*	239	121	118	62	94	71	167	13
	%	100	50.63	49.37	25.94	39.33	29.71	69.87	5.44

Table 26: Distribution of Demographic Characteristics of Tested Accommodated Population, ELA Reading

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	691	234	457	169	328	169	582	145
	%	100	33.86	66.14	24.46	47.47	24.46	84.23	20.98
4	*N*	772	290	482	198	351	190	643	139
	%	100	37.56	62.44	25.65	45.47	24.61	83.29	18.01

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
5	*N*	866	334	532	216	374	240	693	117
	%	100	38.57	61.43	24.94	43.19	27.71	80.02	13.51
6	*N*	452	187	265	112	203	116	356	40
	%	100	41.37	58.63	24.78	44.91	25.66	78.76	8.85
7	*N*	375	181	194	76	173	114	271	22
	%	100	48.27	51.73	20.27	46.13	30.40	72.27	5.87
8	*N*	279	132	147	68	111	87	192	11
	%	100	47.31	52.69	24.37	39.78	31.18	68.82	3.94
9	*N*	290	136	154	63	104	115	200	5
	%	100	46.90	53.10	21.72	35.86	39.66	68.97	1.72
10	N	1033	410	623	246	465	288	799	153
	%	100	39.69	60.31	23.81	45.01	27.88	77.35	14.81

Table 27: Distribution of Demographic Characteristics of Tested Accommodated Population, Mathematics EOC

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Algebra 1	*N*	394	169	225	101	157	124	252	16
	%	100	42.89	57.11	25.63	39.85	31.47	63.96	4.06
Geometry	*N*	310	163	147	63	120	116	217	8
	%	100	52.58	47.42	20.32	38.71	37.42	70.00	2.58

Table 28: Distribution of Demographic Characteristics of Tested Accommodated Population, Science and Social Studies DEI

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Biology 1	*N*	312	165	147	69	117	116	202	9
	%	100	52.88	47.12	22.12	37.50	37.18	64.74	2.88
Grade 5 Science	*N*	826	323	503	207	359	225	659	113
	%	100	39.10	60.90	25.06	43.46	27.24	79.78	13.68
Grade 8 Science	*N*	235	110	125	62	95	65	158	11
	%	100	46.81	53.19	26.38	40.43	27.66	67.23	4.68
U.S. History	*N*	243	125	118	57	92	86	148	5
	%	100	51.44	48.56	23.46	37.86	35.39	60.91	2.06
Civics	*N*	349	178	171	74	166	99	248	20
	%	100	51.00	49.00	21.20	47.56	28.37	71.06	5.73

Table 29: Distribution of Demographic Characteristics of Tested Accommodated Population, Science and Social Studies TTS

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Biology 1	*N*	17,087	6,767	10,320	3,849	6,110	6,231	14,048	571
	%	100	39.60	60.40	22.53	35.76	36.47	82.21	3.34
Grade 5 Science	*N*	32,874	13,360	19,514	7,740	12,345	10,930	26,048	4,154
	%	100	40.64	59.36	23.54	37.55	33.25	79.24	12.64
Grade 8 Science	*N*	25,649	9,922	15,727	6,372	9,584	8,371	21,066	1,486
	%	100	38.68	61.32	24.84	37.37	32.64	82.13	5.79
U.S. History	*N*	11,070	4,335	6,735	2,586	3,734	4,116	9,283	363
	%	100	39.16	60.84	23.36	33.73	37.18	83.86	3.28
Civics	*N*	25,436	10,014	15,422	5,932	9,520	8,536	20,657	1,611
	%	100	39.37	60.63	23.32	37.43	33.56	81.21	6.33

Table 30: Distribution of Demographic Characteristics of Tested Accommodated Population, B.E.S.T. Writing

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
4	*N*	706	262	444	177	335	161	595	134
	%	100	37.11	62.89	25.07	47.45	22.80	84.28	18.98
5	*N*	736	288	448	179	345	178	599	106
	%	100	39.13	60.87	24.32	46.88	24.18	81.39	14.40
6	*N*	441	173	268	110	211	100	358	45
	%	100	39.23	60.77	24.94	47.85	22.68	81.18	10.20
7	*N*	348	174	174	72	167	98	249	22
	%	100	50.00	50.00	20.69	47.99	28.16	71.55	6.32
8	*N*	237	106	131	61	98	65	166	11
	%	100	44.73	55.27	25.74	41.35	27.43	70.04	4.64
9	*N*	240	124	116	53	95	86	179	3
	%	100	51.67	48.33	22.08	39.58	35.83	74.58	1.25
10	*N*	312	155	157	66	139	102	206	7
	%	100	49.68	50.32	21.15	44.55	32.69	66.03	2.24

3. ADAPTIVE TESTING ADVANTAGES, ALGORITHM, AND SIMULATION STUDIES OVERVIEW

Florida's statewide, standardized assessments transitioned from fixed form to adaptive testing for English Language Arts and mathematics in the 2022–2023 school year and for science and social studies in the 2023–2024 school year. This chapter presents a brief overview of the advantages of adaptive testing, the algorithm that forms the basis of adaptive testing, and simulation studies that inform implementation. Further details, including testing procedures and evaluations, are presented in Volume 2, *Test Development;* and Volume 4, *Evidence of Reliability and Validity,* Section 4, Validity.

3.1 ADAPTIVE TESTING ADVANTAGES

According to Birnbaum (1957, as cited in Baker & Kim, 2004), the item information function is defined as

$$I_i(\theta) = -E\left(\frac{\partial^2 \log P_i(\theta)}{\partial \theta^2}\right).$$

This is also the Fisher information, which extends to the overall log-likelihood of the pattern of responses given a set of items on a test form seen by a student. In particular, the log-likelihood breaks up as the sum of the logarithms of the item characteristic curves of the individual items $P_i(\theta)$:

$$\sum_{i \in I} I_i(\theta) = -\sum_{i \in I} E\left(\frac{\partial^2 \log P_i(\theta)}{\partial \theta^2}\right).$$

Therefore, a well-tailored test for a particular student *s* means having the individual items *i* on the test form *I* have large item information $I_i(\theta)$ for the ability θ of the student *s*. The validity of this equation rests on the fundamental assumption of the local independence of the items given ability in item response theory, which we evaluate using the Q3 statistic in Volume 4. In a fixed form, such as in accommodated forms, as part of form construction, items are selected to shape the overall test information function to provide better test reliability of the test in the portion of the ability scale where most students are scoring or at the achievement-level cuts—which sometimes match. However, the test cannot be tailored for everybody along the entire ability spectrum, which is the problem that adaptive testing solves.

Once this problem is solved, the same amount of information can be obtained with fewer items on the test. However, solving this problem in practice requires a suitable algorithm for controlling exposure, meeting test blueprints, and selecting items based on ability estimated on the fly. This is made especially challenging under the requirement of three test administrations under the same blueprint. Addressing this challenge requires the focused development of enough suitable items to equip the item bank.

3.2 DESCRIPTION OF THE ADAPTIVE ALGORITHM

The implementation details of the adaptive algorithm are endless, as various scenarios have been addressed over the many years this algorithm has been used in other states. For example, the initial student ability estimate, recycling algorithm, passage group constraints, and other factors affect the algorithm, and it is not the goal to elucidate everything here. Both content requirements are mostly expressed in minimum and maximum number requirements at the overall test level and at more specific reporting categories or even higher levels of specificity, and the estimated item information contribution is simultaneously evaluated for a set of items pre-filtered at each stage to first ensure that candidate items are among the best few for satisfying the content requirements. Therefore, the basic principle is to first select items that have maximum content value, prioritizing those categories furthest from meeting minimum requirements, and especially so as the test nears conclusion. Only those items whose number can be adjusted are further evaluated as to the item information as estimated in this volume. Therefore, blueprint considerations always take precedence over adaptiveness, and in the case of initial calibration of the item bank, the adaptive component may have to be turned off entirely to obtain a sample for calibration. The final choice of item is randomized.

3.3 EVALUATION OF SIMULATIONS

The simulation outcomes are evaluated by psychometricians at the Florida Department of Education (FDOE) and Cambium Assessment, Inc. (CAI). Bias, correlation of average item difficulty against ability (as a measure of adaptiveness), item exposure, and blueprint match are the main pillars of the analysis, and special care must be taken regarding item bank depth. If the number of times a student takes a test increases beyond the available number of items in the bank to meet blueprint, items must be recycled from previous administrations to meet test blueprint requirements, which can also affect the adaptiveness of the test. If items need to be reused from previous administrations (recycling feature on), then a multi-opportunity study is necessary to determine accurate results.

More in-depth descriptions, including testing procedures and evaluations, are presented in Volume 2, Section 4, Test Construction and in Volume 4, Section 4, Validity.

4. ITEM BANK MAINTENANCE

This chapter describes the item bank in terms of review of operational and field-test items in spring 2025.

4.1 OVERVIEW OF ITEM DEVELOPMENT

Complete details of the item development plan for Cambium Assessment, Inc. (CAI) are provided in Volume 2, *Test Development,* of this technical report. The test development phase includes a variety of activities designed to produce high-quality assessments that accurately measure student skills and abilities according to the academic standards and blueprints.

New items are developed each year to be field-tested and added to the operational item pool. Several factors determine the development of new items. The item development team conducts a gap analysis for distributions of items across multiple dimensions, such as item counts, item types, item difficulty, Depth of Knowledge (DOK) levels, and numbers in each strand or benchmark.

Summative online assessments contain operational items and embedded field-test (EFT) items that are randomly distributed throughout each test. Operational items are used to calculate student scores. EFT items are non-scored items and are used to populate the bank for future operational use.

The accommodated versions of online assessments can contain filler items in the field-test slots to ensure equal-length assessments. These items are not analyzed as part of field-test calibrations.

4.2 REVIEW OF OPERATIONAL ITEMS

During the operational calibration, both operational and field-test items were reviewed based on their performance during the spring administration. Before the spring administration, *Calibration* and *Scoring Specifications* documents are created by CAI, Florida Department of Education (FDOE), and Human Resources Research Organization (HumRRO). The original versions on which subsequent years' updates are based were reviewed by the Technical Advisory Committee (TAC). The specifications documents outline all details of item calibration, flagging rules for items, equating to the item response theory (IRT)-calibrated item pool, pre-equating of accommodated forms, and scoring. CAI uses the specifications to complete classical item analyses and IRT calibrations (see Section 5, Item Analyses Overview, and Section 6, Item Calibration and Scaling, of this volume of the technical report) for each test and posts results to a secure location for review. Items are reviewed, with special attention given to items that are flagged based on statistical rules described in the *Calibration* document. These flagging rules are outlined in the following sections. Psychometricians and content experts work together to review items and their statistics and determine whether any items should be removed from scoring.

4.3 FIELD TESTING

The bank item pool grows each year through new item field testing. Any item used on an assessment is field-tested before it is used as an operational item.

Embedded Field Test

Approximately three to seven field-test items are assigned to students randomly. Tables 31–33 provide a brief description of each item type.

Table 31: ELA Reading Item Types and Descriptions

Response Type	Description
Multiple-choice (MC)	Student selects one correct answer from several options.
Multiple-select (MS)	Student selects all correct answers from several options.
Table match (MI)	Student checks a box to indicate if information from a column header matches information from a row. On accommodated forms, the student fills in a bubble to indicate if information from a column header matches information from a row.
Hot-text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. On accommodated forms, the student fills in bubbles to indicate which sentences are correct.
MC HT selectable (two-part HT)	Student selects the correct answers from Part A and Part B. Part A is an MC or an MS item, and Part B is a selectable HT item.
External copy interaction (ECI)	Student selects information directly from the passage to support an analysis or make an inference.
Evidence-based selected-response (EBSR)	Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A.

Table 32: Mathematics and Mathematics EOC Item Types and Descriptions

Response Type	Description
Equation (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. On accommodated forms, the student uses an empty response box to write in an answer.
Edit task inline choice (ETIC)	Student chooses the replacement for an incorrect number, word, phrase, or blank from several options. This includes items with one or more ETIC interactions. On accommodated forms, the student fills in a bubble to indicate the correct number, word, or phrase that should fill in the blank.
Graphing	Student creates a graph or number line. The student can create a bar graph, line plot, or histogram by clicking parts of a display. The student can plot a point, create a graph of an equation, draw a shape, or construct other responses by clicking a grid or a coordinate grid. The student can plot a point on a number line by clicking the number line or graph an inequality by clicking the number line and choosing an arrow. This item type is not used on accommodated forms.
Grid (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot-text (HT)	Student is directed to select text to support an analysis or make an inference. On accommodated forms, the student fills in bubbles to indicate which sentences are correct.
Multiple-choice (MC)	Student selects one correct answer from four options.

Response Type	Description
Table match (MI)	Student checks a box to indicate if information from a column header matches information from a row. On accommodated forms, the student is directed to fill in a bubble that matches a correct option from a column with a correct option from a row.
Multiple-select (MS)	Student selects all correct answers from several options.
Multi-interaction (MULTI)	This is an item that contains more than one response type. It could contain more than one of the same interaction type (except for multiple combinations of ETIC) or a combination of interaction types.

Table 33: Science and Social Studies Item Types and Descriptions

Response Type	Description
Multiple-choice (MC)	Student selects one correct answer from four options.

Table 34 shows the number of mathematics and mathematics End-of-Course (EOC) items by grade and item type that are included in spring for field testing. Table 35 shows the number of reading items by grade and item type that were included in spring for field testing. Table 36 shows the number of science and social studies items by grade and item type that were included in spring for field testing.

During calibrations, some items were dropped from the initial item pool due to poor performance. Appendix B, Field-Test Item Statistics, provides the number of field-test items remaining after removing, during calibrations, items with poor performance.

Table 34: Mathematics and Mathematics EOC Field-Test Items by Item Type and Grade

Item Type	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra 1	Geometry
EQ	51	69	79	71	79	65	64	74
ETIC	15	12	6	9	16	20	13	25
GI	0	1	0	0	0	1	0	0
Graphing	1	0	2	0	3	3	0	0
HT	0	0	0	0	0	1	1	0
MC	33	34	44	52	64	80	66	51
MI	6	8	5	0	2	4	2	2
MS	29	18	7	9	6	10	11	9
MULTI	12	7	9	9	10	14	15	8
Total number of items	147	149	152	150	180	198	172	169

Table 35: ELA Reading Field-Test Items by Item Type and Grade

Item Type	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
EBSR	23	16	9	20	12	20	22	20
HT	6	8	6	4	10	23	14	14
MC	124	100	65	157	128	159	130	98

Item Type	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
MI	13	10	9	17	19	13	10	13
MS	14	16	11	16	11	8	16	12
Two-part HT	0	0	1	0	0	0	0	0
ECI	0	0	0	1	1	0	0	1
Total number of items	180	150	101	215	181	223	192	158

Table 36: Science and Social Studies Field-Test Items by Item Type and Grade

Item Type	Grade 5	Grade 8	Biology 1	U.S. History	Civics
MC	190	177	232	300	311
Total number of items	190	177	232	300	311

Volume 2, *Test Development,* provides a detailed overview of the development and review process for new items and additional details on the development and maintenance of the item pool.

Writing Independent Field Test

In 2023, writing was administered as an independent field test (IFT) to a sample of Florida students. Results from this field test will be used for operational writing administrations moving forward, including in 2024.

A scientific sampling design was used to identify and select the sample students for the IFT. A stratified random sample of intact schools participated, one representative of the population and testing conditions, and the writing sample selected represented the state population with respect to ethnicity and gender distribution. Each prompt was administered randomly and only to the students from the sample schools. The students' responses were then handscored by two scorers based on the B.E.S.T. rubric. Table 37 shows the number of prompts that were field-tested and the total number of students.

Table 37: Number of Prompts and Sample Size

Grades	Number of Prompts	Sample Size per Prompt	Total Expected Sample Size	Final Calibration Sample Size
4	10	5,000 (+10%)	55,000	49,431
5	10		55,000	50,408
6	11		60,500	58,448
7	10		55,000	49,883
8	10		55,000	50,089
9	12		66,000	59,107
10	16		88,000	74,794

The generalized selection methods are described as follows:

Let $k_{(j)g}$ denote the number of students in grade g in the jth school $j = \{1, 2, \ldots N_g\}$, and $K_g = \sum_{J=1}^{N_g} k_{(j)g}$ is the total number of students in grade g across all schools. N_g is the total number of eligible schools in grade g. CAI proposed the writing sample size for each grade (see Table 1).

Let the total sample size for grade g be t_g. Hence, assuming a typical sample size of students in each school at grade *g,*

$$\bar{k}_g = \frac{K_g}{N_g},$$

the total number of schools required for sampling was obtained:

$$M_g = \frac{t_g}{\bar{k}_g}.$$

Rather than making an arbitrary assumption regarding the value of $\bar{k}_g$, CAI derived the value for each grade from the data provided in the State Student Results (SSR) files.

Stratified Sampling

To use a proportionate stratification method, CAI used the number of students, $l_{n,g}$, to first identify the proportion of schools across the state within stratum l:

$$P_{l,g} = \frac{l_{n,g}}{K_g}.$$

Schools were then sampled within each stratum:

$$m_{l,g} = P_{l,g} M_g.$$

The sampling method used an explicit stratum as well as implicit strata. The implicit strata were binned as quintiles. Within each explicit stratum, schools will be sorted in a serpentine order (alternating ascending and descending) by the implicit strata, and $m_{l,g}$ schools were selected systematically from this sorted list.

In hierarchical serpentine sorting, within a stratum, the first variable was sorted in ascending order. Then, within the first level of the first variable, the second variable was sorted in ascending order. Within the second level of the first variable, the second variable was sorted in descending order. This procedure continued for all levels and all variables so that it was equivalent to alternate ascending and descending order by each variable.

To yield a representative sample of students from the testing population, the sampling strata must identify and capture the most important characteristics of the state population. For this reason, the strata outlined in the following list were used.

Explicit Strata

- **Region:** The state was divided into various geographic regions. This variable is intended to capture the differences in student populations across the state.

Implicit Strata

- **Percent proficient in the school on the prior year's reading test:** This variable is intended to capture the ability of students across the population.

- **School size:** This variable is intended to ensure that the sample represents schools of various sizes.
- **Curriculum group:** Standard, English language learner [ELL], exceptional student education. [ESE])
- **Gender:** Male and female.
- **Percent ethnicity:** White, African American, Hispanic, and Other.

Post hoc analysis was performed to evaluate the representativeness of the sample and submitted for approval. *N* counts within each region, mean scaled scores, and proportion of demographic groups listed in the implicit strata above were matched between the sample schools and the target.

5. ITEM ANALYSES OVERVIEW

This chapter summarizes the classical item analyses and differential item functioning (DIF) analyses. Classical and item response theory (IRT) statistics were derived during the spring administration, after students had gone through a year's worth of instruction and had the opportunity to learn.

5.1 CLASSICAL ITEM ANALYSES

Item analyses examine whether test items function as intended. Overall, classical item analysis and IRT analysis require a minimum sample of 1,500 responses (Kolen & Brennan, 2014) per item. In fact, many more than 1,500 responses are always available. Similarly, a minimum sample of 200 responses (Zwick, 2012) per item in each subgroup is applied for DIF analyses.

Several item statistics are used to evaluate multiple-choice (MC) and non-multiple-choice (non-MC) items, generally referred to as constructed-response (CR) items, for integrity and appropriateness of the items' statistical characteristics. Table 38 presents the thresholds used to flag an item for further review based on classical item statistics.

Table 38: Thresholds for Flagging Items in Classical Item Analysis

Rule	Flagging Criteria	Rationale
p-value	For 1-point items, flag if $p < 0.20$ or $p > 0.90$	Item is too difficult and *p*-value is less than expected from random chance, or item is too easy for population.
Relative mean	For polytomous items, flag if the relative mean is < 0.15 or > 0.95	Item is too difficult or too easy.
Correlation with test for a key	Flag if < 0.25	Item is non-discriminating.
Distractor *p*-value	Flag if the *p*-value for the distractor is larger than the *p*-value for the key	Item is potentially problematic.
Correlation with test for distractors	Flag if correlation for any distractor is larger than correlation for the key	Distractor is more discriminating than the keyed response.
DIF	Flag if DIF statistics fall into the C category for any group	Item shows evidence of significant DIF.

Item Discrimination

The item discrimination index indicates the extent to which each item is differentiated between the test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item can differentiate between high- and low-achieving students. The discrimination index for MC items is calculated as the correlation between the item score and the IRT theta ability estimate for students. Point-biserial or point-polyserial correlations for

operational items are presented in Appendix A, Operational Item Statistics, of this volume of the technical report.

Distractor Analysis

Distractor analysis for MC items is used to identify items that may have had marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracted high-scoring students. For MC items, the correct response should have been the option most frequently selected by high-scoring students. The discrimination value of the correct response should have been substantial and positive, and the discrimination values for distractors should have been lower and generally negative.

Item Difficulty

Extremely difficult or extremely easy items are flagged for review but are not necessarily deleted if they are grade-level appropriate and aligned with the test specifications. For MC items, the proportion of students in the sample selecting the correct answer (the *p*-value) is computed in addition to the proportion of students selecting incorrect responses. For CR items, item difficulty is calculated using the item's relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the maximum possible score). Section 6.6, Results of Calibrations Including Field-Test Items, of this volume summarizes the conventional item *p*-values and IRT parameters. Appendix A of this volume presents the *p*-values for operational items.

5.2 Differential Item Functioning Analysis

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014) provides a guideline to determine when sample sizes permitting subgroup differences in performance should be examined and when appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors. To identify such potential problems, items were evaluated in terms of DIF statistics.

DIF analysis was conducted for all items to detect potential item bias across major gender, ethnic, and special population groups. Because of the limited number of students in some groups, DIF analyses were performed for the following groups:

- Male/female
- White/African American
- White/Hispanic
- Student with disabilities (SWD)/not SWD
- English language learner (ELL)/not ELL

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF was important because it provided a statistical indicator that an item may contain cultural or other biases. DIF-flagged items were further

examined by content experts who were asked to re-examine each flagged item to decide whether the item should have been excluded from the item pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in certain areas are less likely to offer rigorous geometry classes, students at those schools might perform more poorly on geometry items than would be expected given their proficiency on other types of items. In this example, the instruction, not the item, exhibits bias. However, DIF can indicate bias, so all items were evaluated for DIF.

A generalized Mantel–Haenszel (MH) procedure was applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's IRT theta ability estimate on the operational items on a given test is used as the ability-matching variable. DIF analyses were performed on field-test items using IRT ability estimates as the ability-matching variable during field-test calibrations. The corresponding scores were divided into 10 intervals to compute the MHχ^2 DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})},$$

where $k = \{1, 2, \ldots K\}$ for the strata, $n_{R1k}$is the number of correct responses for the reference group in stratum k, and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k} n_{R+k}}{n_{++k}},$$

where n_{+1k} is the total number of correct responses, $n_{R+k}$is the number of students in the reference group, and n_{++k} is the number of students in stratum k, and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k} n_{F+k} n_{+1k} n_{+0k}}{n_{++k}^2 (n_{++k} - 1)},$$

n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses in stratum k.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k} n_{F0k} / n_{++k}}{\sum_k n_{R0k} n_{F1k} / n_{++k}}.$$

The MH-delta (Δ_{MH}, Holland & Thayer, 1988) is then defined as

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH}).$$

The *GMH* statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)' \left(\sum_k var(\boldsymbol{a}_k)\right)^{-1} \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right),$$

where $\boldsymbol{a}_k$ is a $(T-1)\,X\,1$ vector of item response scores, corresponding to the T response categories of a polytomous item (excluding one response). $E(\boldsymbol{a}_k)$ and $var(\boldsymbol{a}_k)$, a $(T-1)\times(T-1)$ variance matrix, are calculated analogously to the corresponding elements in $MH\chi^2$, in stratum k.

The SMD (Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK} m_{FK} - \sum_k p_{FK} m_{RK},$$

where

$$p_{FK} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum k,

$$m_{FK} = \frac{1}{n_{F+k}}\left(\sum_t a_t n_{Ftk}\right)$$

is the mean item score for the focal group in stratum k, and

$$m_{RK} = \frac{1}{n_{R+k}}\left(\sum_t a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum k.

Items are classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. DIF classification rules are illustrated in Table 39. Items were also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African American, Hispanic, female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., White, male). If the DIF statistics fell into the "C" category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged on the basis of DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance in field testing.

Table 39: DIF Classification Rules

Dichotomous Items	
Category	*Rule*
C	MH_{X^2} is significant, and $\left\|\hat{\Delta}_{MH}\right\| \geq 1.5$.
B	MH_{X^2} is significant, and $1 \leq \left\|\hat{\Delta}_{MH}\right\| < 1.5$.
A	MH_{X^2} is not significant, or $\left\|\hat{\Delta}_{MH}\right\| < 1$.
Polytomous Items	
Category	*Rule*
C	GMH_{X^2} is significant, and $\|SMD\| / \|SD\| > .25$.
B	GMH_{X^2} is significant, and $.17 < \|SMD\| / \|SD\| \leq .25$.
A	GMH_{X^2} is not significant, or $\|SMD\| / \|SD\| \leq .17$.

DIF summary tables are presented in Appendix A, Operational Item Statistics, for operational items, and Appendix B, Field-Test Item Statistics, for field-test items. Across all tested grades and DIF comparison groups, less than 1% were classified as C DIF for operational items. Content specialists and psychometricians reviewed items to ensure that they were free of bias.

In addition to the classical item summaries described in this section, IRT-based statistical summaries (i.e., item fit and item fit plots) were used during item review. These methods are described in Section 6.5, IRT Item Summaries.

6. ITEM CALIBRATION AND SCALING

Item response theory (IRT) was used to calibrate all items and derive scores for all FAST, B.E.S.T., Science, and Social Studies assessments. IRT is a general framework that models test responses that result from interactions between students and test items. One advantage of IRT models is that they allow for item difficulty to be scaled on the same metric as test-taker ability.

IRT encompasses many related measurement models. Models can be grouped into two families. While both families include models for dichotomous and polytomous items, they differ in their assumptions about how student ability interacts with items. The Rasch family of models includes the Rasch model and Masters' partial credit model. The Rasch family is distinguished in that the models do not incorporate a pseudo-guessing parameter, and they assume that all items have the same discrimination.

Extensions to the Rasch model include the two-parameter logistic (2PL) and three-parameter logistic (3PL) models and the generalized partial credit model (GPCM). These models differ from the Rasch family of models by including a parameter that accounts for the varied slopes between items, and in some instances, models also include a lower asymptote that varies to account for pseudo-guessing that may occur with some items. A discrimination parameter is included in all models in this family and accounts for differences in the amount of information items may provide along different points of the ability scale (the varied slopes). The 3PL model is characterized by a lower asymptote, often referred to as a *pseudo-guessing parameter*, which represents the minimum expected probability of answering an item correctly. The 3PL model is often used with multiple-choice (MC) items, but it can be used with any item where there is a possibility of guessing. Therefore, content and psychometric teams perform additional reviews on all non-MC items to evaluate the possibility of guessing. If an item involves guessing, a more generalized version of the IRT model (e.g., 3PL) is selected to account for pseudo-guessing.

Two general approaches, pre-equating and post-equating, are used in IRT to calibrate items and score students based on the estimated item parameters. The difference in these two types depends on when the equating practice is conducted. Pre-equating occurs before the operational testing, and post-equating happens after the operational testing. Both are extensively used in K–12 large-scale assessment programs (Tong et al., 2008). In pre-equating, the statistical characteristics of the items estimated from one representative student group are applied to score all future groups of students by relying on the IRT assumption of parameter invariance. Pre-equating has been adopted in large-scale assessments for various practical and policy reasons. The advantages of pre-equating include rapid score reporting, more time for quality control, and more flexibility in the assessment (Tong et al., 2008). In post-equating, the statistical characteristics of the items are estimated by using the post-administration data and are assumed to apply only to this student group. Therefore, the statistics of the items are sometimes considered more accurate than those in pre-equating (Tong et al., 2008). New item statistics are collected each year when items are used, thus assuming that the statistical characteristics of the item may change when the ability of the tested population changes.

In prior years, Florida used the pre-equated method for retake administrations and the post-equating method for non-retake administrations. Beginning with the 2023 spring administration for English Language Arts (ELA) and mathematics and the 2024 spring administration for science

and social studies, due to the transition to computer-adaptive testing (CAT), the pre-equating method became necessary for all tests moving forward.

6.1 ITEM RESPONSE THEORY METHODS

The generalized approach to item calibration was to use the 3PL model (Lord & Novick, 1968) for MC items, use the 2PL model (Lord & Novick, 1968) for binary items that assume no guessing, and use the GPCM (Muraki, 1992) for items scored in multiple categories.

For items with some probability of guessing, such as MC items, the 3PL model was used because it incorporates a parameter to account for guessing. For non-MC binary items, item content was reviewed. If it was determined that there was no probability of guessing, the 2PL model was used; however, the 3PL model was used if guessing was in fact possible.

The 3PL model is typically expressed as

$$\mathrm{P_i}(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta_j - b_i)\right]},$$

where $\mathrm{P_i}(\theta_j)$ is the probability of test taker j answering item i correctly; c_i is the lower asymptote of the item response curve (the pseudo-guessing parameter); b_i is the location parameter; a_i is the slope parameter (the discrimination parameter); and D is a constant fixed at 1.7, bringing the logistic into coincidence with the probit model. Student ability is represented by θ_j. For the 2PL model, the pseudo-guessing parameter (c_i) is set to 0.

The GPCM is typically expressed as the probability for individual j of scoring in the $(z_i + 1)$th category to the *i*th item as

$$\mathrm{P}(z_i|\theta_j) = \frac{\exp\sum_{k=0}^{z_i} Da_i(\theta_j - \delta_{ki})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h} Da_i(\theta_j - \delta_{ki})},$$

where δ_{ki} is the *k*th step value; $z_i = 0,1,..,m_i$, m_i is the maximum possible score of the item; and $\sum_{k=0}^{0} Da_i(\theta_j - \delta_{ki}) = 0$.

All item parameter estimates were obtained with IRTPRO version 5.0 (Cai, Thissen, & du Toit, 2011). IRTPRO employed the marginal maximum likelihood estimation (MMLE) procedure to estimate item parameters.

6.2 ELA AND MATHEMATICS–ESTABLISHING A NEW SCALE

6.2.1 On-Grade Calibrations ELA and Mathematics

Reading and Mathematics

In 2023, a new score scale was established to replace the Florida Standards Assessment (FSA) scale for ELA reading, mathematics, Algebra 1, and geometry to reflect the implementation of the new assessments measuring Florida's B.E.S.T. On-grade calibrations were completed first to establish a new base IRT scale for FAST and B.E.S.T., followed by vertical linking calibrations and calibration of the field-tested writing items.

Initially, Cambium Assessment, Inc.'s (CAI) proposed calibration of the new FAST assessments called for an operational field-test design employing the entire pool of FAST items. In this design, the item selection algorithm is guided only by blueprint weights, ensuring that each test administration meets all blueprint specifications. The adaptive weights, however, are set to zero, so that item selection is independent of item difficulty and student performance. This approach results in a linking design in which all progress monitoring (PM) bank items are linked to all other bank items, and the sample of responses to each item is a random and representative sample of Florida students. In this approach, all bank items would be calibrated concurrently, with the item parameters effectively modeling the full breadth and depth of the measurement model assessed in the FAST assessments. Over time, the overall project plan evolved. The Florida Department of Education (FDOE), in consultation with CAI, committed to immediate scoring and reporting of summative assessment results in spring 2023 based on the existing FSA reporting scale and performance standards. Consequently, the original calibration plan was untenable. Moreover, FDOE preferred to continue adaptive test administration for the summative test administrations, so the revised calibration plan sought to preserve that approach where possible.

Two approaches for establishing the new FAST scales were chosen: one approach for calibrating the new mathematics assessments that was based on the administration of discrete items, and a second approach for ELA that required the administration of passage sets (or item groups) where, when the item group is selected, all items associated with that item group are administered. Both approaches provided immediate scoring and reporting of summative test results on the current FSA scale and performance-level classifications. The approach for mathematics also allowed for continued adaptive test administration of the summative test items, using field-test items administered in embedded field-test (EFT) slots for calibration of the new FAST scale. Because ELA is passage-based, and students are administered only a single passage set or item group in the summative assessment, there is no possibility of linking bank items in the context of the EFT design. For ELA, it was therefore necessary to administer the summative test items as an operational field test, with each test administration meeting all blueprint specifications, but with item selection being independent of item difficulty. As noted in this report, FDOE prefers to maintain adaptive test administration of summative test items where possible. Because the mathematics item pools are made up of discrete items (e.g., items that are not bound to a common stimulus), it was possible to establish the new FAST scale (as well as the high school end-of-course [EOC] tests) using field-test items administered in the EFT slots of the summative test administration. The newly developed items, administered in the EFT slots in the summative assessment, were freely calibrated to construct the new FAST scale. To ensure robust linkages between the items administered in the EFT slots, the plan called for 10 EFT slots per test administration.

Ten EFT slots allowed for each item in the mathematics pool to be paired with every other item in the pool across hundreds of test administrations to ensure a strong linkage between items in the FAST mathematics pool. In addition, the newly developed FAST mathematics items could be linked to the FSA scale by anchoring the summative test items to their FSA bank parameters and then calibrating the field-test item parameters under that constraint.

This procedure resulted in the field-test items having two sets of item parameters, one on the new FAST scale and a second on the current FSA scale, allowing FDOE to establish a linkage between the FSA and FAST scales. These linking constants were then applied to the FSA item parameters for items in the current summative pool to place those item parameters on the new FAST scale as

well. Although indirect, this approach to equating the summative test items to the FAST scale provided a mechanism for deploying the full FAST item pool in the 2023–2024 school year. To provide a check on the quality of the linked item parameters, a sample of the current summative items could be field-tested again in spring 2024 to evaluate whether there is evidence of systematic item drift for indirectly linked item parameters.

In the context of ELA, each student was administered field-test items bound to a common stimulus, in this case a passage set. Because students were administered items from only a single field-test passage set, there was no possibility of linking ELA items in the context of the EFT design. Calibrating the ELA pool to the new FAST scale required an operational field-test design. In this approach, the current FAST ELA pool was configured to be administered as an operational field test. Item selection was configured to achieve a blueprint match for each test administration, but item selection proceeded independently of item difficulty. Therefore, each passage set, and thus each item in the current summative pool, was administered to a random and representative sample of Florida students, supporting calibration of items to the new PM scale.

Since all summative items were already calibrated on the FSA scale, the test administrations supported immediate scoring and reporting of assessment results on the FSA scale and performance-level classification. In addition, the newly developed FAST passage sets and items were randomly selected for administration in the EFT slots in the summative test administration. This resulted in a random and representative sample of student responses to each item. In this approach, all FAST items, including summative and field-test items, could be concurrently calibrated. This placed all ELA items on the new FAST scale with item parameters that (1) robustly model the breadth and depth of the measurement model FAST assesses and (2) were consistent with the originally proposed approach. This approach supported robust and adaptive test administration of the three-opportunity PM assessments. This approach also supported the calibration of the new FAST writing prompts, since writing items must be linked by anchoring summative test item parameters to their FAST bank values and calibrating the writing items under that constraint.

Before the on-grade calibration, classical item statistics were reviewed. The following items were dropped: items not certified from Rubric Evaluation and Verification for Items Scored Electronically (REVISE), items missing score categories, and items with a negative biserial or sample size of less than one thousand. During calibrations, priors were put on b-parameters for any item with convergence issues or the number of iterations increased. Items with negative a-parameters and/or b-parameters larger than 10 were dropped and the calibrations were re-run. The standard error (SE) for the b-parameter larger than 1.0 was also considered. If these SEs were equal to or larger than the b-parameter, priors on the b-parameter were also added if they improved estimates.

Writing

Summative reading items from the on-grade calibrations were calibrated concurrently with the writing prompts. FAST parameters were used as anchors for the calibration of the writing prompts for each dimension (convention, elaboration, and organization). Each dimension was calibrated separately due to the high local dependence between the dimensions. For the spring 2024 administration onwards, B.E.S.T. Writing was reported as a raw score test and these parameters were not used.

6.2.2 Vertical Linking ELA and Mathematics

Vertical linking places test scores from different grade levels on the same measurement scale to track the growth of individual students and groups of students. To establish a new vertical scale for the FAST tests, Grades 3–8 mathematics were linked on a vertical scale. Grades 3–10 reading were also placed on a vertical scale. In addition, the Grade 2 reading and mathematics tests were linked to the FAST vertical scale.

During the spring 2023 administration, linking items from the upper grades and the lower grades were placed onto the on-grade forms. This enabled the forward-linking, backward-linking, and mixed-linking methods. In the mixed-linking method, both the forward- and backward-linking methods were combined to create a vertical scale. Items measuring content from below and above grade were placed onto the on-grade forms. The goal was to administer a linking set that represented the content of the tests from which the items were derived. For example, the Grade 4 items placed onto the Grade 3 test were intended to represent the Grade 4 test blueprint. This design supports the inference that the scaled score from the vertical scale represents both the on-grade performance and the location of a student's performance on the upper-grade test.

A chain-linking approach was used to link the grade-level assessments in each subject area. Following the anchored calibrations, each vertical linking item has two sets of item parameters. One set consists of the on-grade parameters, and the other consists of the off-grade parameters. Grade 3 was used as the base (or anchor) grade for the vertical linking.

The vertical linking calibration used on-grade summative items and vertical linking items from both the lower and the upper grades. No field-test items were included. All items dropped from the previous on-grade calibration steps were excluded. For the off-grade vertical linking items, items were dropped after examination of the criteria outlined in Table 40 from the grades in which they were flagged. In contrast with ELA, mathematics summative items were not flagged because they were administered adaptively. Summative and vertical linking items were concurrently calibrated by fixing the summative items on their on-grade FAST scale parameters. Items with convergence issues were dropped, and the other items were re-calibrated. The *A* and *B* linking constants were obtained using the Stocking–Lord method (Stocking & Lord, 1983) for adjacent grades for the mixed-, forward-, and backward-linking methods (with the lower grade always serving as the reference form).

Stocking–Lord Method

The Stocking–Lord method (Stocking & Lord, 1983) is commonly used alongside the 3PL model and the GPCM and finds the linking constants (*A* and *B*) that minimize the squared distance between two test characteristic curves. *A* is often referred to as the *slope,* and *B* is often referred to as the *intercept*. The approach evaluates the following integral, where the indices *I* denote a common item, and *a* and *b* denote separate forms:

$$SL = \int \left[\sum_{i=1}^{I} p(\theta; a_{ia}, b_{ia}, c_{ia}) - \sum_{i=1}^{I} p(\theta; \frac{a_{ib}}{A}, Ab_{ib} + B, c_{ib}) \right]^2 f(\theta) d(\theta)$$

Calculating the D2 Statistic

After performing the Stocking–Lord method (Stocking & Lord, 1983), the equated parameters were compared by rescaling the items to be on the same scale. D^2, the sum of the squared differences between item characteristic curves (ICCs), was calculated. The D^2, or the mean square deviation (MSD), is computed by integrating out θ as follows:

$$D^2 = \int (ICC_{ai}(\theta) - ICC_{bi}(\theta))^2 f(\theta; \mu, \sigma^2) d\theta.$$

The integral does not have a closed-form solution, and so its approximation is based on the weighted summation over j={1, 2, …, 30} quadrature points, all taken from equally spaced points interior to the normal density, w, between –4 and 4 of the marginal distribution.

$$D^2 = \sum_{j=1}^{30} w_j (ICC_{ai}(\theta_j) - ICC_{bi}(\theta_j))^2$$

D^2 was calculated, and ICCs were plotted. Items with D^2 values more than three standard deviations were flagged for review, as they excessively impact the scale transformation constants.

Table 40: Flagging Criteria for Vertical Linking Items

Rule	Flagging Criteria	Rationale
p-value	For MC items, flag if $p < 0.25$ or $p > 0.95$	Item is too difficult and *p*-value is less than expected from random chance, or item is too easy for population.
Relative mean	For polytomous items, flag if the relative mean is < 0.15 or > 0.95	Item is too difficult or too easy.
Biserial/polyserial	Flag if < 0.15	Item is low-discriminating.
Distractor *p*-value	Flag if the *p*-value for the distractor is larger than the *p*-value for the key	Item is potentially problematic.
Distractor biserial	Flag if the biserial for any distractor is larger than the biserial for the key	Distractor is more discriminating than the key.
Convergence issues	Flag the IRT statistics if IRTPRO does not converge	The number of iterations and convergence should be noted in a table.
D2 and ICCs	Flag if D2 is greater than three standard deviations	Difference between grades is too large.

Final Linking Set

After inspection of the preliminary *A* and *B* constants from the forward-, backward-, and mixed-linking methods, the mixed-linking set was chosen for further evaluation. For ELA, items were further dropped based on Q1, *p*-value reversal between grades, D^2, adequate blueprint representation, and coherent articulation (differences in scores) between grades to achieve a smooth, final solution.

For mathematics, this procedure was not suitable because it resulted in inadequate blueprint proportions and incoherent articulation between grades. Instead, items were dropped based on the a-parameter ratio between grades being too big or too small, reversal of *p*-values and b-parameter between grades, adequate blueprint representation, and coherent articulation between grades to achieve a smooth, final solution. The a-parameter was evaluated based on the consideration that items used in linking should be stable across the grades. The discrimination parameter ratio should be close to 1 if the linking slope is near 1. If the ratio is too far away from 1, the item parameter can be judged as being too unstable, and the item can be tagged as a candidate for removal. The cuts of 0.6 to 1.4 were used. Evaluation of the items was performed iteratively by checking the blueprint at the reporting category level and the removal of the most unstable candidate items first, then checking the blueprint again, and then adding back any necessary items.

In addition to this, for Grades 7 and 8 mathematics, anchor calibrations were re-run with all items (including those previously dropped due to the criteria in Table 40). Items were instead dropped based on the a-parameter ratio between grades being too big or too small, reversal of *p*-values and b-parameters between grades, adequate blueprint representation, and coherent articulation between grades. Table 41 lists the number of items remaining in the final vertical linking set for each ELA reading and mathematics grade combination.

Appendix G, Vertical Linking Grades 3–10 Blueprint Match, presents the results of the initial blueprint violations and final blueprint match.

Table 41: Number of Items Administered, Removed, and Remaining in the Final Vertical Linking Sets

Subject	Grade	Vertical Linking Items Administered	Number of Vertical Linking Items Removed	Final Vertical Linking Set
ELA Reading	4 to 3	77	20	57
	5 to 4	76	25	51
	6 to 5	73	49	24
	7 to 6	70	22	48
	8 to 7	75	32	43
	9 to 8	77	57	20
	10 to 9	78	50	28
Mathematics	4 to 3	70	36	34
	5 to 4	70	21	49
	6 to 5	74	12	62
	7 to 6	72	48	24
	8 to 7	72	31	41

The final vertical linking constants for ELA reading and mathematics are shown in Tables 42 and 43, respectively.

Table 42: Final Vertical Linking Constants for ELA Reading

Grade	Slope	Intercept
3	1.00000	0.00000
4	0.96223	0.60245
5	0.99412	0.98565
6	1.02819	1.12642
7	1.05743	1.41558
8	1.09508	1.72445
9	1.07704	1.92753
10	1.07324	2.15999

Table 43: Final Vertical Linking Constants for Mathematics

Grade	Slope	Intercept
3	1.00000	0.00000
4	0.98467	0.69312
5	1.05306	1.08148
6	0.99186	1.36995
7	0.94724	1.57334
8	0.89911	1.86851

Tables 44 and 45 show descriptive statistics for ELA reading and mathematics across grades on the vertical scale with mean ability. To evaluate the properties of the vertical linking scale for ELA reading and mathematics, the mean ability (theta), growth, and articulation between grades on the vertical scale were examined. Figures 1 and 2 show the separation between the grades at different thetas for ELA reading and mathematics, respectively. The growth and separation are in an acceptable range and direction. The results of the vertical linking appear to be similar to those developed in 2010 and 2015 (see *Florida Statewide Assessments 2014–2015 Technical Report*).

Table 44: Descriptive Statistics for ELA Reading on the Vertical Scale

Grade	*N*	Theta Mean	Theta Standard Deviation	Growth	Effect Size
3	220,125	–0.05729	1.17790		
4	199,860	0.57831	1.07993	0.63560	0.58856
5	206,230	0.97628	1.09024	0.39796	0.36502
6	215,473	1.10367	1.14690	0.12740	0.11108
7	208,172	1.38930	1.18652	0.28562	0.24072
8	213,915	1.69449	1.23179	0.30519	0.24776
9	220,852	1.88886	1.22318	0.19437	0.15891
10	210,980	2.13830	1.21280	0.24944	0.20567

Table 45: Descriptive Statistics for Mathematics on the Vertical Scale

Grade	*N*	Theta Mean	Theta Standard Deviation	Growth	Effect Size
3	219,589	–0.03776	1.10388		
4	196,520	0.67060	1.11232	0.70836	0.63683
5	201,951	1.03755	1.18343	0.36695	0.31007
6	206,192	1.31391	1.12175	0.27636	0.24637
7	146,439	1.44090	1.19483	0.12698	0.10628
8	124,497	1.72319	1.18137	0.28229	0.23895

Figure 1: ELA Reading Trend Lines for Final Solution

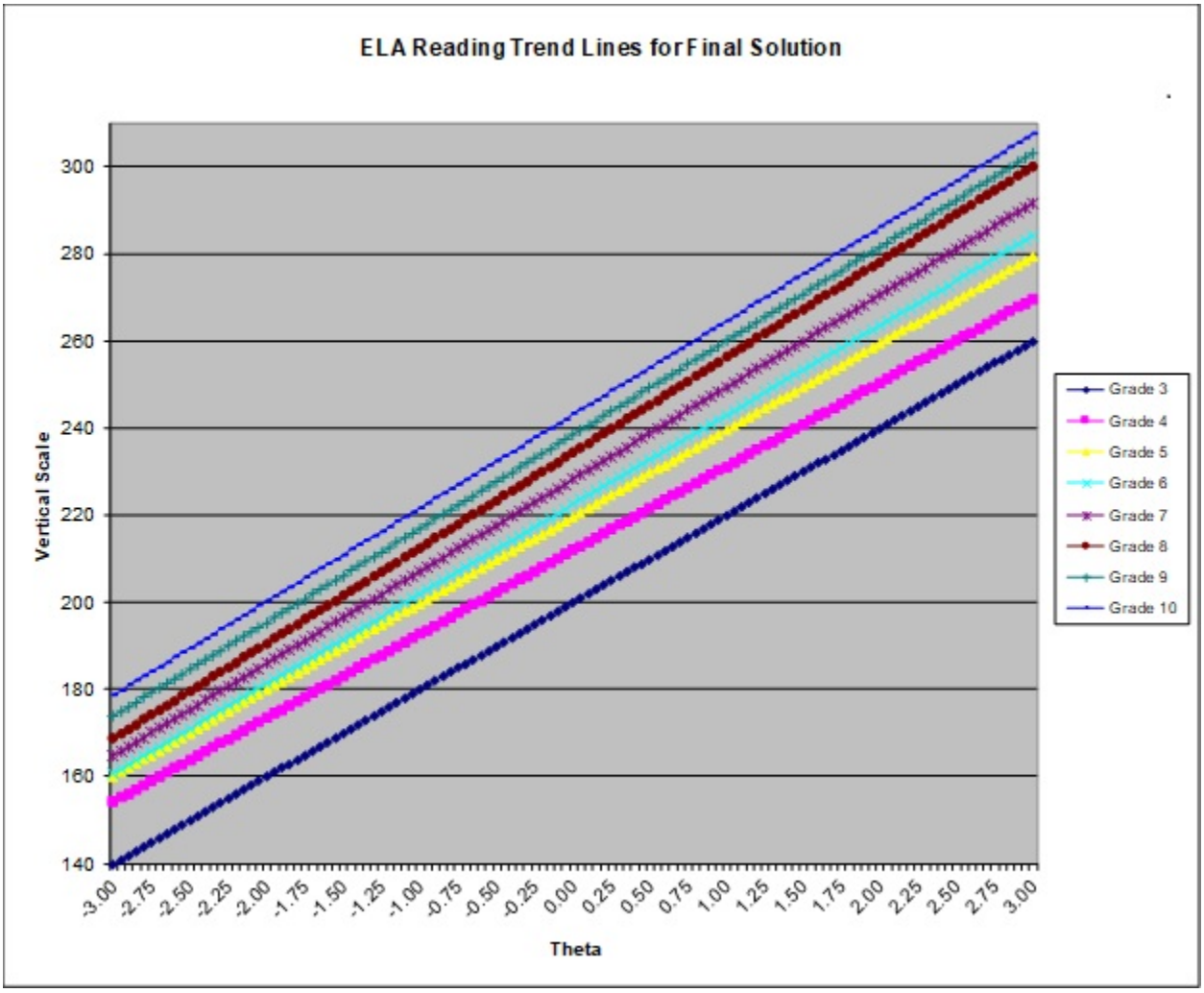
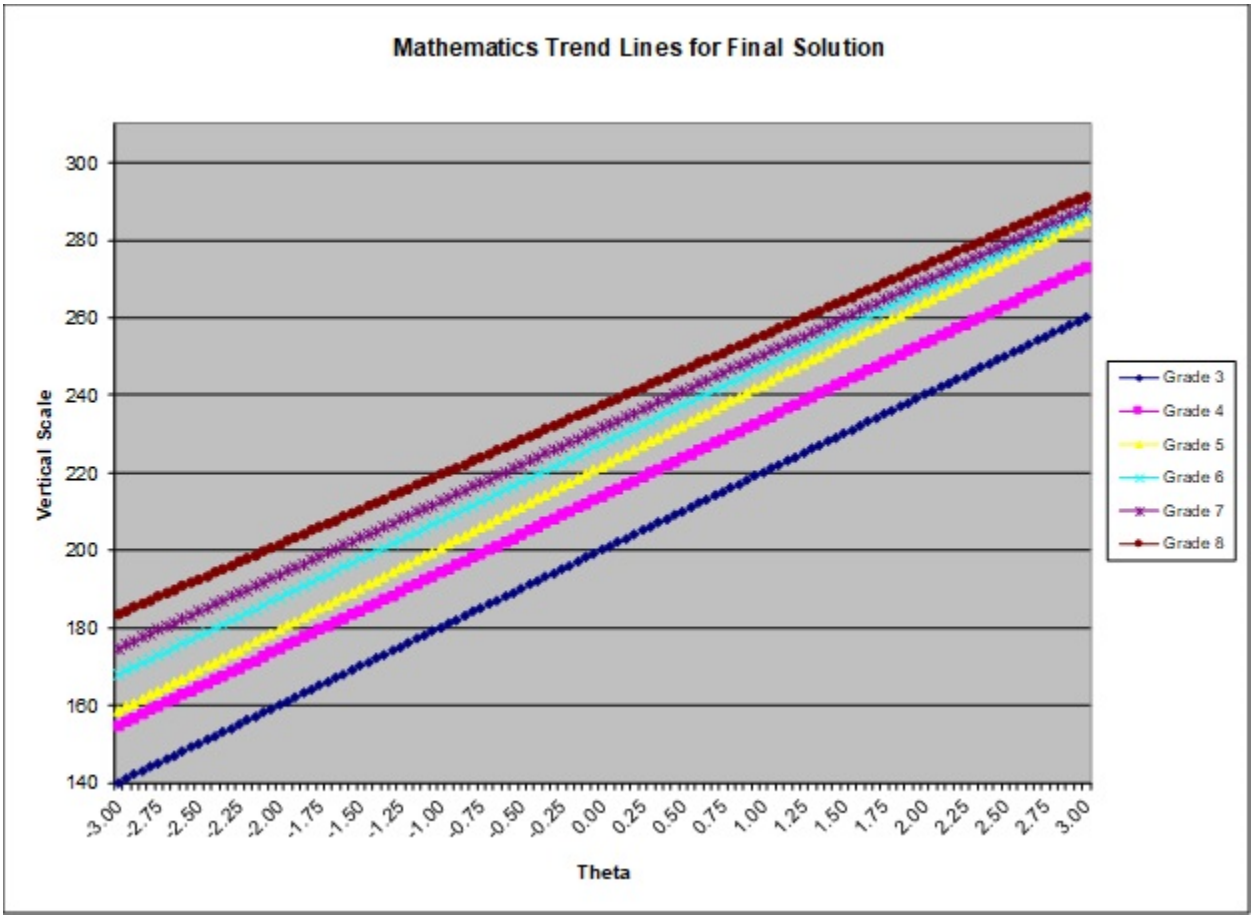


Figure 2: Mathematics Trend Lines for Final Solution



Vertical Linking Between Grades 2 and 3

CAI and Renaissance conducted a linking study to establish a linkage between the Grade 2 Star assessments in reading and mathematics and the new Grade 3 FAST PM assessments in ELA and mathematics. A chain-linking approach was used to establish a linkage between the Star and FAST scales. This embeds operational test items from adjacent grade-level assessments into the field-test slots of each grade's operational test administration. To implement this linking design, a set of Grade 2 Star items (31 reading items and 27 mathematics items) was embedded in the Grade 3 FAST tests, and a set of Grade 3 FAST items (42 reading items and 36 mathematics items) was embedded in Renaissance's Grade 2 Star tests.

The linking calibration used operational summative items and vertical linking items. For the linking items, items were dropped after examination of the criteria outlined in Table 40. Summative and vertical linking items were concurrently calibrated by fixing the operational summative item parameters. After the calibration of the linking items, the linking items between the two grades had two sets of item parameters: one set of parameters on the FAST scale and another set on the Star scale. The linking constants were then calculated with the two sets of item parameters. The challenge in linking Grade 2 to Grade 3 is that the Star and the FAST tests are based on different IRT models. The Star assessments use the Rasch model to scale the Star tests, while the FAST assessments use the 3PL model to scale the FAST tests. To avoid linking between

the Rasch model parameters used in the Star assessment and the 3PL model parameters in the FAST assessment, only forward-linking and backward-linking methods were implemented. For forward linking, Grade 2 Star assessment items embedded in the Grade 3 FAST tests were calibrated and anchored on the FAST operational summative item parameters. For backward linking, Grade 3 FAST assessment items embedded in the Grade 2 Star tests were calibrated and anchored on the Star Grade 2 operational item parameters. The *A* and *B* linking constants were obtained using mean-mean and mean-sigma methods for forward linking with the Grade 2 Star items in the Rasch model. For backward linking, the Stocking–Lord method (Stocking & Lord, 1983) was used with the Grade 3 FAST items in the 3PL model. After the preliminary review of linking results, items were further adjusted based on *p*-value reversal between grades, D^2, and adequate blueprint representation to achieve a final solution.

The linking results showed that the forward mean-mean method did not perform well in reading, and the Stocking–Lord (Stocking & Lord, 1983) method backward-linking results showed comparable growth to the mean-sigma results in ELA and mean-mean results in mathematics. Considering these results, as well as the fact that the Grade 2 Star linking items represented the blueprint content area better than the Grade 3 FAST linking items, FDOE elected to adopt the mean-sigma linking results for ELA and the mean-mean linking results for mathematics. These are the results from forward linking. Table 46 shows the number of items remaining in the final vertical linking set for ELA reading and mathematics. Tables 47 and 48 show the final vertical linking constants and vertical linking results, including theta mean and growth on the FAST scale, and growth from Renaissance's national data as reference.

Table 46: Number of Items Administered, Removed, and Remaining in the Final Linking Sets for Grades 2 and 3

Subject	Vertical Linking Items Administered	Number of Vertical Linking Items Removed	Final Vertical Linking Set
ELA Reading	34	9	25
Mathematics	34	17	17

Table 47: Final Linking Constants Between Star and FAST Assessments for Grades 2 and 3

Subject	Grade	Linking Method	Slope	Intercept
ELA Reading	2 to 3	Mean-Sigma	0.72745	–0.43737
Mathematics	2 to 3	Mean-Mean	1.00000	0.38240

Table 48: Descriptive Statistics for Star Assessments on the FAST Vertical Scale

Subject	Grade	*N*	FAST Scale		Growth from Renaissance's National Data
			Theta Mean	Growth	
ELA Reading	2	207,179	–1.01591	0.95862	1.11810
Mathematics	2	205,437	–1.33223	1.29447	1.17663

After the final linking constants were selected, a concordance table was constructed containing Star scaled scores and corresponding FAST equivalent scaled scores. Star assessments for

Grades K–2 are linked on a common vertical scale referred to as the Star unified scale, and a concordance table is used to provide equivalent FAST scores for the Star assessments in Grades K–2. Because the linking constants were calculated based on the Star unadjusted theta scale to the FAST theta scale, the FAST equivalent scores were calculated based on the Star unadjusted theta scores that correspond to each Star unified scaled score point.

Because FAST and Star assessments have different score ranges, some Star unified scaled scores map to multiple FAST scaled scores or negative FAST scaled scores in ELA. FDOE proposed using the highest FAST scaled score of multiple scores mapped to a single Star scaled score and capping negative ELA scores at zero. The final concordance table is provided in Appendix H, Concordance Tables for Star and FAST.

More information about the Star reporting scale is presented in *Renaissance Learning Star Assessments™ for Reading Technical Manual – Florida* and *Star Assessments™ for Math Technical Manual – Florida*.

Table 49: Final Theta-to-Scaled Score Transformation Equations Between Star and FAST Assessments

Subject	Grade	Theta-to-Scaled Score Transformation
ELA Reading	K–2	FAST scaled score = round (Star Reading theta *14.549044 + 191.252651)
Mathematics	K–2	FAST scaled score = round (Star Mathematics theta *20.000000 + 207.648091)

6.3 SCIENCE AND SOCIAL STUDIES—UPDATING THE SCALE

FDOE decided to move the science and social studies assessments to fixed-length CAT starting with the spring 2024 administration. Before spring 2024, the science and social studies assessments were linear tests: science for Grades 5 and 8 was administered on paper, and Biology 1, Civics, and U.S. History EOC were administered online. There were no changes to the blueprints, content standards, and Achievement-Level Descriptors (ALDs). The same reporting scales and cut scores were kept. For the spring 2024 administration, pre-equated scale scores based on item parameters from the CAT item bank were released on score reports.

To calibrate the complete item pools for future CAT, an operational field-test (OPFT) design was implemented for spring 2024. Item selection was configured to achieve a blueprint match for each test administration, but item selection will proceed independently of item difficulty. Each item (and each passage set for science) in the current pool was administered randomly to Florida students, supporting calibration of the complete item pools. In addition, the newly developed items (and passage sets for science) were randomly selected for administration in the embedded field-test slots in the spring 2024 test administration, resulting in a random and representative sample of student responses to each item.

Following the spring 2024 administration, several calibration activities took place:

- Operational calibration for all tests, which included all operational field-test items that were referred to as operational items in all calibration and equating activities

- Equating to the baseline scale (i.e., spring 2012 for Grades 5 and 8 science and Biology 1, spring 2013 for U.S. History, and spring 2014 for Civics)
- Field-test item calibration for all tests, which involved concurrent calibration of the operational and embedded field-test items, with operational item parameters fixed

The sample consisted of all data from the spring administration, except for students who were retaking the test or part of exclusion rules.

Operational Calibrations and Equating to Calibrated Pool

Operational calibrations were completed first and included all the operational items. All items were freely calibrated and equated back to the IRT-calibrated item pool.

Post-equating the CAT item parameters to the IRT-calibrated item pool was completed in a manner like the post-equating conducted in spring 2016–2019 and spring 2021–2023. The anchor set consisted of all operational items except those that had minor edits since their last use, those with poor statistics identified through key check or calibration check (e.g., a-parameter < 0.5, –6 < b-parameter < 6, Q1_Flag), or those with a small sample size (< 3000). Using the calibrated item statistics from IRTPRO, the anchor items were used to identify the equating constants to place the 2024 CAT item parameters in the IRT-calibrated item pool. The Stocking–Lord (Stocking & Lord, 1983) method was used. The Stocking–Lord method is commonly used alongside the 3PL model and GPCM and establishes the linking constants. This was the same method used for ELA and mathematics described in Section 6.2.2 Vertical Linking ELA and Mathematics of this volume of this technical report.

Table 50 presents the final equating results, including the number of items in the equating design, the number of dropped items, and the number of items in the final equating solution. The last two columns show the slope and intercept from the final Stocking–Lord equating solution. The intercept and slope represent the first and second moments of the ability distribution, respectively. Hence, slope values greater than 1 indicate greater heterogeneity in the population relative to the baseline year, and values less than 1 indicate greater homogeneity than previously observed. Similarly, intercept values greater than 0 indicate an improvement in mean performance relative to the baseline group, and values less than 0 denote the opposite. The *Number of Items in Design* column refers to the size of the equating set for a given test.

Table 50: Final Equating Results

Grade/Course	Number of Items in Design	Number of Items Dropped	Number of Items in Final Solution	Slope	Intercept
Science 5	784	0	784	1.09700	0.141520
Science 8	688	0	688	1.07429	–0.035357
Biology 1	830	0	830	1.04625	0.300680
U.S. History	615	0	615	1.08621	0.410890
Civics	546	0	546	1.10558	0.367970

Post-Administration Analyses

Following the spring 2024 administration, Pearson and CAI conducted the following post-administration analyses for validation purposes and to decide on the CAT scale for future CAT scoring.

1. Student performance comparison of 2023 and 2024:
    - Spring 2024 reported scale scores based on bank parameters (pre-equated): means, standard deviations, and cumulative frequency distributions
    - Historical impact data based on existing cut scores: percentage of students at each achievement level and percentage at Levels 3 and above
2. Comparison of the two sets of 2024 scale scores derived from bank parameters and the freely calibrated CAT item parameters, with the following analyses:
    - Scale score means, standard deviations, and correlations
    - Impact data comparison based on existing cut scores:
        - Percentage of students at each achievement level and percentage at Levels 3 and above
        - A 5-x-5 achievement classification table (comparing old and new performance levels) to check classification consistency
3. Comparison of the two sets of 2024 scale scores derived from bank parameters and the new scaled CAT item parameters, with the following analyses:
    - Scale score means, standard deviations, and correlations
    - Impact data comparison based on existing cut scores
        - Percentage of students at each achievement level and percentage at Levels 3 and above
        - A 5-x-5 achievement classification table (comparing old and new performance levels) to check classification consistency
4. Impact data from solutions 1–3 were compared to spring 2023 impact data and historical data, starting from the baseline year.

Tables 51–55 show the descriptive statistics, impact data, and slope and intercepts for each historical year and 2024 calibration method. The scores based on the post-equated scale are similar to the scores based on the existing pre-equated scale. Table 56 shows the correlation between the calibration methods is highly correlated. The full results from the equating and post-administration analysis are presented in Appendix I, Science and Social Studies Equating Reports.

Table 51: Grade 5 Science Impact Data

Admin	*N*	Mean	Standard Deviation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 3 & Above	Slope	Intercept
2012	199,158	200.27	21.57	22.39	25.60	27.14	12.45	12.42	52.01	1	0
2013	195,130	201.33	22.09	21.35	25.53	26.74	12.35	14.03	53.12	1.02541	0.03746
2014	195,645	201.40	21.41	20.58	25.37	27.90	13.05	13.10	54.05	0.99128	0.04535
2015	198,515	200.43	21.60	21.96	25.32	27.48	12.91	12.33	52.72	1.00654	0.00600
2016	202,655	199.63	21.45	22.82	25.97	27.32	12.51	11.38	51.21	0.99637	–0.02939
2017	212,952	199.54	22.11	23.97	24.91	26.47	12.42	12.23	51.12	1.02999	–0.02828
2018	211,927	201.59	21.58	20.34	24.74	28.06	13.41	13.45	54.92	0.99937	0.06408
2019	218,715	200.20	21.86	22.56	24.74	27.31	12.97	12.41	52.69	1.01134	0.01966
2021	195,881	197.09	22.63	27.88	25.35	25.10	11.18	10.49	46.77	1.05386	–0.15354
2022	211,856	197.32	24.27	28.94	23.05	24.24	11.46	12.32	48.02	1.15043	–0.14973
2023	204,670	199.42	23.32	25.07	23.57	25.41	12.74	13.21	51.36	1.10282	–0.04265
2024 pre-equated	174,446	202.35	23.72	20.51	21.23	27.13	14.90	16.24	58.27	—	—
2024 free calibration	174,446	200.26	21.72	21.60	25.09	28.31	13.03	11.97	53.31	—	—
2024 post-equated	174,446	203.08	23.56	20.09	21.72	26.56	14.32	17.31	58.19	1.09700	0.141522

Table 52: Grade 8 Science Impact Data

Admin	*N*	Mean	Standard Deviation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 3 & Above	Slope	Intercept
2012	193,394	200.14	21.65	22.23	30.54	21.96	13.45	11.82	47.23	1	0
2013	195,683	199.95	21.36	21.99	30.76	22.93	13.30	11.01	47.24	0.97717	0.0024
2014	197,208	200.59	21.30	21.13	30.15	23.28	13.82	11.62	48.72	0.98063	0.02119
2015	196,513	200.56	21.36	21.67	30.26	22.59	13.37	12.11	48.07	0.9896	0.02407
2016	190,668	200.67	21.24	21.50	30.35	22.88	13.31	11.96	48.15	0.97669	0.02543
2017	190,652	200.23	22.47	23.10	28.76	21.68	13.51	12.94	48.13	1.04173	0.0042
2018	193,801	201.00	21.88	21.74	28.17	22.59	14.68	12.83	50.10	1.01453	0.03484
2019	195,621	200.69	20.86	21.24	30.31	23.35	13.73	11.38	48.46	0.96203	0.06078
2021	188,147	198.44	21.95	25.04	30.18	21.72	12.53	10.54	44.79	1.01618	–0.07865
2022	198,831	198.30	22.98	26.23	28.78	21.25	12.44	11.30	44.99	1.07117	–0.10571
2023	200,961	197.88	23.00	26.83	29.20	20.67	12.08	11.22	43.97	1.06867	–0.13371
2024 pre-equated	178,237	199.16	24.06	25.04	26.20	21.31	14.21	13.23	48.75	—	—
2024 free calibration	178,237	200.32	22.26	22.11	28.37	22.86	14.31	12.34	49.51	—	—
2024 post-equated	178,237	199.74	23.56	24.17	27.27	21.34	13.68	13.54	48.56	1.074288	–0.035357

Table 53: Biology 1 Impact Data

Admin	N	Mean	Standard Deviation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 3 & Above	Slope	Intercept
2012	190,340	398.65	27.88	14.16	26.81	36.74	10.80	11.49	59.03	1	0
2013	186,884	403.63	27.62	10.37	23.47	38.18	12.44	15.55	66.17	0.97940	0.23880
2014	195,495	404.93	27.38	9.50	22.96	37.79	12.95	16.80	67.54	0.98806	0.23292
2015	204,250	402.88	28.45	11.93	23.29	36.69	12.46	15.63	64.78	1.02191	0.16174
2016	197,663	402.59	28.88	12.20	23.94	35.93	11.70	16.22	63.85	1.03985	0.15057
2017	195,316	402.31	29.51	12.61	23.59	35.40	12.23	16.17	63.80	1.04515	0.14259
2018	192,870	404.22	28.67	11.05	23.50	35.19	12.66	17.60	65.45	1.03960	0.20179
2019	199,690	405.63	29.40	10.98	21.63	34.84	12.76	19.79	67.39	1.06669	0.26302
2021	188,281	400.73	29.45	13.47	25.32	34.94	11.46	14.81	61.21	1.05573	0.07658
2022	205,696	400.89	30.62	14.40	24.81	32.95	11.36	16.47	60.78	1.10678	0.07831
2023	215,366	402.77	30.46	13.26	23.49	33.70	11.80	17.75	63.25	1.11692	0.13727
2024 pre-equated	199,680	406.06	30.00	11.17	19.69	35.02	13.80	20.32	69.14	—	—
2024 free calibration	199,680	399.43	27.85	13.59	25.66	38.37	10.96	11.42	60.75	—	—
2024 post-equated	199,680	406.79	29.21	10.17	20.56	35.30	13.47	20.50	69.27	1.046247	0.300677

Table 54: U.S. History Impact Data

Admin	N	Mean	Standard Deviation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 3 & Above	Slope	Intercept
2013	133,573	398.74	28.36	21.02	22.41	28.7	16.77	11.11	56.58	1	0
2014	163,887	405.77	28.60	15.62	18.91	27.37	19.92	18.18	65.47	1.01314	0.19872
2015	169,500	406.53	28.14	15.13	19.21	27.12	19.49	19.06	65.67	1.01405	0.27360
2016	168,914	406.64	28.32	15.18	18.81	27.01	19.92	19.07	66.00	1.01949	0.26042
2017	176,039	407.53	28.42	14.63	18.35	26.46	20.27	20.29	67.02	1.02257	0.31435
2018	181,143	408.39	29.18	14.65	17.29	26.00	20.11	21.94	68.05	1.05315	0.32386
2019	177,696	409.74	29.48	13.56	16.61	25.98	20.75	23.10	69.83	1.06156	0.39195
2021	155,675	404.68	30.29	17.54	19.32	26.13	18.48	18.53	63.14	1.08883	0.22896
2022	178,770	406.49	29.55	16.41	18.38	25.53	19.62	20.06	65.21	1.08114	0.29094
2023	189,980	404.92	30.81	18.86	18.11	24.91	18.48	19.64	63.04	1.13477	0.22709
2024 pre-equated	183,147	408.93	30.65	15.01	15.69	24.99	20.25	24.06	69.30	—	—
2024 free calibration	183,147	399.46	27.72	20.72	21.52	29.29	17.48	10.99	57.76	—	—
2024 post-equated	183,147	409.56	30.14	14.57	16.07	24.95	20.13	24.29	69.37	1.08621	0.410889

Table 55: Civics Impact Data

Admin	N	Mean	Standard Deviation	Level 1	Level 2	Level 3	Level 4	Level 5	Level 3 & Above	Slope	Intercept
2014	200,604	398.93	28.57	18.99	20.50	27.10	18.51	14.89	60.50	1	0
2015	196,818	402.39	28.30	16.25	19.09	26.28	20.03	18.35	64.66	1.00885	0.13221
2016	197,966	404.09	28.45	15.09	17.72	26.80	20.35	20.04	67.19	1.01387	0.20464
2017	200,980	406.27	28.35	13.05	17.43	25.95	20.92	22.66	69.53	1.01146	0.28596
2018	199,288	407.64	28.85	12.75	16.50	25.16	20.85	24.74	70.75	1.04218	0.33898
2019	213,183	407.74	28.82	13.01	15.98	25.10	20.99	24.93	71.02	1.04938	0.33403
2021	200,618	402.35	30.47	17.52	18.39	25.46	18.36	20.27	64.09	1.09207	0.13500
2022	209,801	406.43	29.95	15.11	15.82	24.46	20.27	24.34	69.07	1.09536	0.28702
2023	206,946	403.68	30.80	17.59	16.73	24.50	19.34	21.85	65.68	1.12469	0.18601
2024 pre-equated	188,311	407.82	31.07	14.39	14.38	23.41	20.78	27.04	71.23	—	—
2024 free calibration	188,311	399.42	27.99	18.25	20.04	28.57	18.70	14.45	61.72	—	—
2024 post-equated	188,311	408.41	30.74	13.69	14.90	23.68	20.42	27.31	71.41	1.105578	0.36797

Table 56: Scale Score Correlations for Each Scale

Grade	Scale	Pre-Equated	Post-Equated	Free
Grade 5 Science	Pre-equated	1.000	—	—
	Post-equated	0.996	1.000	—
	Free	0.996	0.999	1.000
Grade 8 Science	Pre-equated	1.000	—	—
	Post-equated	0.992	1.000	—
	Free	0.991	0.999	1.000
Biology 1	Pre-equated	1.000	—	—
	Post-equated	0.992	1.000	—
	Free	0.992	0.999	1.000
U.S. History	Pre-equated	1.000	—	—
	Post-equated	0.993	1.000	—
	Free	0.993	0.999	1.000
Civics	Pre-equated	1.000	—	—
	Post-equated	0.994	1.000	—
	Free	0.993	0.999	1.000

Considering these results, FDOE chose the post-equated scale as the scale for future CAT administrations. The item parameters will be used for CAT as the new pre-equated parameters. In spring 2024 and beyond, all field-test items are concurrently calibrated, with the scaled operational item parameters fixed to the post-equated scale. This option updates item bank parameters (taking into account changes in student ability over time) while still tying the new scale back to pre-established cut scores.

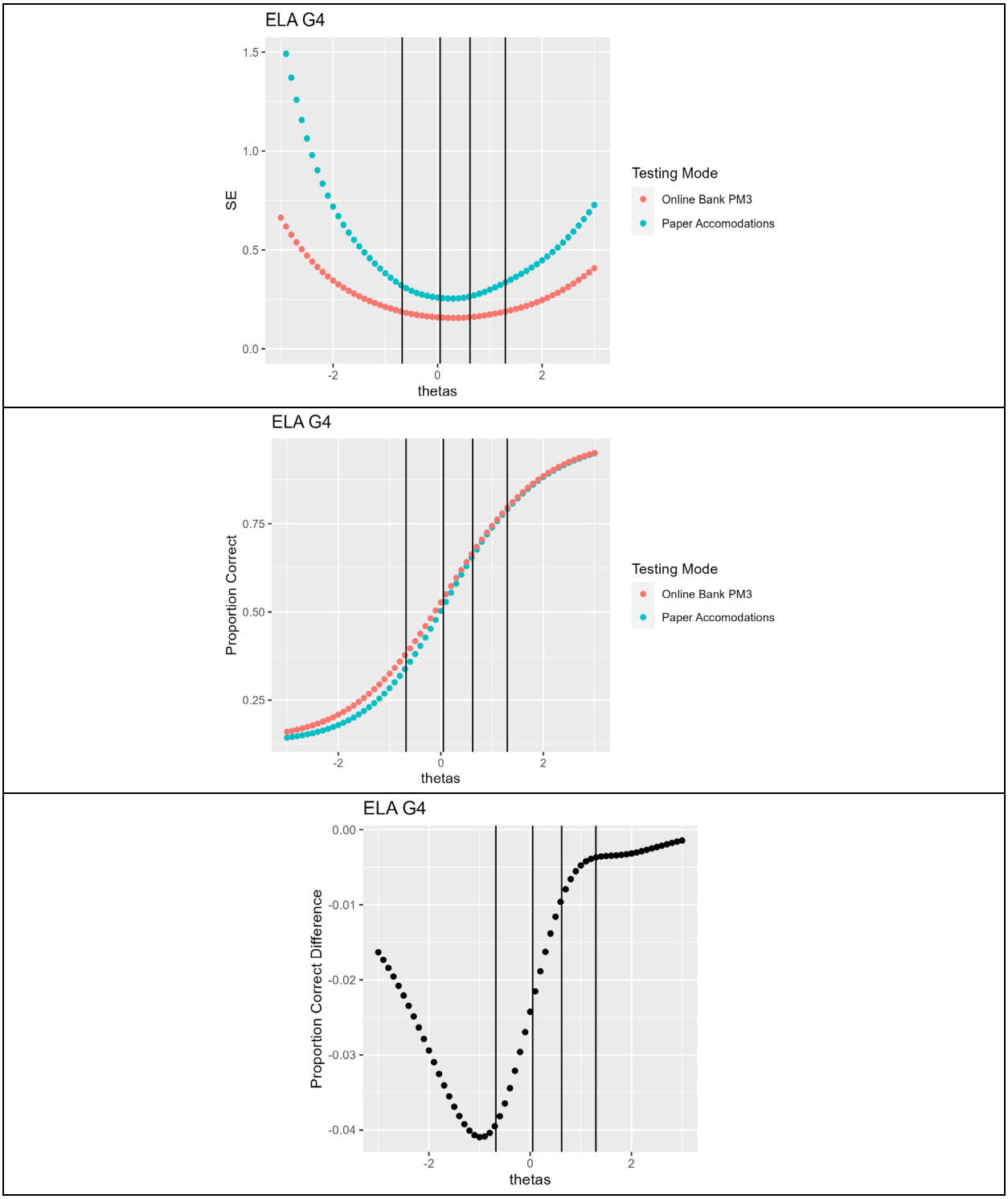
6.4 ACCOMMODATED FORMS

Accommodated forms used online pre-equated parameters for scoring purposes, and no calibrations were performed on the accommodated forms.

To create the spring 2025 accommodated forms for ELA and mathematics, CAI's automated form-building tool inside CAI's Item Tracking System (ITS) was used. ITS is a web-based software application. In conjunction with the Item Authoring Tool (IAT), which can be accessed through the system, ITS enables users to create items and stimuli for testing purposes. Once the items and stimuli have been created, users can review, approve, and publish these items in ITS so they can be administered to students through the Test Delivery System (TDS).

Psychometric targets were set by CAI psychometricians at the individual item level and overall form level in conjunction with FDOE, using the ITS automated form-building tool. The tool constructs the forms based on the selection of individual items (after evaluation of their statistics and blueprint match) and comparison of the forms against bank averages and characteristics, in addition to minimizing the SE at the grade-level cut. Figure 3 is a sample representation of that evaluation. In these evaluations, there are no expectations that the statistical characteristics for a form composed of a limited number of items would overlap completely with the entire bank. However, the patterns observed should be consistent. All parties review two forms per grade and make recommendations to the CAI Content team, who then make necessary item replacements, taking into account suitability for inclusion in an accommodated form and psychometric feedback. FDOE then selects the final form. More detailed information about accommodated form construction is presented in Section 4.4, Accommodation Form Construction, of Volume 2, *Test Development.* Accommodated forms for science and social studies built in 2024 and beyond also follow this procedure. Further psychometric information about the 2024 accommodated forms is presented in Appendix C, Test Characteristic Curves with SEMs.

Figure 3: Sample Psychometric Curves for Fixed Forms with Performance-Level Cuts



6.5 IRT ITEM SUMMARIES

6.5.1 Item Fit

Yen's Q1 is used to evaluate the degree to which observed data fit the item response model (Yen, 1981). Q1 is a fit statistic that compares observed and expected item performance. To calculate fit statistics before scores were available from CAI's scoring engine, Maximum A Posteriori (MAP) estimates from IRTPRO were used for student ability estimates in the calculations. IRTPRO does not calculate the maximum likelihood estimation (MLE); however, the prior mean and variance for the MAP were set to 0 and 10,000, respectively, so that the resulting MAP estimates approximate the MLE.

Q1 is calculated as

$$\mathrm{Q_{1i}} = \sum_{\mathrm{j=1}}^{\mathrm{J}} \frac{\mathrm{N_{ij}}(\mathrm{O_{ij}} - \mathrm{E_{ij}})^2}{\mathrm{E_{ij}}(1 - \mathrm{E_{ij}})},$$

where N_{ij} is the number of test takers in cell j for item i, and O_{ij} and E_{ij} are the observed and predicted proportions of test takers in cell j for item i. The expected or predicted proportion is calculated as

$$\mathrm{E_{ij}} = \frac{1}{\mathrm{N_{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N_{ij}}} \mathrm{P_i}(\hat{\theta}_\mathrm{a}),$$

where $P_i(\hat{\theta}_a)$ is the item characteristic function for item i and test taker a. The summation is taken over test takers in cell j. The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$$gen\ \mathrm{Q_{1i}} = \sum_{\mathrm{j=1}}^{\mathrm{J}} \sum_{\mathrm{k=1}}^{\mathrm{m_i}} \frac{\mathrm{N_{ij}}(\mathrm{O_{ikj}} - \mathrm{E_{ikj}})^2}{\mathrm{E_{ikj}}}$$

with

$$\mathrm{E_{ikj}} = \frac{1}{\mathrm{N_{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N_{ij}}} \mathrm{P_{ik}}(\hat{\theta}_\mathrm{a}).$$

To determine acceptable fit, the results for both the Q1 and Generalized Q1 are transformed into the statistic $\mathrm{ZQ_1}$,

$$\mathrm{ZQ_1} = \frac{\mathrm{Q_1} - \mathrm{df}}{\sqrt{2\mathrm{df}}},$$

and are compared to a criterion $\mathrm{ZQ_{crit}}$ (FDOE, 1998),

$$\mathrm{ZQ_{crit}} = \frac{\mathrm{N}}{1500} * 4,$$

where Q is either Q1 or Generalized Q1 and *df* is the degrees of freedom for the statistic. The degrees of freedom are calculated as $J * (K - 1) - m$ where J is the trait interval, K is the number of score categories, and m is the number of estimated item parameters in the IRT model. In Yen (1981), the trait interval of 10 is used. For example, MC items have $df = 10 * (2 - 1) - 3 = 7$. Poor fit is indicated where ZQ_1 is greater than ZQ_{crit}.

The number of items flagged by Q1 is presented in Appendix A, Operational Item Statistics, for operational items and in Appendix B, Field-Test Item Statistics, for field-test items.

No more than one operational item was flagged for fit as measured by Q1 in each test. Psychometricians and content specialists reviewed the items before a final decision was made about their inclusion in student score calculations.

Appendix B lists the number of field-test items by grade and subject flagged by Q1. Before field-test items are placed on forms for operational use in future test administrations, content specialists and psychometricians will review them. More information about test construction and item review is presented in Volume 2, *Test Development,* of this technical report.

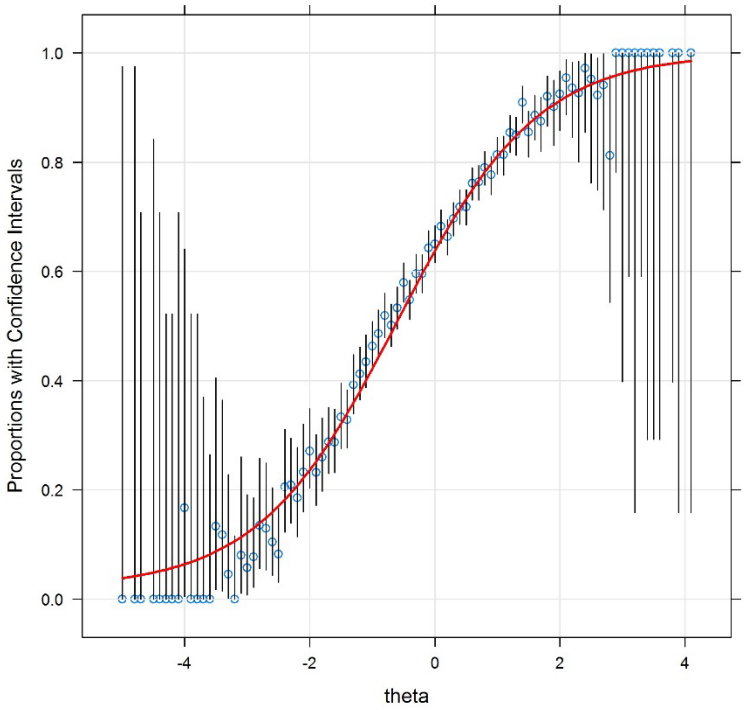
6.5.2 Item Fit Plots

Another way to evaluate item fit is to examine empirical fit plots for each item. The plots in this section are only examples of the types of fit plots used during item calibrations to add to the collection of evidence to evaluate item quality.

Fit plots were created for all items during calibration and are available on request. Along with classical item statistics and Q1 flags, item fit plots were used to review items.

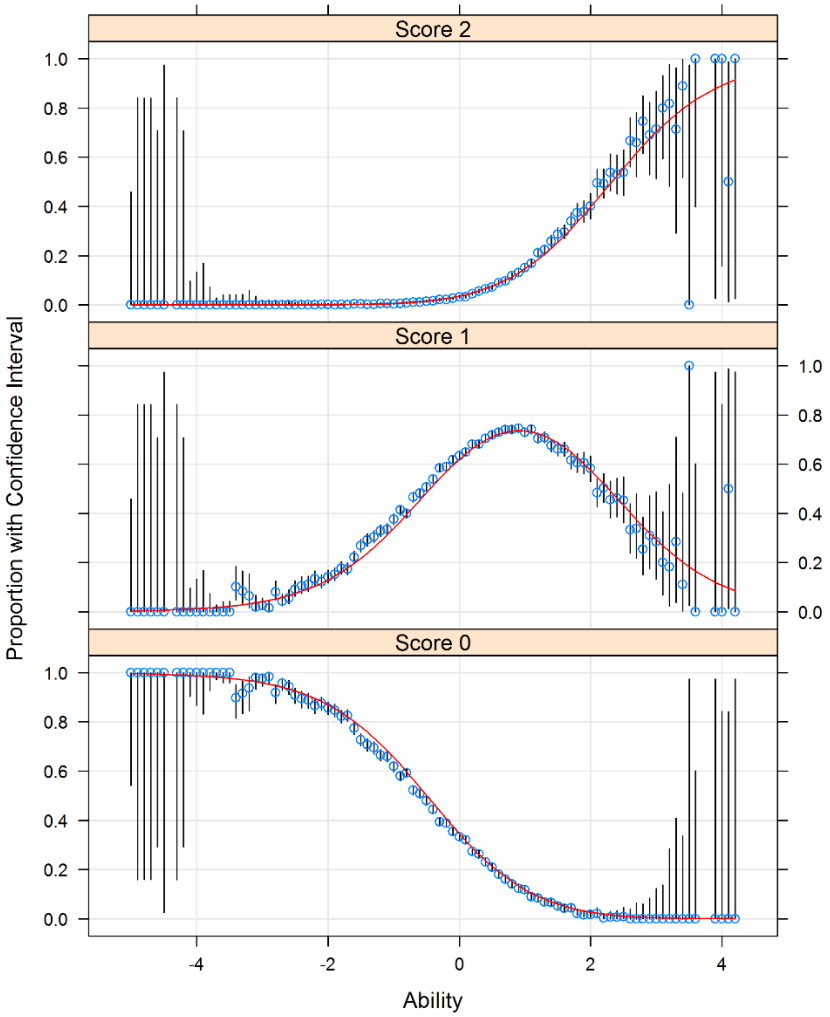
The fit plot in Figure 4 illustrates a 1-point item that fits the item response model well. The blue dots represent the proportion of students within a score bin correctly answering the item. The red solid line is the IRT-based ICC. The black lines indicate the error bands associated with the ICC for each theta point. A “good” item is one in which the observed dots follow the red solid line in the error bands across the range of ability.

Figure 4: Example Fit Plot—1-Point Item



The plot in Figure 5 is provided for items worth 2 points or more. Again, the red lines represent the IRT-based ICC. Here, the dots represent the percentage of students within a score bin, at each score point. Like Figure 4, a "good" item is one in which the observed dots follow the red solid line within the error bands across the range of ability.

Figure 5: Example Fit Plot—2-Point Item



6.6 Results of Calibrations Including Field-Test Items

The item pools will grow each year as a result of field testing for new items. Any item used in an assessment is field-tested before it is used as an operational item. In spring 2025, the tests included EFT items.

To put the field-test items on the operational scale, all operational and field-test items were concurrently calibrated, with the scaled operational item parameters fixed. Convergence was reviewed, and where issues arose, priors were placed on items. Item statistics and parameters from these analyses were uploaded into CAI's ITS.

Classical item analyses ensure that items function as intended with respect to the underlying scales. CAI's analysis program, Workspace, computes the required item and test statistics for each MC item to check the integrity of the item and to verify the appropriateness of the item's difficulty level.

The results of the classical item analysis and IRT analysis are described in Section 5, Item Analyses Overview, and are presented in Appendix A for the spring 2025 operational items and in Appendix B for the spring 2025 field-test items.

7. SCORING

This chapter provides the scoring procedure used in tests administered in the 2024–2025 school year. It covers the computational details of the maximum likelihood estimation (MLE), standard error (SE) of estimate, scale scores, and performance levels reported.

7.1 FLORIDA ASSESSMENTS SCORING

7.1.1 Maximum Likelihood Estimation

The tests were based on the three-parameter logistic (3PL) model and generalized partial credit model (GPCM) of item response theory (IRT) models, with the two-parameter logistic (2PL) model treated as a special case of the 3PL model. Theta scores were generated using pattern scoring, a method that scores students differently depending on how they answer individual items.

Likelihood Function

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

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

where

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

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

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

$$Q_i = 1 - P_i,$$

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 item, *h* indexes the step of the item, m_i is the maximum possible score point (starting from 0), δ_{ki} is the *k*th step for item *i* with *m* total categories, and $D = 1.7$.

A student's theta based on the MLE estimate is defined as $\arg\max_{\theta} log\big(L(\theta)\big)$ given the set of items administered to the student.

Derivatives

Finding the maximum likelihood requires an iterative method, such as Newton–Raphson iterations. The estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} \Big/ \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t},$$

where

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} = \frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} = \sum_{i=1}^{N_{3PL}} Da_i \frac{(P_i - c_i)Q_i}{1 - c_i}\left(\frac{z_i}{P_i} - \frac{1 - z_i}{Q_i}\right)$$

$$\frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} = -\sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1 - c_i)^2}\left(1 - \frac{z_i c_i}{P_i^2}\right)$$

$$\frac{\partial lnL(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N_{CR}} Da_i \left(z_i - \frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}\right)$$

$$\frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}\right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}\right)$$

and where θ_t denotes the estimated θ at iteration *t,* N_{CR} is the number of items that are scored using the GPCM, and N_{3PL} is the number of items scored using the 3PL or 2PL model.

SEs of Estimate

When the MLE is available, the SE of the MLE is estimated by:

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

where

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

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

Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded, and an MLE cannot be generated. In addition, when a student's raw score is lower than the expected raw score due to guessing, the likelihood is not identified. The extreme cases were handled as follows:

1. Assign the lowest obtainable theta (LOT) value of –3 to a raw score of 0.
2. Assign the highest obtainable theta (HOT) value of 3 to a perfect score.
3. 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.

SE of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the SE is estimated based on Fisher information.

When the MLE is not available (such as for extreme score cases) or the MLE is censored to the LOT or HOT, the SE for student *s* is estimated by:

$$se(\theta_s) = \frac{1}{\sqrt{I(\theta_s)}},$$

where $I(\theta_s)$ is the test information for student *s.* Tests included items that were scored using the 3PL model, 2PL model, and GPCM from IRT. The 2PL model can be visualized as either a 3PL item with no pseudo-guessing parameter or a dichotomously scored GPCM item. The test information was calculated as

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

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

For SE of LOT/HOT scores, theta in the formula above is replaced with the LOT/HOT values.

A global maximum of 1.5 is applied to all SEs.

7.1.2 Scale Scores

Three scale types are created for the assessments:

1. A vertical scale score for Grades 3–10 English Language Arts (ELA) and Grades 3–8 mathematics
2. A within-test scaled score for mathematics End-of-Course (EOC) tests, science, and social studies
3. Raw score reporting for writing

Table 57 shows the theta-to-scale score transformation equations.

Table 57: Theta-to-Scale Score Transformation Equations

Subject	Grade	Theta-to-Scale Score Transformation
ELA	3	Scale Score = round(theta * 20 + 200)
	4	Scale Score = round(theta * 19.24464 + 212.04895)
	5	Scale Score = round(theta * 19.88239 + 219.71302)
	6	Scale Score = round(theta * 20.56381 + 222.52838)
	7	Scale Score = round(theta * 21.14869 + 228.31157)
	8	Scale Score = round(theta * 21.90164 + 234.48903)
	9	Scale Score = round(theta * 21.54087 + 238.55054)
	10	Scale Score = round(theta * 21.46475 + 243.19982)
Mathematics	3	Scale Score = round(theta * 20.000000 + 200.000000)
	4	Scale Score = round(theta * 19.69341 + 213.86243)
	5	Scale Score = round(theta * 21.06118 + 221.62960)
	6	Scale Score = round(theta * 19.83724 + 227.39906)
	7	Scale Score = round(theta * 18.94480 + 231.46678)
	8	Scale Score = round(theta * 17.98219 + 237.37017)
Algebra 1		Scale Score = round(theta * 25 + 400)
Geometry		Scale Score = round(theta * 25 + 400)
Grade 5 Science	5	Scale Score = round(theta * 20 + 200)
Grade 8 Science	8	Scale Score = round(theta * 20 + 200)
Biology 1	10	Scale Score = round(theta * 25 + 400)
U.S. History	9	Scale Score = round(theta * 25 + 400)
Civics	7	Scale Score = round(theta * 25 + 400)

When calculating the scale scores, the following rules were applied:

1. The same linear transformation was used for all students in a grade.
2. Scale scores were rounded to the nearest integer (e.g., 302.4 to 302; 302.5 to 303).
3. An SE was provided for each score, using the same set of items used to derive the score.

The SE of the scaled score is calculated as:

$se(SS) = se(\theta) * slope,$

where $slope$ is the slope from the theta-to-scaled score transformation equation in Table 57.

Appendix D, Distribution of Scale Scores and Standard Errors, summarizes the scale scores.

Writing responses are scored on three dimensions based on a rubric, each scored independently. These are shown in Table 58 along with the range of scores that may be assigned for each. Total raw score is equal to 12. Minimum raw score for a valid response is 3. When a response does not meet the minimal requirements to receive a rubric score, it is assigned a 0. The distribution for writing scores based on dimensions is presented in Appendix J, Writing Scores.

Table 58: B.E.S.T. Writing Dimension Scores for Valid Responses

Dimension	Possible Scores
Purpose/Structure (Organization)	1, 2, 3, or 4
Development (Elaboration)	1, 2, 3, or 4
Language (Conventions)	1, 2, 3, or 4

7.1.3 Performance Levels

Each student is assigned a performance category according to his or her accountability scale score. Tables 59–62 provide the cut scores for performance levels for mathematics, ELA reading, mathematics EOC, science, and social studies.

Table 59: Cut Scores for Mathematics by Grade

Grade	Cut Between Levels 1 and 2	Cut Between Levels 2 and 3	Cut Between Levels 3 and 4	Cut Between Levels 4 and 5
3	183	198	209	225
4	200	211	221	238
5	207	222	234	246
6	213	229	239	254
7	223	235	247	258
8	227	244	254	263

Table 60: Cut Scores for ELA Reading by Grade

Grade	Cut Between Levels 1 and 2	Cut Between Levels 2 and 3	Cut Between Levels 3 and 4	Cut Between Levels 4 and 5
3	186	201	213	225
4	199	213	224	237
5	206	222	232	246
6	209	225	237	250
7	215	232	242	257
8	220	238	251	262
9	224	242	254	267
10	230	247	258	271

Table 61: Cut Scores for Mathematics EOC

Grade	Cut Between Levels 1 and 2	Cut Between Levels 2 and 3	Cut Between Levels 3 and 4	Cut Between Levels 4 and 5
Algebra 1	379	400	418	435
Geometry	385	404	423	432

Table 62: Cut Scores for Science and Social Studies

Grade	Cut Between Levels 1 and 2	Cut Between Levels 2 and 3	Cut Between Levels 3 and 4	Cut Between Levels 4 and 5
Grade 5 Science	185	200	215	225
Grade 8 Science	185	203	215	225
Biology 1	369	395	421	431
U.S. History	378	397	417	432
Civics	376	394	413	428

Performance levels were not used for writing. Students received a rubric-based raw score for each dimension (with a range from 1–4 for each dimension) and an overall raw score that ranged from 3–12 for valid responses, as shown in Table 63.

Table 63: B.E.S.T. Writing Achievement–Score Ranges

Grade	Overall Raw Score Range	Dimension 1: Purpose/Structure (Organization)	Dimension 2: Development (Elaboration)	Dimension 3: Language (Conventions)
4	3–12	1–4	1–4	1–4
5	3–12	1–4	1–4	1–4
6	3–12	1–4	1–4	1–4
7	3–12	1–4	1–4	1–4
8	3–12	1–4	1–4	1–4
9	3–12	1–4	1–4	1–4
10	3–12	1–4	1–4	1–4

7.1.4 Alternate Passing Score

The Alternate Passing Score (APS) is the Florida Standards Assessment (FSA) equivalent score reported on the FAST and B.E.S.T. scaled score. When scores were reported in the 2022–2023 and fall 2023–2024 school years, there were no approved FAST and B.E.S.T. reporting scales, so cut scores were reported as the FSA-linked equivalent. The FAST and B.E.S.T. scale transformation constants are now known, so the passing scores can be reported on the FAST and B.E.S.T. scales. The State Board of Education has adopted the Commissioner's proposed score scale for the FAST and B.E.S.T. assessments on October 18, 2023. Since the cut scores recommended from the summer 2023 standard-setting process have been approved, it is important to note that these APS cut scores were used only with students who were retaking the test. The new FAST and B.E.S.T. cut scores will apply to students taking the FAST and B.E.S.T. assessments for the first time in 2023–2024 and beyond.

The APS was derived from the equipercentile relationship between the FSA EOC Level 2/3 cut scores, FAST and B.E.S.T. score scales, and corresponding alternative passing scores on the FAST and B.E.S.T. score scales. The following are the scores for the tests:

- The alternate passing score for FAST Grade 10 ELA is **246** and above on the FAST scale, which corresponds to the passing score of 350 and above on the FSA Grade 10 ELA.
- The alternate passing score for B.E.S.T. Algebra 1 EOC is **398** and above on the B.E.S.T. scale, which corresponds to the passing score of 497 and above on the FSA Algebra 1 EOC.
- The alternate passing score for B.E.S.T. Geometry EOC is **401** and above on the B.E.S.T. scale, which corresponds to the passing score of 499 and above on the FSA Geometry EOC.

A student's passing indicator is based on whether the scale score meets the passing requirement, and the performance level is based on the scale score and the scale score cut point exclusively.

APS Eligibility on the FAST and B.E.S.T. Assessments

Grade 10 ELA Reading

Eligibility for using the Grade 10 FSA ELA APS cut score on the B.E.S.T. score scale is based on student cohort. Students who entered Grade 9 in 2021–2022 (or prior), regardless of their first attempt taking the assessment, are eligible to use the FAST APS for graduation purposes. Also, students who entered Grade 10 in fall 2023 (or prior), regardless of their first attempt taking the assessment, are eligible to use the APS for graduation purposes. In addition, students who took the Grade 10 FAST ELA assessment in spring 2023 as above-grade-level testers (e.g., Grade 9 students receiving Grade 10 instruction) are also eligible to use the APS, even though they are *not* in the 2022–2023 cohort.

Algebra 1 and Geometry

Eligibility for using the APS for the B.E.S.T. Algebra 1 and B.E.S.T. Geometry tests is based on **when students first participated in the assessment.** Students who took one of these assessments prior to the adoption of the new passing scores (i.e., prior to winter 2023) are eligible to use the APS for Algebra 1 for graduation purposes or the APS for geometry for scholar designation/CAP purposes. Students who participated in the B.E.S.T. Algebra 1 or B.E.S.T. Geometry assessments

for the first time in winter 2023 and beyond must obtain new passing scores for graduation/CAP and scholar designation/CAP purposes, respectively.

Students who took B.E.S.T. EOC for the first time in winter 2023 and onwards will *not* be APS eligible and will need to earn the passing score based on the new B.E.S.T. cut scores accordingly.

The APS that applies to a particular student will vary depending on student cohort, when students first participated in the assessment, and the test administration season. Not all historical APS have been included here. Information about the full range of prior APS scores is presented in *The B.E.S.T. and FAST 2023–2024 Administration Summative Scoring Specifications.*

7.1.5 Reporting Category Scores

In addition to overall scores, students also receive a performance classification on each of the reporting categories.

Reporting category scores will be calculated using MLE. These subscores, however, will be based only on the items contained in the reporting category.

Reporting Category Scores Using MLE Scoring

Theta scores for reporting categories will be estimated with the same MLE methods used to calculate overall theta scores.

Standard Error of Measurement for the Reporting Category

As with the total score, the standard error of measurement (SEM) for student *i* in the reporting category is

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

where

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

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.

Note that the calculation depends on the unique set of items that each student answers and their estimate of θ. Different students will have different SEM values even if they have the same raw score and/or theta estimate.

SE Transformation

SEs of the MLEs are similarly transformed to be placed onto the reporting scale. This transformation is defined as:

$$SEM_{SS} = a * SEM_{\theta_i},$$

where SEM_θ is the SE of the ability estimate on the θ scale and a is the slope of the scaling constants. The SEM is calculated on the basis of all item(s) that test takers saw for both complete and incomplete tests (Attempted = Y). The upper bound of the SEM is set to 1.5 on the theta metric. Any value larger than 1.5 is truncated at 1.5 on the theta metric for both overall theta scores and reporting category theta scores.

Subscale Performance Classification

Cambium Assessment, Inc. (CAI) will report relative strengths and weaknesses for each student at the reporting category (domain) level. The strengths and weaknesses will be computed relative to the student's reporting category scores. SEs will be based on the SE for the subscore.

Subscale-level classifications are computed to classify student achievement levels for each of the content standard subscales. For each subscale, the band is generally defined as a range extending one-and-a-half SEM below and one-and-a-half SEM above the proficient cut score. The rules surrounding classification are:

- If $(\theta_{tt} < \theta_{Proficient} - 1.5 * SEM)$, then performance is classified as Below Standard
- If $(\theta_{Proficient} - 1.5 * SEM \leq \theta_{tt} < \theta_{Proficient} + 1.5 * SEM)$, then performance is classified as At/Near Standard
- If $(\theta_{tt} \geq \theta_{Proficient} + 1.5 * SEM)$, then performance is classified as Above Standard

where $\theta_{Proficient}$ is the proficient cut score of the overall test, θ_{tt} is the student's score on a given reporting category, and SEM is the SEM for a given student's subscale theta estimate. Zero and perfect scores (as well as lowest observable scale score [LOSS] and highest observable scale score [HOSS]) would always be assigned *Below Standard* and *Above Standard*, respectively. Truncated scale scores use actual SEMs from the vertical scale theta estimates.

Appendix E, Distribution of Reporting Category Scores, summarizes the scores.

Target Scores

The target scores are used for standards-level strengths and weaknesses reports and are produced for the online tests only. Target scores are computed on the basis of responded items. If a test has unanswered items, then they are ignored.

Target scores will be computed in two ways: (1) target scores relative to a student's overall estimated ability (θ) and (2) target scores relative to the proficiency standard (Level 3 cut).

Target Scores Relative to Student's Overall Estimated Ability

The formula $p_{ij} = p(z_{ij} = 1)$ represents the probability that student *j* responds correctly to item *i* (z_{ij} represents the *j*th student's score on the *i*th item). For items with one score point, the 3PL IRT model is used to calculate the expected score on item *i* for student *j* with estimated ability $\hat{\theta}_j$:

$$E(z_{ij}) = c_j + (1 - c_j)\frac{exp\left(Da_j(\hat{\theta}_i - b_j)\right)}{1 + exp\left(Da_j(\hat{\theta}_i - b_j)\right)}$$

.

For items with two or more score points, the GPCM is used to calculate the expected score for student *j* with estimated ability $\hat{\theta}_j$ on an item *i* with a maximum possible score of m_i:

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{l\exp\left(\sum_{k=1}^{l} Da_i(\hat{\theta}_j - b_{i,k})\right)}{1 + \sum_{l=1}^{m_i}\exp\left(\sum_{k=1}^{l} Da_i(\hat{\theta}_j - b_{i,k})\right)}$$

.

For each item *i*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, *T*:

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score is computed by averaging individual student target scores for the target, across students of different abilities receiving different items measuring the same target at different levels of difficulty:

$$\bar{\delta}_{Tg} = \frac{1}{n_g}\sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g - 1)}\sum_{j \in g}\left(\delta_{jT} - \bar{\delta}_{Tg}\right)^2},$$

where n_g is the number of students who responded to any of the items that belongs to the target *T* for an aggregate unit *g*. If a student did not happen to see any items on a particular target, then the student is *not* included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target. Direct reporting of the statistic $\bar{\delta}_{Tg}$ is not suggested. Instead, reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target is recommended. In some cases, insufficient information will be available and that will be indicated as well.

Target-level strengths and weaknesses are reported as follows:

- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is worse than on the overall test.
- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is better than on the overall test.
- Otherwise, performance is similar to the overall test.
- If $se(\bar{\delta}_{Tg}) > 0.2$, then data are insufficient.

Target Scores Relative to On-Grade Standard (Level 3 cut)

The formula $p_{ij} = p(z_{ij} = 1)$ represents the probability that student *j* responds correctly to item *i* (z_{ij} represents the *j*th student's score on the *i*th item). Items with 1 score point are scored using the 3PL IRT model to calculate the expected score on item *i* for student *j* with $\theta_{Level\ 3\ cut}$:

$$E(z_{ij}) = c_j + (1 - c_j)\frac{exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}{1 + exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}$$

.

For items with 2 or more score points, the GPCM is used to calculate the expected score for student *j* with *Level 3 cut* on an item *i* with a maximum possible score of m_i:

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{l\exp(\sum_{k=1}^{l} Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^{l} Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}$$

.

For each item *i*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, *T*:

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score is computed by averaging individual student target scores for the target, across students of different abilities receiving different items measuring the same target at different levels of difficulty:

$$\bar{\delta}_{Tg} = \frac{1}{n_g}\sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)}\sum_{j \in g}(\delta_{jT} - \bar{\delta}_{Tg})^2},$$

where n_g is the number of students who responded to any of the items that belongs to the target *T* for an aggregate unit *g*. If a student did not happen to see any items on a particular target, then the student is *not* included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target. Direct reporting of the statistic δ_{Tg} is not suggested. Instead, reporting whether, in

the aggregate, a group of students performs better, worse, or as expected on this target is recommended. In some cases, insufficient information will be available and that will be indicated as well.

Target-level strengths and weaknesses are reported as follows:

- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is *below* the proficiency standard.
- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is *above* the proficiency standard.
- Otherwise, performance is *near* the proficiency standard.
- If $se(\bar{\delta}_{Tg}) > 0.2$, then data are insufficient.

8. QUALITY CONTROL FOR DATA, ANALYSES, SCORING, AND SCORE REPORTS

This chapter documents the data preparation and quality-control procedures used in analyses, scoring, and reporting.

8.1 DATA PREPARATION AND QUALITY CONTROL

Cambium Assessment, Inc.'s (CAI) quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

Before any analysis, data are first extracted from the Database of Record (DOR). Processing and exclusion rules are then applied to determine the final data file to be used in psychometric analyses.

After data files are finalized, they are passed to two psychometricians who used the files for all analyses independently. Each psychometrician independently implements classical and item response theory (IRT) analyses. The results from the two psychometricians (i.e., the IRTPRO output files) are formally compared. Any discrepancies are identified and resolved.

When all classical and IRT results match the findings from the independent analysts, the results are uploaded to the Secure File Transfer Protocol (SFTP) site for review. Independent replications are also completed by Florida Department of Education (FDOE) psychometricians, the Human Resources Research Organization (HumRRO), and the Buros Institute of Mental Measurements (Buros). Meetings are held with CAI, FDOE, the Test Development Center (TDC), HumRRO, and Buros to discuss classical statistics and IRT analyses when needed. Content experts from CAI and the TDC also review classical statistics and provide input. FDOE approves the results when there is replication and verification from all parties.

CAI uploads item statistics to the item bank after receiving final confirmation from all parties that the IRT statistics are accurate and that the items are appropriate for use in operational scoring.

8.2 SCORING QUALITY CONTROL

Before the operational testing window opens, CAI's scoring engine is tested to ensure that the maximum likelihood estimations (MLEs) the engine produced are accurate. This process is referred to as the *mock data* process. During mock data, CAI establishes all systems and simulates item response data as if real students responded to the test items. CAI then tests all programs and verifies all results before implementing the operational test. Simulated data are posted to the SFTP site for FDOE, HumRRO, and Buros to allow all parties to test their systems.

Once final operational item calibrations are complete and approved by FDOE, item parameters are uploaded to CAI's Item Tracking System, and student scores—including MLEs, scale scores, and reporting category raw scores—are generated via the scoring engine.

Like the verification process with calibrations, CAI, FDOE, and HumRRO independently check scores. FDOE approves scores only when there is three-way replication and verification.

9. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* https://www.aera.net/Portals/38/1999%20Standards_revised.pdf

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Scientific Software International.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. RR-91-47). Wiley. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1991.tb01414.x

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates. https://files.eric.ed.gov/fulltext/ED272577.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education/Praeger.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://hdl.handle.net/11299/115645

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*(2), 106–108. https://doi.org/10.1080/00031305.1986.10475369

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Tong, Y., Wu, S.-S., & Xu, M. (2008, March). *A comparison of pre-equating and post-equating using large-scale assessment data* [Presentation]. American Educational and Research Association Annual Meeting, New York, NY, United States.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. RR-12-08). Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1109842.pdf