State of Florida

Benchmarks for Excellent Student Thinking (B.E.S.T.)

2022–2023

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

Major contributors to this technical report include the following staff from CAI: Dr. Gary Phillips, Dr. Stephan Ahadi, Dr. Myvan Bui, Dr. Yuan Hong, Dr. Hyesuk Jang, Dr. Sherry Li, Dr. Peter Diao, Melissa Boyanton, Matt Gordon, and Zoe Dai. The major contributors from the FDOE are as follows: Vince Verges, Susie Lee, Racquel Harrell, Sally Donnelly, Shakia Johnson, Kristina Lamb, Jenny Black, Dr. Esra Kocyigit, Dr. Salih Binici, Wenyi Li, Jielin Ming, Saeyan Yun, and Yiting Yao.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Beginning with the 2022–2023 school year, Florida's statewide, standardized assessments in reading, writing, and mathematics were aligned with the Benchmarks for Excellent Student Thinking (B.E.S.T.). Beginning in fall 2022, the Florida assessments are referred to as the FAST (Florida Assessment of Student Thinking) and B.E.S.T. The FAST is administered as a progress monitoring assessment and includes Voluntary Prekindergarten (VPK) through grade 10 English language arts (ELA) and VPK through grade 8 mathematics assessments. B.E.S.T. assessments that are not part of the FAST progress monitoring program include grades 4–10 writing and end-of-course (EOC) assessments in Algebra 1 and Geometry. This technical report describes the FAST assessments for grades 3–10 ELA and grades 3–8 mathematics, and the B.E.S.T. assessments. The details of the 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*.

The *Florida Benchmarks for Excellent Student Thinking 2022–2023 Technical Report* is provided to document all methods used in test construction, outline psychometric properties of the tests, summarize student results, and document 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 and 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.
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, the Florida Department of Education (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 the FDOE in support of any requests to further investigate test quality, validity, or other issues as identified. As of now, there are no reports to include in this volume for 2022–2023.

1.1 PURPOSE AND INTENDED USES OF THE FAST AND B.E.S.T. 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 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 FAST and B.E.S.T. standards and test blueprints can be found in Volume 2, Test Development. Additional verification of content validity can also be found in Volume 4, Evidence of Reliability and Validity. The documentation about the comparability of online and accommodated tests can be found in Volume 4, Evidence of Reliability and Validity.

The Florida FAST and B.E.S.T. 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 can be found in Volume 4, Evidence of Reliability and Validity.

The Florida FAST and B.E.S.T. assessments are criterion-referenced tests intended to measure whether students have progressed on the B.E.S.T. standards in ELA and mathematics. The Florida FAST and B.E.S.T. assessment standards and test blueprints are discussed in Volume 2, Test Development.

Table 1 outlines the required uses of the FAST and B.E.S.T. assessments.

Table 1: Required Uses and Citations for the Florida FAST and B.E.S.T. Assessments

Assessment	Assessment Citation	Required Use	Required Use Citation
Statewide Assessment Program	s. 1008.22, F.S. Rule 1.09422, F.A.C. Rule 1.0943, F.A.C Rule 1.09432, F.A.C.	Third Grade Retention; Student Progression; Remedial Instruction; Reporting Requirements	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.
		High School Standard Diploma	s. 1003.4282, F.S.
		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.

Assessment	Assessment Citation	Required Use	Required Use Citation
		District Grades	s. 1008.34, F.S.
		Differentiated Accountability	s. 1008.33, F.S. Rule 6A-1.099811, F.A.C.
		Opportunity Scholarship Hope Scholarship Florida Tax Credit Scholarship Family Empowerment Scholarship	s. 1002.38, F.S. s. 1002.40, F.S. s. 1002.395, F.S. s. 1002.394
		The New Worlds Reading Initiative	s. 1003.485

Appendix F 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, the FDOE began transitioning from the Florida Standards Assessment (FSA) to the FAST assessment, in accordance with changes to state statute.

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., Progress Monitoring (PM) 3) 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 at grades 3–10 and in mathematics at 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, the FDOE will start reporting scores on the new FAST scale.

FAST is administered as a progress monitoring assessment. Students participate three times per year: once at the beginning of the year (PM1, August 15–October 7, 2022), once in the middle of the year (PM2, December 5, 2022–January 27, 2023), and once at the end of the year (PM3, May 1–June 2, 2023).

- 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 (FDOE, 2022).
- 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 (FDOE, 2022).

- PM3 produces summative scores that will accurately measure student mastery of the B.E.S.T. standards at the end of the school year. While PM1 and PM2 are for informational purposes only, PM3 is used for school accountability in Grade 3 and higher beginning with the 2023–2024 school year (FDOE, 2022). Assessments in Grades PreK-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 progress monitoring program.

For PM1 and PM2, the FAST assessments in ELA and mathematics were administered as online computer-adaptive assessments. For mathematics grades 3–8 (as well as EOC Algebra 1 and Geometry), the calibration plan allowed for continued adaptive administration for PM3. For ELA, however, while the adaptive algorithm continued to be used to deliver assessments online, the adaptive weights were set to zero so that item selection was independent of item difficulty, allowing for a representative sample of student responses to each item with respect to the ability distribution. In addition, a representative sample of students in grades 4–10 responded to a single writing prompt as part of a field test that did not contribute to a summative score in 2023.

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 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 set in 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. Geometry EOC for 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.

The transition to the FAST and B.E.S.T. assessments is 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 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.

Psychometricians and content experts from CAI, the 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 the 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 both perspectives were considered for inclusion on Florida's assessments and for constructing Florida's vertical scale.

In 2014 and going forward 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 an accommodation 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 grade 3 and grade 10 Florida Comprehensive Assessment Test (FCAT) 2.0 reading and Next Generation Sunshine State Standards (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 were 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 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.

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 was 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 school year (SY) 2019–2020. Specifically, as of the cancellation, only grade 10 ELA writing and reading Retake and Algebra 1 EOC Retake were completed, while the spring 2020 regular assessments were canceled, including grades 3–10 ELA reading, grades 4–10 ELA writing, grades 3–8 mathematics, Algebra 1, and Geometry. Because of the cancellation, no empirical data that depend 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 the prior year for processes that were uncompleted prior to the cancellation, whereas results were reported based on spring 2020 for processes that were completed before the cancellation.

Developments in 2021

Because of the cancellation of the spring 2020 regular assessments, the FDOE could not field test numerous newly-developed items across all subjects in 2020 and so 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 around 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*. The 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 standards, the 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, the 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 were 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 and beyond, the FDOE will start 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 progress monitory 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.

1.3 Participants in the Development and Analysis of the FAST and B.E.S.T. Assessments

The FDOE manages the FAST and B.E.S.T. program with the assistance of several participants, including multiple offices within the FDOE, Florida educators, a Technical Advisory Committee (TAC), and vendors. The 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 (FDOE)

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 (TDC): Funded by the 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 develop the academic standards and clarify how these standards will be assessed, aid in test design, and review test items and passages.

Technical Advisory Committee

The 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 was responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the FAST and B.E.S.T. 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 various federal and state agencies, corporate and not-for-profit organizations and foundations. For the FAST and B.E.S.T. assessments, HumRRO conducts independent checks on the equating and linking activities and reports its findings directly to the FDOE. HumRRO also provides consultative services to the 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. For the 2022–2023 FAST and B.E.S.T., Buros provided independent operational checks on the equating procedures, writing handscoring activities, and the scanning and editing 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 FAST and B.E.S.T. 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 onsite monitoring of test administration in samples of school districts.

1.4 Available Test Formats and Special Versions

For the summative assessment, students in grades 3–10 reading, grades 3–8 mathematics, and students taking Algebra 1 and Geometry EOC assessments are administered online computer-adaptive tests during each spring, which 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 indicated 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 FAST and B.E.S.T. test 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, or the mathematics EOC tests in the FAST and B.E.S.T. assessments administered in the spring. Retake administrations for the EOC assessments occur in the summer, fall, and winter, and grade 10 ELA retake administrations occur only in the fall and spring.

Tables 2–4 show the number of students who were tested and the number of students who were reported in the spring 2023 FAST and B.E.S.T. by grade and subject area for online tests. Information for students who took accommodated forms are available in this volume, Section 3.2, FAST and B.E.S.T. 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 5–7 present the percentages of students in each performance level for grades and subjects that were reported for the spring 2023 FAST and B.E.S.T. Please refer to Appendix D for descriptive statistics on the scale score distributions across all students and subgroups.

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

Mathematics			ELA Reading		
Grade/Test	Number Tested	Number Reported	Grade	Number Tested	Number Reported
3	221,144	221,088	3	221,662	221,585
4	197,888	197,851	4	201,164	201,109

Mathematics			ELA Reading		
5	206,364	206,300	5	207,502	207,452
6	211,163	210,954	6	216,412	216,236
7	151,498	151,197	7	209,599	209,397
8	170,944	170,630	8	215,735	215,503
Algebra 1	254,244	253,628	9	223,090	222,786
Geometry	222,829	222,493	10	262,343	261,928

Table 3: Number of Students Participating in B.E.S.T. Assessments (PM1)

Mathematics			ELA Reading		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**
3	219,450	219,347	3	219,969	219,841
4	192,505	192,378	4	195,864	195,768
5	203,166	203,088	5	204,214	204,144
6	208,249	207,840	6	213,422	213,251
7	155,695	155,002	7	206,420	206,223
8	159,208	158,853	8	211,892	211,765
			9	219,274	218,964
			10	210,434	210,155

Table 4: Number of Students Participating in B.E.S.T. Assessments (PM2)

Mathematics			ELA Reading		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**
3	219,465	219,394	3	220,034	219,929
4	195,463	195,415	4	198,894	198,848
5	204,701	204,613	5	205,671	205,641
6	208,925	208,747	6	214,230	214,095
7	149,712	149,007	7	207,407	207,244
8	168,720	168,294	8	212,647	212,490
			9	220,932	220,610
			10	210,875	210,607

Table 5: Percentage of Students Taking Operational Forms by Performance Level (PM3)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	20.0	22.8	21.9	24.7	10.6
	4	23.3	18.5	20.2	26.4	11.6
	5	24.6	24.0	21.6	16.7	13.1
	6	24.2	27.5	18.2	20.2	9.9
	7	33.4	22.2	21.6	13.9	8.9
	8	30.7	28.0	20.7	12.7	7.9
ELA Reading	3	25.1	23.4	21.8	17.5	12.1
	4	25.3	22.5	21.3	20.1	10.8
	5	24.3	25.9	18.9	20.6	10.3
	6	25.7	24.6	21.8	17.9	10.0
	7	26.3	26.2	17.6	20.2	9.6
	8	25.4	26.4	21.9	14.6	11.6
	9	25.7	26.1	20.6	17.2	10.3
	10	26.1	26.4	19.9	17.5	10.1
EOC	Algebra 1	24.1	25.3	36.5	5.6	8.5
	Geometry	29.8	24.6	26.0	9.0	10.6

Table 6: Percentage of Students Taking Operational Forms by Performance Level (PM1)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	65.6	26.8	6.1	1.3	0.2
	4	72.5	19.1	5.7	2.5	0.2
	5	60.3	27.1	9.2	2.7	0.6
	6	54.7	30.9	9.9	4.1	0.4
	7	60.3	22.3	12.9	3.6	0.9
	8	58.0	32.1	7.1	2.2	0.6
ELA Reading	3	49.4	26.6	14.2	7.1	2.6
	4	46.8	25.6	16.0	9.4	2.2
	5	40.6	30.2	15.5	10.9	2.7
	6	32.8	27.9	20.7	14.0	4.6
	7	36.3	28.7	16.0	14.3	4.8
	8	37.5	30.8	18.1	8.7	4.9
	9	35.6	29.6	18.1	11.7	4.9
	10	37.9	26.8	16.7	12.4	6.1

Table 7: Percentage of Students Taking Operational Forms by Performance Level (PM2)

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	40.5	32.3	18.1	7.9	1.2
	4	48.4	27.0	15.3	8.1	1.2
	5	41.4	30.9	17.5	7.7	2.4
	6	36.1	33.8	17.2	10.8	2.1
	7	45.8	25.4	18.2	7.9	2.7
	8	38.0	31.1	18.2	8.7	4.1
ELA Reading	3	36.7	26.4	18.8	12.0	6.2
	4	36.2	24.8	18.9	14.7	5.4
	5	33.5	30.0	17.0	14.6	5.0
	6	30.8	27.0	21.1	14.8	6.3
	7	32.9	27.9	16.7	16.2	6.3
	8	33.5	28.7	19.8	11.0	6.9
	9	33.1	28.3	18.8	13.3	6.5
	10	36.2	26.1	16.9	13.4	7.5

1.6 DEMOGRAPHICS OF TESTED POPULATION

Tables 8–14 present the distribution of students, in counts and in percentages, who participated in the spring administration of the 2022–2023 FAST and B.E.S.T. by grade and subject. The numbers presented here 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 3.2, FAST and B.E.S.T. Accommodations. The subgroups reported are gender, ethnicity, Students with Disabilities (SWD), and ELL. Section 1.2, Testing Accommodations of Volume 5 of this technical report provides explicit definitions for the two major subgroups to which accommodations are available: ELL and SWD. Students offered accommodations may choose to not use the accommodation.

Table 8: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM3)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,726	107,047	112,679	45,617	81,631	75,884	27,688	35,974
	%	100	48.72	51.28	20.76	37.15	34.54	12.60	16.37
4	*N*	196,727	96,544	100,183	39,057	72,319	70,133	24,170	26,005
	%	100	49.08	50.92	19.85	36.76	35.65	12.29	13.22
5	*N*	205,203	101,105	104,098	41,369	76,450	71,814	26,523	23,756
	%	100	49.27	50.73	20.16	37.26	35.00	12.93	11.58
6	*N*	210,451	102,829	107,622	44,800	80,458	70,416	30,272	21,115
	%	100	48.86	51.14	21.29	38.23	33.46	14.38	10.03
7	*N*	150,783	73,881	76,902	35,122	58,952	47,727	25,140	17,807
	%	100	49.00	51.00	23.29	39.10	31.65	16.67	11.81
8	*N*	170,201	83,106	87,095	38,722	66,178	54,175	25,429	17,342
	%	100	48.83	51.17	22.75	38.88	31.83	14.94	10.19

Table 9: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM3)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	220,208	107,258	112,950	45,762	81,797	76,016	27,809	36,052
	%	100	48.71	51.29	20.78	37.15	34.52	12.63	16.37
4	*N*	199,966	98,023	101,943	39,501	73,278	71,566	24,263	26,128
	%	100	49.02	50.98	19.75	36.65	35.79	12.13	13.07
5	*N*	206,344	101,597	104,747	41,386	76,662	72,516	26,594	23,695
	%	100	49.24	50.76	20.06	37.15	35.14	12.89	11.48
6	*N*	215,726	105,191	110,535	45,445	81,842	72,774	30,498	21,250
	%	100	48.76	51.24	21.07	37.94	33.73	14.14	9.85
7	*N*	208,913	102,615	106,298	42,718	78,688	72,230	27,097	18,802
	%	100	49.12	50.88	20.45	37.67	34.57	12.97	9.00
8	*N*	215,002	105,343	109,659	44,454	81,188	73,854	26,901	17,649
	%	100	49.00	51.00	20.68	37.76	34.35	12.51	8.21
9	*N*	222,275	109,526	112,749	45,739	83,007	77,694	26,256	17,621
	%	100	49.27	50.73	20.58	37.34	34.95	11.81	7.93
10	*N*	261,288	127,346	133,942	58,300	99,674	86,028	26,260	25,146
	%	100	48.74	51.26	22.31	38.15	32.92	10.05	9.62

Table 10: Distribution of Demographic Characteristics of Tested Population, Mathematics EOC

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
Algebra 1	*N*	253,009	123,545	129,464	54,057	96,993	84,553	28,564	23,054
	%	100	48.83	51.17	21.37	38.34	33.42	11.29	9.11
Geometry	*N*	221,937	109,228	112,709	45,312	82,135	78,762	25,074	15,818
	%	100	49.22	50.78	20.42	37.01	35.49	11.30	7.13

Table 11: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM1)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,347	106,730	112,617	46,225	80,049	76,440	24,256	33,638
	%	100	48.66	51.34	21.07	36.49	34.85	11.06	15.34
4	*N*	192,378	94,438	97,940	38,113	69,199	69,865	22,560	23,856
	%	100	49.09	50.91	19.81	35.97	36.32	11.73	12.40
5	*N*	203,088	100,050	103,038	41,240	74,170	72,022	25,454	22,669
	%	100	49.26	50.74	20.31	36.52	35.46	12.53	11.16
6	*N*	207,840	101,583	106,257	44,348	77,756	71,037	29,759	21,382
	%	100	48.88	51.12	21.34	37.41	34.18	14.32	10.29
7	*N*	155,002	75,866	79,136	35,481	58,656	51,328	25,117	15,555
	%	100	48.95	51.05	22.89	37.84	33.11	16.20	10.04
8	*N*	158,853	77,632	81,221	37,390	60,373	50,766	24,715	15,473
	%	100	48.87	51.13	23.54	38.01	31.96	15.56	9.74

Table 12: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM1)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,841	106,948	112,893	46,369	80,192	76,600	24,356	33,678
	%	100	48.65	51.35	21.09	36.48	34.84	11.08	15.32
4	*N*	195,768	96,018	99,750	38,514	70,230	71,415	22,684	24,051
	%	100	49.05	50.95	19.67	35.87	36.48	11.59	12.29
5	*N*	204,144	100,468	103,676	41,236	74,325	72,723	25,517	22,551
	%	100	49.21	50.79	20.20	36.41	35.62	12.50	11.05
6	*N*	213,251	103,990	109,261	44,942	79,321	73,421	29,981	21,464
	%	100	48.76	51.24	21.07	37.20	34.43	14.06	10.07

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
7	*N*	206,223	101,149	105,074	42,190	75,931	72,946	26,951	16,780
	%	100	49.05	50.95	20.46	36.82	35.37	13.07	8.14
8	*N*	211,765	103,653	108,112	43,719	78,410	74,357	26,431	16,047
	%	100	48.95	51.05	20.65	37.03	35.11	12.48	7.58
9	*N*	218,964	108,127	110,837	45,031	79,886	78,453	25,829	15,215
	%	100	49.38	50.62	20.57	36.48	35.83	11.80	6.95
10	*N*	210,155	104,608	105,547	41,848	76,683	76,782	21,947	13,826
	%	100	49.78	50.22	19.91	36.49	36.54	10.44	6.58

Table 13: Distribution of Demographic Characteristics of Tested Population, Mathematics (PM2)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,394	106,788	112,606	45,814	80,718	76,462	26,181	34,675
	%	100	48.67	51.33	20.88	36.79	34.85	11.93	15.80
4	*N*	195,415	95,859	99,556	38,899	71,061	70,497	23,982	24,593
	%	100	49.05	50.95	19.91	36.36	36.08	12.27	12.59
5	*N*	204,613	100,716	103,897	41,458	75,527	72,248	26,397	22,631
	%	100	49.22	50.78	20.26	36.91	35.31	12.90	11.06
6	*N*	208,747	101,828	106,919	44,484	78,918	70,815	30,319	19,956
	%	100	48.78	51.22	21.31	37.81	33.92	14.52	9.56
7	*N*	149,007	72,873	76,134	34,907	57,168	48,022	25,284	16,270
	%	100	48.91	51.09	23.43	38.37	32.23	16.97	10.92
8	*N*	168,294	82,190	86,104	38,449	64,573	54,473	25,343	16,138
	%	100	48.84	51.16	22.85	38.37	32.37	15.06	9.59

Table 14: Distribution of Demographic Characteristics of Tested Population, ELA Reading (PM2)

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	219,929	107,047	112,881	45,942	80,914	76,629	26,280	34,762
	%	100	48.67	51.33	20.89	36.79	34.84	11.95	15.81
4	*N*	198,848	97,421	101,427	39,375	72,112	71,980	24,104	24,706
	%	100	48.99	51.01	19.80	36.26	36.20	12.12	12.42
5	*N*	205,641	101,146	104,495	41,433	75,696	72,927	26,486	22,567
	%	100	49.19	50.81	20.15	36.81	35.46	12.88	10.97

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
6	*N*	214,095	104,291	109,804	45,108	80,415	73,193	30,550	20,013
	%	100	48.71	51.29	21.07	37.56	34.19	14.27	9.35
7	*N*	207,244	101,733	105,511	42,390	77,083	72,712	27,270	17,276
	%	100	49.09	50.91	20.45	37.19	35.09	13.16	8.34
8	*N*	212,490	103,966	108,524	43,820	79,491	74,019	26,707	16,523
	%	100	48.93	51.07	20.62	37.41	34.83	12.57	7.78
9	*N*	220,610	108,677	111,933	45,300	81,624	78,151	26,365	16,328
	%	100	49.26	50.74	20.53	37.00	35.42	11.95	7.40
10	*N*	210,607	104,617	105,990	42,185	77,555	76,219	21,961	14,271
	%	100	49.67	50.33	20.03	36.82	36.19	10.43	6.78

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.

During the 2022–2023 school year, the Florida Department of Education (FDOE) began transitioning from the Florida Standards Assessment (FSA) to the Florida Assessment of Student Thinking (FAST) and B.E.S.T. FAST refers to the new Coordinated Screening and Progress Monitoring (CSPM) System assessments, which are aligned to the B.E.S.T. standards. FAST assessments include Voluntary Prekindergarten (VPK) through grade 10 English language arts (ELA) and VPK through grade 8 mathematics. End-of-Course (EOC) assessments are not part of FAST but are also aligned to B.E.S.T.

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) Adds grades 9 and 10 to the ELA assessments administered as part of the CSPM system.
2) 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.
3) Requires the results for the FAST ELA and mathematics assessments be available no later than May 31 each year beginning with the 2023–2024 school year.

Per s. 1008.25(8), F.S., FAST assessments will be administered three times per year, the first (Progress Monitoring [PM]1) 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 questions incorrectly. Each PM event is tied to a blueprint for the full grade-level content. Many of the same computer-based item types that students are already familiar with will be used on the FAST assessments. As part of FAST, writing will be administered in grades 4–10. Writing will be reported separately from reading and will not contribute to an overall ELA score. FAST writing will be computer-based in all assessed grades, and prompts will be in response to text. In 2022–2023, writing was administered as a field test to a representative sample of Florida students during the spring 2023 administration.

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 will be used for informational purposes only and will not be used for accountability. PM3 will be 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 will be applied. Tests will be computer adaptive through the Test Delivery System (TDS) secure browser. As with FSA, a Level 3 achievement level on the FAST assessments will be considered passing. However, SB 1048 revised the definition of a Level 3 score from a “satisfactory performance” to “grade-level performance.”

3. SUMMARY OF OPERATIONAL PROCEDURES

This chapter summarizes the spring administration procedures, the number of students taking accommodated tests, and students' performance levels based on the spring 2023 administration.

3.1 SPRING ADMINISTRATION PROCEDURES

Table 15 shows the schedule for the spring administration of the 2022–2023 B.E.S.T. assessments, broken down by testing window and subject area.

Table 15: Testing Windows by Subject Area

Assessment	Testing Window
Algebra 1 Retake	September 12–30, 2022 February 20–March 10, 2023
ELA Retake Reading and Writing	September 12–30, 2022 February 20–March 10, 2023
Grades 3–10 FAST ELA Reading Grades 3–8 FAST Mathematics	**First Administration ("PM1"):** August 15–September 30, 2022 **Second Administration ("PM2"):** December 5, 2022–January 27, 2023 **Third Administration ("PM3"):** May 1–June 2, 2023
Algebra 1 and Geometry	November 28–December 16, 2022 January 9–20, 2023 May 1–26, 2023 July 10–21, 2023

In accordance with state law, students were required to participate in the spring assessment, and all testing took place during the designated testing window. The Florida Assessment of Student Thinking (FAST) and B.E.S.T. 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. 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 FAST and B.E.S.T. administration included the district assessment coordinators (DACs), school administrators, 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 can be found in Volume 5 of the *Florida Benchmarks for Excellent Student Thinking 2022–2023 Technical Report*.

A secure browser developed by CAI (CAI Secure Browser) was required to access the online FAST and B.E.S.T. The browser provided a secure environment for student testing by disabling the hot keys, 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 of this technical report.

Students were able to participate in FAST and B.E.S.T. online tests via multiple platforms, such as Windows, Chrome, Mac, and iPad. Prior to the test administration, a series of user acceptance testing is performed on all platforms on which FAST and B.E.S.T. 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), Cambium Assessment, Inc. (CAI) conducted a device comparability study to provide evidence of comparability of the Florida Statewide Assessments scores across devices. This study can be found in Volume 7 of the *Florida Standards Assessments 2019–2020 Technical Report.*

Prior to test administration, a series of user acceptance testing is performed on all approved platforms to ensure that items are rendered as expected and have similar appearance across platforms to minimize potential device effects. A rigorous review is in place to ensure that the content of the items on accommodated tests matches the content of the items as administered online (i.e., wording, graphics, paragraph breaks, and option order).

3.2 FAST AND B.E.S.T. 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 of Volume 5: Testing Accommodations of this technical report. Also, please refer to Section 6.2.2 of this volume for the details of the accommodated form construction, in addition to Appendix C of Volume 1.

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 of Volume 4 of this technical report.

The number of students who took the accommodated version of the 2022–2023 FAST and B.E.S.T. varied between 414 and 1,377 across grades and subjects, as shown in Table 16. In the 2022–2023 administration, accommodations were only available for the Progress Monitoring (PM) 3/spring summative assessments. Accommodations for the PM1 and PM2 will be available from 2023–2024 onwards.

Table 16: Counts of Accommodated Assessments by Grades and Subjects

Subject	Grade	Spring 2023
Mathematics	3	1,362
	4	1,124
	5	1,097
	6	503
	7	414
	8	429
ELA Reading	3	1,377
	4	1,143
	5	1,108
	6	510
	7	484
	8	501
	9	511
	10	640
EOC	Algebra 1	619
	Geometry	556

Tables 17–19 present the distribution of accommodated students, in counts and in percentages, who participated in the spring administration of the 2022–2023 FAST and B.E.S.T. by grade and subject. The subgroups reported are gender, ethnicity, Students with Disabilities (SWD), and English language learners (ELLs). During this year, accommodated forms were not available for other test administrations. From 2023–2024 onwards, accommodated forms will be available for all test administrations.

Table 17: Distribution of Demographic Characteristics of Tested Accommodated Population, Mathematics

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	1,362	490	872	320	645	325	1170	337
	%	100	35.98	64.02	23.49	47.36	23.86	85.90	24.74
4	*N*	1,124	418	706	257	546	262	978	226
	%	100	37.19	62.81	22.86	48.58	23.31	87.01	20.11
5	*N*	1,097	437	660	258	541	248	939	180
	%	100	39.84	60.16	23.52	49.32	22.61	85.60	16.41
6	*N*	503	212	291	123	250	109	408	55
	%	100	42.15	57.85	24.45	49.7	21.67	81.11	10.93
7	*N*	414	170	244	119	169	113	352	22
	%	100	41.06	58.94	28.74	40.82	27.29	85.02	5.31
8	*N*	429	183	246	129	178	113	331	14
	%	100	42.66	57.34	30.07	41.49	26.34	77.16	3.26

Table 18: Distribution of Demographic Characteristics of Tested Accommodated Population, ELA Reading

Grade	Group	All Students	Female	Male	African American	Hispanic	White	SWD	ELL
3	*N*	1,377	501	876	322	656	326	1,182	349
	%	100	36.38	63.62	23.38	47.64	23.67	85.84	25.34
4	*N*	1,143	422	721	256	552	276	994	229
	%	100	36.92	63.08	22.40	48.29	24.15	86.96	20.03
5	*N*	1,108	441	667	262	553	243	953	186
	%	100	39.8	60.20	23.65	49.91	21.93	86.01	16.79
6	*N*	510	212	298	125	255	110	417	55
	%	100	41.57	58.43	24.51	50.00	21.57	81.76	10.78
7	*N*	484	207	277	123	203	145	396	23
	%	100	42.77	57.23	25.41	41.94	29.96	81.82	4.75
8	*N*	501	211	290	131	213	144	364	18
	%	100	42.12	57.88	26.15	42.51	28.74	72.65	3.59
9	*N*	511	220	291	126	228	142	360	12
	%	100	43.05	56.95	24.66	44.62	27.79	70.45	2.35
10	*N*	640	254	386	179	238	197	392	12
	%	100	39.69	60.31	27.97	37.19	30.78	61.25	1.88

Table 19: 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*	619	241	378	167	252	176	411	14
	%	100	38.93	61.07	26.98	40.71	28.43	66.40	2.26
Geometry	*N*	556	218	338	135	211	181	373	8
	%	100	39.21	60.79	24.28	37.95	32.55	67.09	1.44

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. For eligible students participating in computer-based assessments, accommodations such as masking, text-to-speech, and regular or large print passage booklets were made available. Students could use these accommodations only as dictated on their IEPs or Section 504 Plans. Additional accommodations guidelines can be found in Volume 5 of this technical report.

4. ITEM BANK MAINTENANCE

This chapter describes the item bank in terms of review of operational and field-test items in spring 2023.

4.1 OVERVIEW OF ITEM DEVELOPMENT

Complete details of the item development plan for Cambium Assessment, Inc. (CAI) are provided in the *Florida Benchmarks for Excellent Student Thinking 2022–2023 Technical Report*, Volume 2, Test Development. The test development phase includes a variety of activities designed to produce high-quality assessments that accurately measure student skills and abilities with respect 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 randomly distributed throughout the test. Operational items are used to calculate student scores. EFT items are nonscored items and are used to populate the Florida Assessment of Student Thinking (FAST) and B.E.S.T. test bank for future operational use.

The accommodated versions of online assessments 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 2023 operational calibration, both operational and field-test items were reviewed based on their performance during the spring administration. Before the spring administration, a *Calibration and Scoring Specifications* document is created by CAI, the Florida Department of Education (FDOE), and the Human Resources Research Organization (HumRRO) and reviewed by the Technical Advisory Committee (TAC). The specifications document outlines 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 post results to a secure location for review. Items are reviewed, with special attention being paid to items flagged based on the statistical rules described in the *Calibration and Scoring Specifications* 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 are to be removed from scoring.

4.3 FIELD TESTING

The FAST and B.E.S.T. 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 6–12 field-test items are assigned to students randomly, as described in the following paragraphs.

Table 20 shows the number of mathematics and mathematics EOC items by grade and item type that are included in spring 2023 CAT for field testing. Table 21 shows the number of Reading items by grade and item type that were included on spring 2023 Computer-Adaptive Test (CAT) 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 removal of items during calibrations. The item types are described in Section 3.2 of Volume 2 of the *Florida Benchmarks for Excellent Student Thinking 2022–2023 Technical Report.*

Table 20: 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	99	71	122	120	57	26	36	67
ETC	37	18	32	27	13	8	7	22
GI	0	0	1	0	0	1	6	10
HT	0	0	0	0	0	0	0	1
MC	98	68	128	85	67	32	23	47
MI	31	14	18	14	3	2	2	7
MS	36	25	20	18	5	6	1	5
Multi	2	2	4	4	8	0	9	10
Total Number of Items	303	198	325	268	153	75	84	169

Table 21: 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	17	42	49	31	31	33	33	22
HT	1	4	2	3	6	2	12	7
MC	147	266	245	206	211	175	222	169
MI	11	11	19	14	20	8	15	9
MS	21	54	48	36	17	20	22	19
Two-Part HI	1	1	0	0	0	0	1	0
Total Number of Items	198	378	363	290	285	238	305	226

A detailed overview of the development and review process for new items is provided in the *Florida Benchmarks for Excellent Student Thinking 2022–2023 Technical Report*, Volume 2, Test Development. Additional details on development and maintenance of the item pool are also given in the same volume.

Writing Independent Field Test

In 2023, writing was administered as an independent field test (IFT) to a sample of Florida students.

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 scored by two human raters based on the B.E.S.T. rubric.

Table 22 shows the number of prompts that were field-tested and the total number of students.

Table 22: The 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 *j*th 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

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

we obtain the total number of schools required for sampling to be

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

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

In order to use a proportionate stratification method, we first identified the proportion of schools across the state within stratum l using the number of students $l_{n,g}$ as

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

and then within each stratum $m_{l,g} = P_{l,g}M_g$ schools were sampled. 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 (alternating ascending and descending) order by the implicit strata and $m_{l,g}$ schools were selected systematically from this sorted list.

In hierarchical serpentine sorting, within a stratum, we sorted the first variable in ascending order. Then, within the first level of the first variable, we sorted the second variable in ascending order. Within the second level of the first variable, we sorted the second variable in descending order. We continued to apply this procedure to all levels and all variables so that it is 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 Reading Test:** This variable is intended to capture the ability of students across the population.
- **School size:** This variable is intended to ensure that schools of various sizes are represented in the sample.
- **Curriculum Group** (Standard, English language learner [ELL], Exceptional Student Education [ESE])
- **Gender** (Male and Female)
- **Percent Ethnicity:** The following demographic variables were used:
    - Percent White
    - Percent African American
    - Percent Hispanic
    - 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) stats were derived from the Progress Monitoring (PM) 3 administrations, after students had gone through a year worth of instruction and had 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. The thresholds used to flag an item for further review based on classical item statistics are presented in Table 23.

Table 23: Thresholds for Flagging Items in Classical Item Analysis

Rule	Flagging Criteria	Rationale
p-value	For 1-point items, flag if $p < 0.20$ or $p > 0.90$	Items are 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	Non-discriminating item
Distractor *p*-value	Flag if the *p*-value for the distractor is larger than the *p*-value for the key	Potentially problematic item
Correlation with test for distractors	Flag if correlation for any distractor is larger than correlation for 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 differentiated between those 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 can be found 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 points). Conventional item *p*-values and IRT parameters are summarized in Section 6.4, Results of Calibrations, of this volume. The *p*-values for operational items can be found in Appendix A, Operational Item Statistics, of this volume.

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) document 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, Florida Assessment of Student Thinking (FAST) and B.E.S.T. 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 bias. 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, it is not the item that exhibits bias but rather the instruction. 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. For field-test items, we performed DIF analyses using IRT ability estimates as the ability-matching variable during field-test calibrations. The corresponding scores are 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 24. 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 based on 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 24: 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	MH_{X^2} is significant and $\|SMD\| / \|SD\| > .25$.
B	MH_{X^2} is significant and $.17 < \|SMD\| / \|SD\| \leq .25$.
A	MH_{X^2} is not significant or $\|SMD\| / \|SD\| \leq .17$.

DIF summary tables can be found 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% of mathematics, mathematics end-of-course (EOC), and English language arts (ELA) items 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 6.3, IRT Item Summaries.

6. ITEM CALIBRATION AND SCALING

Item response theory (IRT) was used to calibrate all items and derive scores for all Florida Assessment of Student Thinking (FAST) and B.E.S.T. tests. IRT is a general framework that models test responses resulting from an interaction 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 it assumes 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, all non-MC FAST and B.E.S.T. items undergo additional reviews by content and psychometric teams 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 being conducted. Pre-equating occurs before the operational testing, whereas post-equating happens after the operational testing. Both are extensively used in K–12 large-scale assessment programs (Tong, Wu, & Xu, 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, Wu, & Xu, 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, Wu, & Xu, 2008). New item statistics are collected each year when items are used, thus assuming the statistical characteristics of the item may change when the ability of tested population changes.

In prior years, Florida used the pre-equated method for retake tests and post-equating for non-retake administrations. For the 2023 spring administration and future test administrations, due to the transition to computer-adaptive testing, the pre-equating method became necessary for all tests.

6.1 ITEM RESPONSE THEORY METHODS

The generalized approach to item calibration was to use the 3PL model (Lord & Novick, 1968) for MC items; to use the 2PL model (Lord & Novick, 1968) for binary items that assume no guessing; and to 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 since 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

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

where $\mathrm{P_i}(\theta_\mathrm{j})$ is the probability of test taker *j* answering item *i* correctly, $\mathrm{c_i}$ is the lower asymptote of the item response curve (the pseudo-guessing parameter), $\mathrm{b_i}$ is the location parameter, $\mathrm{a_i}$ 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 ($\mathrm{c_i}$) 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

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

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

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 ON-GRADE CALIBRATION

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 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, the 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 for calibrating the new mathematics assessments that was based on the administration of discrete items, and a second approach for English language arts (ELA) that required 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 for 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/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, the 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 also 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 the 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 are 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. To calibrate the ELA pool to the new FAST scale, therefore, 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 blueprint match for each test administration, but item selection proceeded independently of item difficulty. Each passage set, and thus each item in the current summative pool, was therefore administered to a random and representative sample of Florida students, supporting calibration of items to the new progress monitoring 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 robustly model the breadth and depth of the measurement model FAST assesses, and that were consistent with the originally proposed approach. This approach supported robust, adaptive test administration of the three-opportunity progress monitoring assessments in the 2023–2024 school year. This approach also supported the calibration of the new FAST writing prompts, since linkage of writing items must be achieved 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 negative biserial or sample size of less than one thousand. During calibrations, priors were put on b-parameters for any items 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 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 remaining 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.

6.2.1 Vertical Linking

Vertical linking places test scores from different grade levels on the same measurement scale so that we can 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 and backward-linking methods as well as the mixed-linking method. 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 upper grades. No field-test items were included. All items dropped from the previous on-grade calibration steps were not included. For the off-grade vertical linking items, items were dropped after examination of criteria outlined in Table 25 from the grades in which they were flagged. Unlike for ELA, mathematics summative items were not flagged, as 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 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, 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), were calculated. The D^2, or the 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 3 standard deviations were flagged for review, as they excessively impact the scale transformation constants.

Table 25: Flagging Criteria for Vertical Linking Items

Rule	Flagging Criteria	Rationale
p-value	For multiple-choice items, flag if $p < 0.25$ or $p > 0.95$	Items are 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.

Rule	Flagging Criteria	Rationale
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 3 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 smoothed, 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 smoothed, final solution. Evaluation of the a-parameter was performed 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 blueprint at the reporting category level and the removal of the most unstable candidate items first, then checking the blueprint again, then adding back any necessary items, etc.

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 25). 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 26 lists the number of items remaining in the final vertical linking set for each ELA reading and mathematics grade combination.

Results of the initial blueprint violations and final blueprint match can be found in Appendix G.

Table 26: 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 Table 27 and Table 28, respectively.

Table 27: 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 28: 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

Descriptive statistics for ELA reading and mathematics across grades on the vertical scale with mean ability are shown in Tables 29 and 30. 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 29: Descriptive Statistics for ELA Reading on the Vertical Scale

Grade	N	Theta Mean	Theta St Dev	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 30: Descriptive Statistics for Mathematics on the Vertical Scale

Grade	N	Theta Mean	Theta St Dev	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

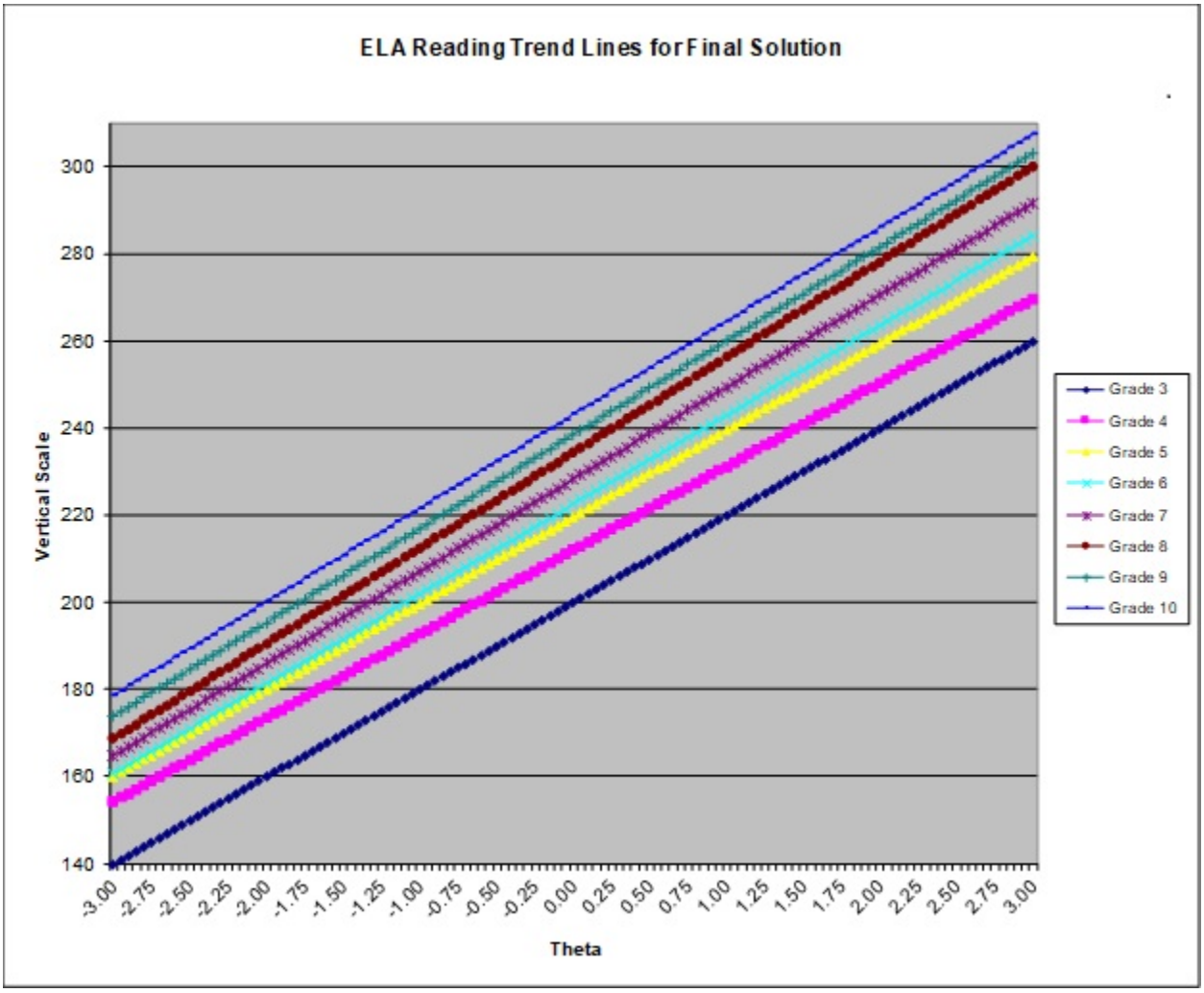
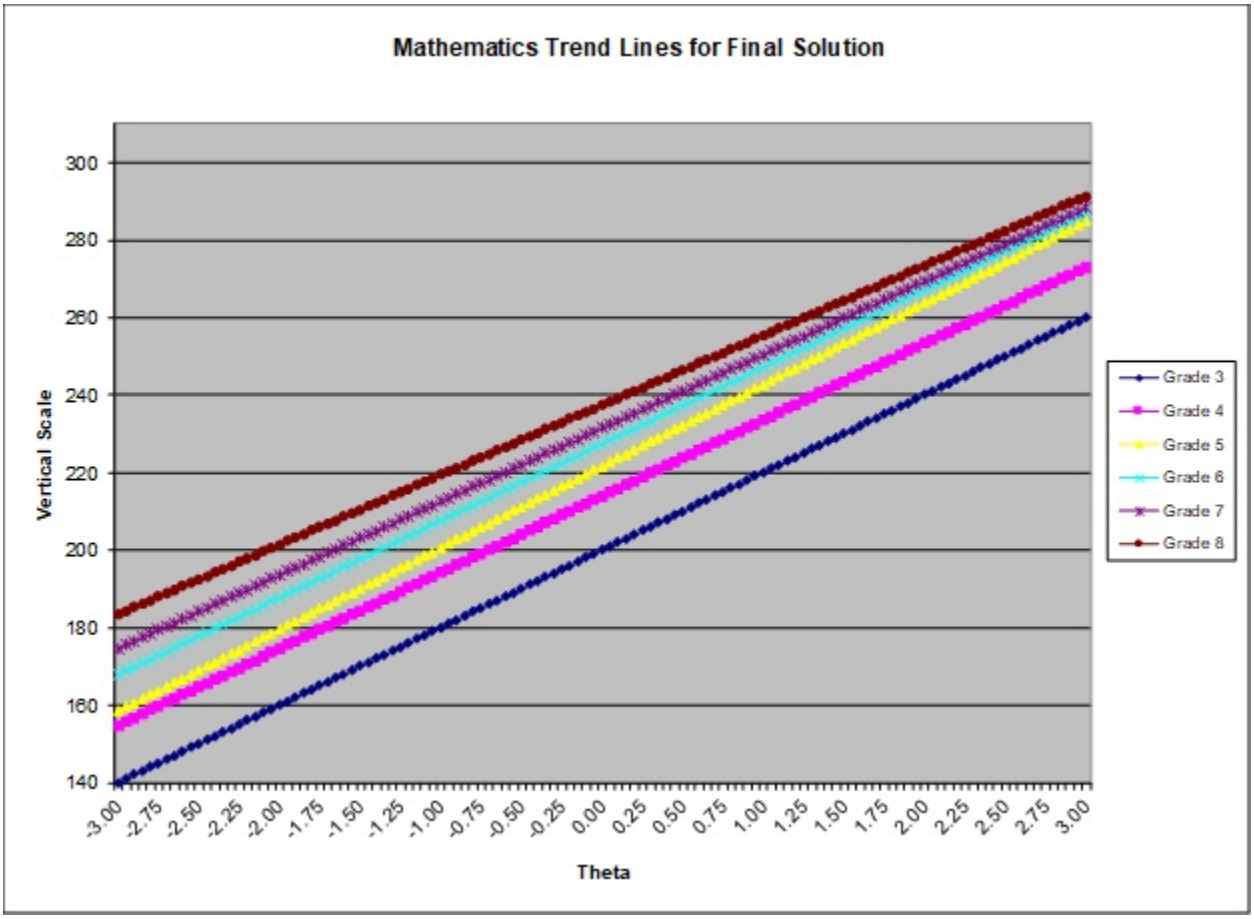


Figure 2: Mathematics Trend Lines for Final Solution



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 progress monitoring 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) were embedded in the grade 3 FAST tests, and a set of grade 3 FAST items (42 reading items and 36 mathematics items) were 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 criteria outlined in Table 18. 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 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 models to scale the FAST tests. To avoid linking between the Rasch model parameters used in the Star assessment and 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 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 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 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 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 better represented the blueprint content area than the grade 3 FAST linking items, the 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 31 shows the number of items remaining in the final vertical linking set for ELA reading and mathematics, and Tables 32 and 33 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 31: 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 32: 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 33: 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	207179	-1.01591	0.95862	1.11810

Subject	Grade	N	FAST Scale		Growth from Renaissance's National Data
			Theta Mean	Growth	
Mathematics	2	205437	-1.33223	1.29447	1.17663

After the final linking constants were selected, a concordance table containing Star scaled scores and corresponding FAST equivalent scaled scores was constructed. 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. Since the linking constants were calculated based on the Star Unadjusted theta scale to 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. The 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.

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

Table 34: 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)

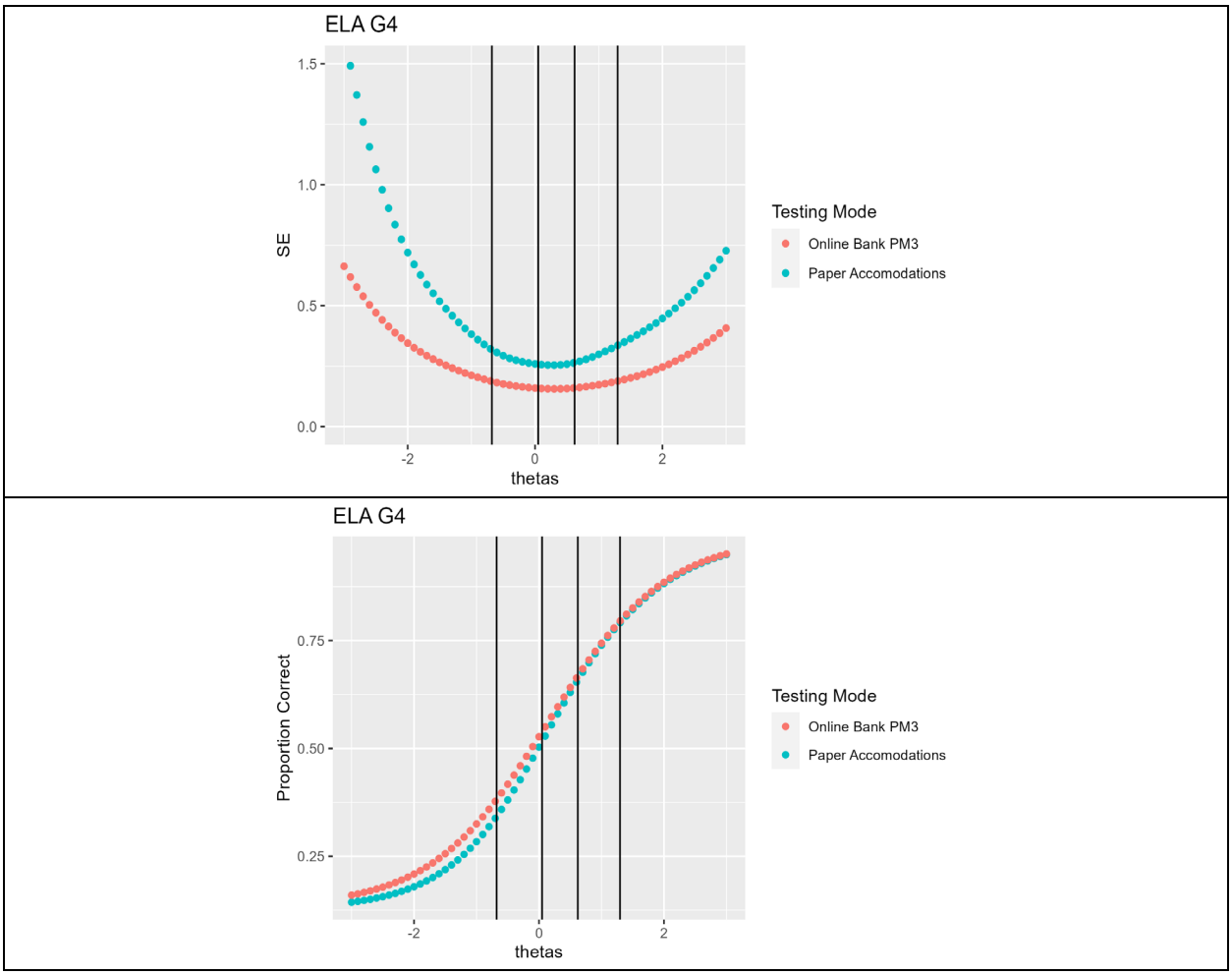
6.2.2 Accommodated Forms

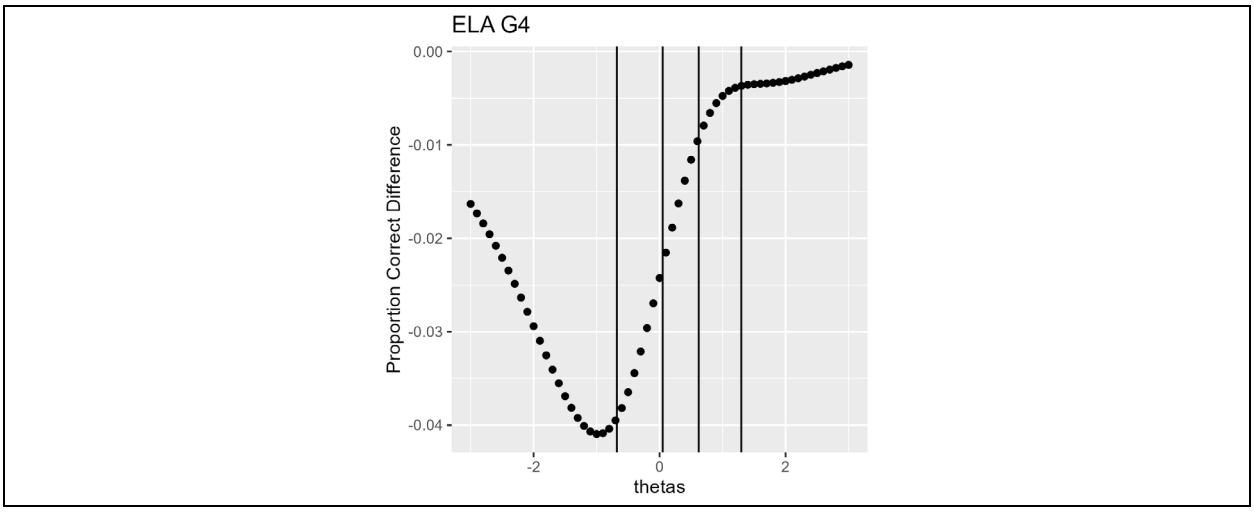
Accommodated forms used online parameters for scoring purposes and no calibrations were performed on the accommodated forms. To create the spring 2023 accommodated forms, CAI ran simulations in summer 2022 (based on the new FAST/B.E.S.T. blueprints and the Computer-Adaptive Test [CAT] algorithm) for each grade. Five resulting forms at each grade that showed the lowest SE at the on-grade cut and within the acceptable range for other statistics (see Section 5, Item Analyses Overview) were selected for further evaluation. Content reviewed the forms and made any necessary item replacements, taking into account suitability for inclusion in an accommodated form and psychometric feedback. Two of the five forms with the best statistics were selected to send to the FDOE for evaluation and selection of a final form. More information about accommodated form construction can be found in Volume 2, Section 4.4, Accommodation Form Construction.

As this was prior to the spring 2023 calibrations and standard setting, the parameters and cuts for the forms were based on the pre-equated FSA scale. Further psychometric information about the 2023 accommodated forms can be found in Appendix C.

Looking ahead to the 2024 accommodated form construction, forms will not be based on simulations but instead be constructed based on 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 SE at the grade-level cut. Figure 3 is a sample of that evaluation. Bank parameters and cuts from the FAST and B.E.S.T scales will be used.

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





6.3 IRT ITEM SUMMARIES

6.3.1 Item Fit

Yen's Q1 (1981) is used to evaluate the degree to which the observed data fit the item response model. 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

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

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

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

where $P_i\left(\hat{\theta}_a\right)$ 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

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

with

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

To determine acceptable fit, both the Q1 and Generalized Q1 results are transformed into the statistic ZQ_1:

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

and are compared to a criterion ZQ_{crit} (FDOE, 1998):

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

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 can be found in Appendix A for operational items and Appendix B 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 for student score calculation.

Appendix B, Field-Test Item Statistics, 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 can be found in Volume 2 of this technical report.

6.3.2 Item Fit Plots

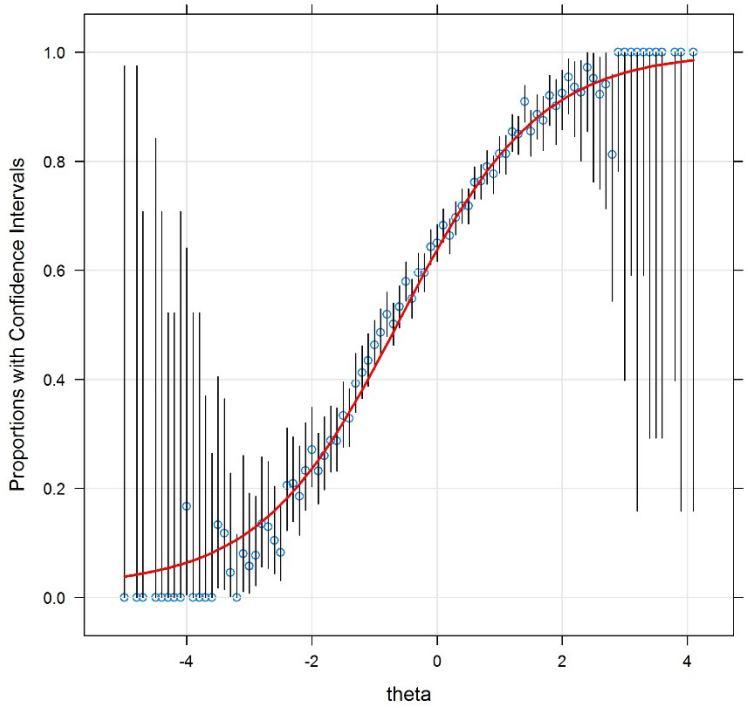
Another way to evaluate item fit is to examine empirical fit plots for each item. The plots in this section are only examples of the types of fit plots used during item calibrations to add to the collection of evidence to evaluate item quality.

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 one-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

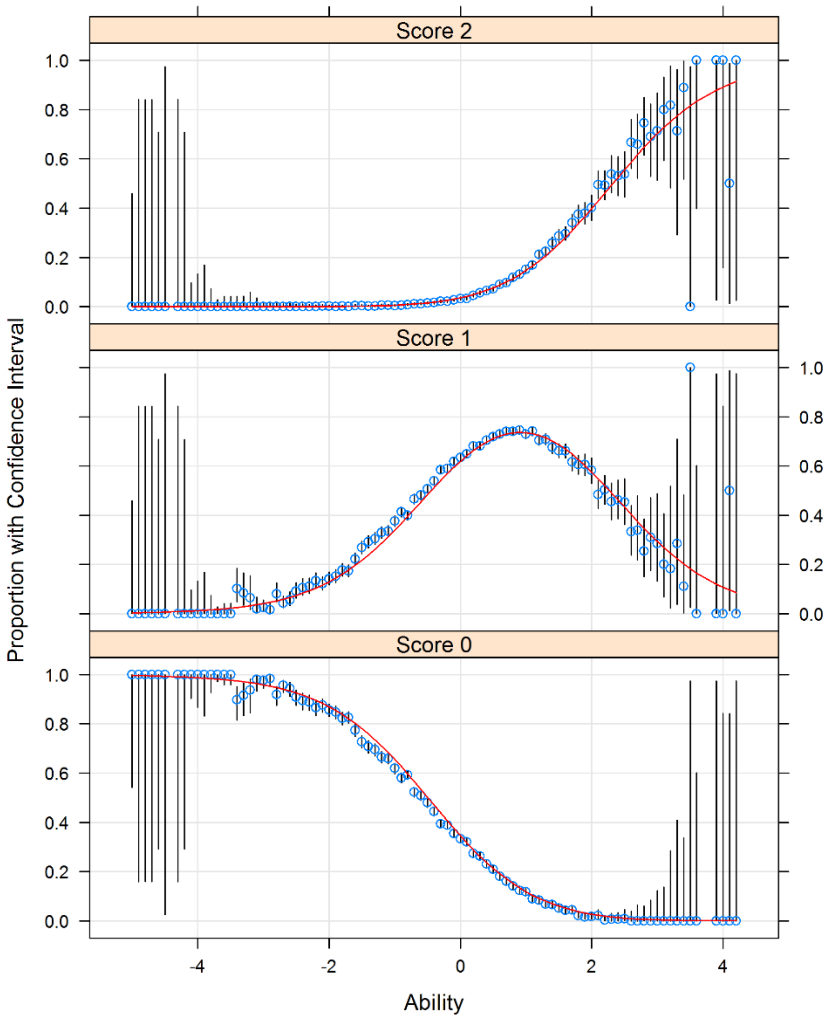
solid line is the IRT-based item characteristic curve. The black lines indicate the error bands associated with the item characteristic curve for each theta point. A "good" item is one in which the observed dots follow the red solid line in the error bands across the range of ability.

Figure 4: Example Fit Plot—One-Point Item



The plot in Figure 5 is provided for items worth two points or more. Again, the red lines represent the IRT-based item characteristic curve. Here, the dots represent the percentage of students within a score bin, at each score point. Like the first plot, 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—Two-Point Item



6.4 Results of Calibrations

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 2023 operational items and Appendix B for the spring 2023 field-test items.

7. SCORING

This chapter provides the scoring procedure used in tests administered in the 2022–2023 school year. It covers the computational details of the maximum likelihood estimation (MLE), standard error of estimate, scale scores, performance level, and subscores reported.

7.1 FAST/B.E.S.T. 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 therefore be expressed as

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

where

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

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

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

$$Q_i = 1 - P_i,$$

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 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

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

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

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

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

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

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

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.

Standard Errors of Estimate

When the MLE is available, the standard error of the MLE is estimated by:

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

where

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

where N_{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. For FAST and B.E.S.T. scoring, the extreme cases were handled as follows:

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

Standard Error of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the standard error (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*. The Florida Assessment of Student Thinking (FAST)/B.E.S.T. 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

There are two scale types created for the FAST/B.E.S.T.:

- A vertical scale score for grades 3–10 English language arts (ELA) and grades 3–8 mathematics
- A within-test scaled score for mathematics end-of-course (EOC) tests

Table 35 shows the theta-to-scale score transformation equations.

Table 35: 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)

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 34.

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

7.1.3 Performance Levels

Each student is assigned a performance category according to his or her accountability scale score. Tables 36–38 provide the cut scores for performance levels for mathematics, ELA reading, and mathematics EOC.

Table 36: 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 37: 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 38: 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

7.1.4 Alternate Passing Score

This section provides information regarding the Alternative Passing Scores (APS) for the FAST and B.E.S.T. assessments for students who took the FAST grade 10 ELA or B.E.S.T. Algebra 1 or Geometry EOC in spring 2023, and who are required to earn a passing score on these tests to meet graduation requirements.

As required, the determination of APS for this group of students was made based on linking the 2021–2022 student performance on the Florida Standards Assessment (FSA) grade 10 ELA and Algebra 1 and Geometry EOC to the 2022–2023 spring student performance on FAST grade 10 ELA and B.E.S.T. Algebra 1 and Geometry EOC, respectively. The following list indicates the APS for the FAST and B.E.S.T. assessments that correspond to the FSA grade 10 ELA and Algebra 1 and geometry EOC passing scores. These alternate passing scores will remain in effect for students in this cohort who participate in the FAST and B.E.S.T. retakes, even after the State Board of Education approves the new achievement-level cut scores for FAST and B.E.S.T. in the coming months. 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 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.

Table 39 indicates the equipercentile relationship between the FSA/EOC level 2/3 cut scores, FAST/B.E.S.T. score scale, and corresponding alternative passing scores on the FAST/B.E.S.T. score scale. The table indicates the new level 2/3 cut scores for the FAST/B.E.S.T. assessments proposed to the State Board of Education for adoption in the coming months. Comparing the APS scores to the proposed new cut scores listed in the last column of the table reveals that the APS scores refer to a student performance similar to or less rigorous than what has been recently proposed for the State Board of Education for adoption for the new assessment system. The table also indicates the equipercentile relationship between the previously mentioned scores and earlier passing scores in Florida.

Table 39: Transitioning from FSA to FAST/B.E.S.T. (2023–2024)

Test Passing Scores by Subject	Level 2/3 Cut Score on FSA Scale for Different Cohorts	Percentile Rank of Level 2/3 Cut Score on the FSA/EOC Score Scale	Percentile Rank of Alternative Passing Cut Score on the FAST/B.E.S.T. Score Scale	Alternative Passing Cut Score on the FAST/B.E.S.T. Score Scale	Approved Level 2/3 Cut Score on the FAST/B.E.S.T. Score Scale (2023 and Later)
FSA Passing Score for Grade 10 ELA	350	52.1	52.1	246	247
FCAT 2.0 Passing Score	349	50.4	50.4	245	
FCAT Passing Score	344	42.0	42.0	240	
Old Passing Score for FCAT	339	35.0	35.0	236	
Old Passing Score for HSCT Students	332	26.0	26.0	229	
FSA Passing Score for Algebra	497	46.9	46.9	398	400
FCAT 2.0 Passing Score for Algebra	489	37.0	37.0	390	
FSA Passing Score for Geometry	499	50.9	50.9	401	404
FCAT 2.0 Passing Score for Geometry	492	42.0	42.0	394	

Note: Rows shaded in gray indicate the outcomes of the recent linking study, which established a connection between the FSA and FAST/B.E.S.T. score scales.

From the Progress Monitoring (PM) 2/winter 2023 administration and beyond, for each test taker, Cambium Assessment, Inc. (CAI) will calculate three passing scores based on each of the three latest cuts (APS Proposed Level 2/3 cut, FSA APS cut, and Florida Comprehensive Assessment Test [FCAT] 2.0 cut) on the FAST/B.E.S.T. scale. The Florida Department of Education (FDOE) will select the relevant passing score for each student to receive. These cut scores are shown in Table 40.

Table 40: Alternate Passing Score Cut Scores

Test	APS Approved Level 2/3 Cut Score on the FAST/B.E.S.T. Score Scale (2023 and Later)	FSA Alternative Passing Cut Score on the FAST/B.E.S.T. Score Scale	FCAT 2.0 Alternative Passing Cut Score on the FAST/B.E.S.T. Score Scale
Grade 10 ELA	247	246	245
Algebra 1	400	398	390
Geometry	404	401	394

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

7.1.5 Reporting Category Scores

In addition to overall scores, students also receive scale scores on each 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. For partial cases, no imputation is made for reporting category scores.

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 (SEM) for the Reporting Category

As with the total score, the SEM for student *i* on 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 each student answers and their estimate of θ, and different students will have different SEM values even if they have the same raw score and/or theta estimate.

Standard Error 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 based on 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

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.

See Appendix E, Distribution of Reporting Category Scores, for the summaries of scores.

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 CHECK

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

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

When all classical and IRT results matched findings from the independent analysts, the results were uploaded to the Secure File Transfer Protocol (SFTP) site for review. Florida Department of Education (FDOE) psychometricians, the Human Resources Research Organization (HumRRO), and Buros also completed independent replications. Meetings were held with CAI, the FDOE, Test Development Center (TDC), HumRRO, and Buros to discuss classical statistics and IRT analyses when needed. Content experts from CAI and the TDC also reviewed classical statistics and provided input. The FDOE approved results when there was replication and verification from all parties.

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

8.2 SCORING QUALITY CHECK

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

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

Like the verification process with calibrations, CAI, the FDOE, and HumRRO performed independent score checks. The FDOE only approved scores when there was three-way replication and verification.

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

In the 2022–2023 school year, Florida's statewide, standardized assessments transitioned from fixed form to adaptive testing. 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, can be found in Volume 2, Section 4, Test Construction and Volume 4, Section 4, Validity.

9.1 ADAPTIVE TESTING ADVANTAGES

According to Birnbaum (1957, as cited in Baker and 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 our accommodated forms, as part of form construction, items are selected to shape the overall test information function so as to provide better reliability of the test in the portion of the ability scale where the most students are scoring or at the achievement-level cuts–which sometimes match. However, it is not possible to tailor the test 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, to solve this problem in practice requires a suitable algorithm for controlling exposure, meeting test blueprints, and selecting items based on ability estimated on the fly, which is made especially challenging under the requirement of three test administrations under the same blueprint. Addressing this challenge requires the focused development of a suitable and sufficient number of items to equip the item bank.

9.2 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, etc., all have an effect on

the algorithm and it is not our goal to elucidate everything here. Both content requirements are mostly expressed in minimum number 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 are simultaneously evaluated for a set of items pre-filtered at each stage to first ensure that candidate items are amongst the best few for satisfying the content requirements. Therefore, the basic principle is to first select items that have maximum content value, with categories that are furthest from meeting minimum requirements prioritized and especially more as the test nears conclusion. Only amongst those, whose number can be adjusted, are any further evaluations made as to their item information as estimated above. 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.

9.3 EVALUATION OF SIMULATIONS

The simulation outcomes are evaluated by psychometricians at the Florida Department of Education (FDOE) as well as Cambium Assessment, Inc. 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 about item bank depth. If the number of times a student takes a test increases, items must be recycled to meet test blueprint requirements, which can also affect the adaptiveness of the test. If items are reused only when necessary (recycling feature on), then a multi-opportunity study is necessary to determine accurate results.

10. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.*

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

Baker, F.B. and Kim, S.H. (2004) Item Response Theory: Parameter Estimation Techniques. 2nd Edition, CRC Press, Boca Raton. (pages 70-71)

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. 91-47). Educational Testing Service. 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://doi.org/10.1177/014662169201600206

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.* Paper presented at the American Educational and Research Association annual meeting, New York, NY.

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. 12-08). Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1109842.pdf