Florida
Statewide Assessments

Florida Statewide Assessments

2021–2022

Volume 1
Annual Technical Report

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

ACKNOWLEDGMENTS

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

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Beginning in fall 2020, all Florida Standards Assessments (FSA) and Next Generation Sunshine State Standards (NGSSS) assessments are collectively referred to as the Florida Statewide Assessments. The *Florida Statewide Assessments 2021–2022 Technical Report* is provided to document all methods used in test construction, outline psychometric properties of the tests, provide summaries of 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 as described below:

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 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 Florida Statewide Assessments standard setting process. This volume is not updated each year because standard setting was finalized in the first year of operational testing.
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 2021–2022.

1.1 PURPOSE AND INTENDED USES OF THE FLORIDA STATEWIDE 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 give 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 forward student learning. 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 (Standards for Educational and Psychological Testing [American Educational Research Association, American Psychological

Association, & National Council on Measurement in Education, 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 Florida Statewide Assessments standards and test blueprints can be found in Volume 2, Test Development. Additional verification of content validity can also be found in Section 4 of Volume 4, Evidence of Reliability and Validity. The documentation about the comparability of online and paper-pencil tests can be found in Section 5 of Volume 4, Evidence of Reliability and Validity.

The Florida Statewide Assessments yield test scores that are useful for understanding whether individual students have a firm grasp of the Florida Standards and whether students are improving in their performance over time. Additionally, scores can be aggregated to evaluate the performance of subgroups, and both individual and aggregated scores can be compared over time using program evaluation methods. The reliability of the test scores can be found in Section 3 of Volume 4, Evidence of Reliability and Validity.

The Florida Statewide Assessments are criterion-referenced tests that are intended to measure whether students have made progress on the Language Arts Florida Standards (LAFS), the Mathematics Florida Standards (MAFS), and the Next Generation Sunshine State Standards (NGSSS). The Florida Statewide Assessments standards and test blueprints are discussed in Volume 2, Test Development.

Table 1 outlines required uses of the FSA and the NGSSS.

Table 1: Required Uses and Citations for the Florida Statewide 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.
		District Grades	s. 1008.34, F.S.
		Differentiated Accountability	s. 1008.33, F.S. Rule 6A-1.099811, F.A.C.
		Opportunity Scholarship	s. 1002.38, F.S.

Appendix J of this volume provides a glossary of terms, abbreviations, and acronyms used throughout the technical report.

1.2 Background and Historical Context of Test

To accompany the development of new Florida educational standards, the FSA was designed to measure students' progress in English Language Arts (ELA), Mathematics, and End-of-Course

(EOC) tests. The FSA was first administered to students during spring 2015, replacing the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0) in English Language Arts and Mathematics. Students in Grade 3–6 Reading and Mathematics were administered fixed operational forms on paper. Students in Grades 7–8 Mathematics and Grades 7–10 Reading were administered fixed operational forms online. Online operational EOC assessments were given to students taking Algebra 1 and Geometry. In 2009, the revisions of the Sunshine State Standards approved by the Florida State Board of Education (SBE) in 2007 and 2008 started to be referred to as the 2007 NGSSS and 2008 NGSSS, respectively. NGSSS assessments were administered to students starting from spring 2012. For all online assessments, paper accommodated versions were available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicated such a need.

Within the current Florida statewide assessments program, students in grade 3 must score at Level 2 or higher on the Grade 3 ELA assessment in order 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 have the opportunity to retake the assessments multiple times; they may also use concordant scores on the ACT or SAT to meet the Grade 10 ELA requirement; or they may earn a comparative passing score on the Postsecondary Education Readiness Test (PERT) 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.

In the rest of this section, the transition to the FSA will be highlighted. This brief background should establish the legislative and curricular framework for the technical analyses described in the remaining sections of this volume and other volumes of the technical report.

Developments in 2012

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

Developments in 2013

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

Developments in 2014

In response to Executive Order 13-276, the state of Florida issued an Invitation to Negotiate in order 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.

During summer 2014, psychometricians and content experts from CAI, the FDOE, and the Department's Test Development Center met to build test forms for spring 2015. Because it was necessary to implement an operational test in the following school year, items from the state of Utah's Student Assessment of Growth and Excellence (SAGE) assessment were used to construct Florida's test forms for the 2014–2015 school year. Assessment experts from FDOE, the Department's Test Development Center, 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 that were deemed suitable from both perspectives were considered for inclusion on Florida's assessments and for constructing Florida's vertical scale.

It is important to note that, in Florida, post-equating is used each year, so all data used for evaluating student performance on the FSA was derived from the Florida population after the spring 2015 administration.

In addition to the operational test items, field-test items were embedded into test forms administered online in order to build the Florida-specific FSA item pool for future use. These items were placed onto test forms using an embedded field-test design in the same fixed positions across all test forms within a grade. A very large number of items were field tested, as described later in this volume, in order to build a substantial bank of items to 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 that occurred during the winter of 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 a large number of prompts for the state. These prompts have been used since spring 2016.

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

Developments in 2015

The first operational administration of the FSA occurred in spring 2015. Grade 3 and Grade 4 ELA and Mathematics assessments were administered entirely on paper, and all other grades and subjects were administered primarily online, with the exception of 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 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 FCAT 2.0 Reading and NGSSS Algebra 1 EOC, respectively. This linking was required to determine student-level eligibility for promotion (Grade 3 ELA) and graduation (Grade 10 ELA and Algebra 1), which are also statutory requirements. This was accomplished using equipercentile linking for Grade 10 ELA and Algebra 1. Further legislation enacted in spring 2015 changed the promotion requirement for Grade 3 ELA, instead requiring that students scoring 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. The most significant implication of this legislation was that a significant number of students in grade 8 participated in the Algebra 1 EOC but not the Grade 8 Mathematics assessment. This will be discussed in more detail in other volumes of the technical report, especially as it relates 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 occurred 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 from 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. More information about the method and procedure can be found in Section 6.2, Equating to the IRT (item response theory)-Calibrated Item Pool.

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*, a major change that affected test administration during school year (SY) 2019–2020 was the cancellation of the spring 2020 assessments due to the COVID-19 pandemic. Specifically, by the time of the cancellation, only Grade 10 ELA Writing Retake, Grade 10 ELA 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, Grades 5 and 8 Science, Algebra 1, Geometry, Biology 1, Civics, and U.S. History EOC. As a consequence 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 prior year (i.e., the spring 2019 regular assessments) for processes that were not completed prior to the cancellation, whereas results were reported based on spring 2020 for processes that were completed prior to the cancellation.

Developments in 2021

As a consequence 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 thus 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 given that Florida's large population sizes totaling around 200,000 students per grade and subject facilitated obtaining sufficient sample sizes for all of the 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 of the field-test items developed in 2020 to make sure they were free from any bias or sensitivity issues due to the ongoing COVID-19 event before they were field tested in spring 2021.

Developments in 2022

New items in Grade 3 Reading, Grades 4–10 ELA, Grades 3–8 Mathematics, and Mathematics EOC tests (i.e., Algebra 1 and Geometry) have been developed under the guidelines of Florida's new standards, the Benchmarks for Excellent Student Thinking (B.E.S.T.) Standards. These items were field tested in spring 2022. The BEST items are used to develop the Florida Assessment of Student Thinking (FAST) in Grades 4-10 Reading and Grades 3-8 Mathematics, and the BEST assessments for Algebra 1 and Geometry EOC.

1.3 Participants in the Development and Analysis of the Florida Statewide Assessments

The FDOE manages the Florida Statewide Assessments 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 of implementing Florida's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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. Funded by the FDOE via a grant, the Test Development Center (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 participate in the development of academic standards, the clarification of how these standards will be assessed, test design, and review of test questions and passages.

Technical Advisory Committee

FDOE convenes a panel once a year (twice if technical issues/concerns arise) to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Florida testing. This committee is comprised of several nationally recognized assessment experts and highly experienced practitioners from multiple Florida school districts.

Cambium Assessment, Inc. and Pearson

Cambium Assessment, Inc. (CAI) and Pearson were the vendors selected through the state-mandated competitive procurement process. CAI and Pearson were responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the Florida Statewide Assessments described in this report. All activities were conducted under the close direction of FDOE staff experts.

Human Resources Research Organization

The Human Resources Research Organization (HumRRO) has provided program evaluation to a wide variety of federal and state agencies as well as corporate and non-profit organizations and foundations. For the Florida Statewide 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 professional assistance, expertise, and information to users of commercially published tests. For the 2022 Florida Statewide Assessments, Buros provided independent operational checks on the equating procedures, Writing hand scoring activities, and the scanning and editing services provided by CAI. Each year, Buros delivers reports on their observations, which are available upon request.

Caveon Test Security

Caveon Test Security analyzes Florida Statewide Assessments 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.

1.4 Available Test Formats and Special Versions

Students in Grades 3–6 Reading and Mathematics and Grades 5 and 8 Science are administered fixed operational forms on paper. Students in Grades 7–8 Mathematics and Grades 7–10 Reading are administered fixed operational forms online. Online operational EOC assessments are given to

students taking Algebra 1, Geometry, Biology 1, U.S. History, and Civics. For all online assessments, paper accommodated versions are available to students whose IEPs or Section 504 Plans indicated such a need.

Administered test forms contain operational items and embedded field-test (EFT) items in pre-determined slots across each form. Operational items are used to calculate student scores. EFT items are non-scored items and are used either to populate the Florida Statewide Assessments test bank for future operational use or to equate the current year forms to the item response theory (IRT; van der Linden & Hambleton, 1997) calibrated item pool. While there is only one operational form in Grades 3–8 Mathematics, Grades 3–10 Reading, and Grades 5 and 8 Science, there are multiple test forms in order to vary the EFT items on each form and build a large item bank.

Students in grades 4–10 respond to a single, text-based Writing prompt; the assessments for Grades 4–6 Writing are administered on paper, and Grades 7–10 Writing are administered online. Writing and Reading item responses are combined so that the data can be calibrated concurrently and subsequently to form an overall ELA score. Scale scores for the separate components are not reported. In this document, the term *ELA* is used when referring to the combined Reading and Writing score and *Reading* is used when referring to only the Reading test form or items.

EOC assessments are administered as online, fixed-form assessments to students enrolled in Algebra 1, Geometry, Biology 1, U.S. History, and Civics. These tests have multiple operational forms and contain EFT items to build future test forms as well as items to equate the current-year forms to the IRT-calibrated item pool.

1.5 STUDENT PARTICIPATION

By statute, all Florida public school students are required to participate in the statewide assessments. Students take Mathematics, Reading, Writing, NGSSS Science, or EOC tests in the Florida Statewide 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.

Table 2 shows the number of students who were tested and the number of students who were reported in the spring 2022 Florida Statewide Assessments by grade and subject area. The participation counts by subgroup, including gender, ethnicity, special education, and English language learner status (ELL), are presented in Section 9, Statistical Summary of Test Administration, of this volume of the technical report. Table 3 presents the percentage of students in each performance level for grades and subjects that were reported for the spring 2022 Florida Statewide Assessments. Please refer to Appendix F for descriptive statistics on the scale score distributions across all students and subgroups.

Table 2: Number of Students Participating in Florida Statewide Assessments

Mathematics			ELA			Science and NGSSS EOC		
Grade/Test	**Number Tested**	**Number Reported**	**Grade**	**Number Tested**	**Number Reported**	**Test**	**Number Tested**	**Number Reported**
3	208,253	207,531	3	210,992	210,396	Science 5	212,877	211,831
4	195,760	195,047	4	202,994	198,594	Science 8	200,017	199,034
5	211,499	210,709	5	216,849	212,492	Biology 1	208,969	208,677
6	187,502	185,275	6	209,003	197,122	Civics	211,706	211,533
7	174,299	171,011	7	214,337	207,191	U.S. History	180,440	180,243
8	154,130	150,778	8	220,602	213,464			
Algebra 1	222,412	217,541	9	219,068	209,276			
Geometry	195,299	191,298	10	218,793	203,493			

Table 3: Percentage of Students Taking Operational Forms by Performance Level

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	24.1	18.0	26.3	21.6	10.0
	4	24.4	14.6	25.5	21.7	13.8
	5	29.0	19.1	21.6	19.0	11.3
	6	29.5	21.8	22.5	18.7	7.5
	7	32.6	21.4	24.6	14.6	6.7
	8	35.9	22.6	23.2	10.3	7.9
ELA	3	24.9	22.2	27.7	19.0	6.2
	4	23.3	19.7	25.1	21.4	10.5
	5	22.7	22.6	25.0	20.0	9.7
	6	24.8	22.9	20.9	22.3	9.0
	7	29.8	21.9	21.3	17.3	9.7
	8	29.9	21.0	23.0	16.1	9.9
	9	27.1	21.5	20.7	20.6	10.1
	10	28.0	23.4	20.2	19.2	9.3
EOC	Algebra 1	35.5	12.0	27.4	13.9	11.2
	Geometry	34.5	16.2	27.8	11.4	10.1
	Biology 1	14.6	24.9	32.8	11.3	16.3
	Civics	15.2	15.9	24.5	20.2	24.2
	U.S. History	16.6	18.4	25.5	19.5	20.0
Science	5	28.9	23.0	24.2	11.5	12.3
	8	26.3	28.8	21.3	12.4	11.3

*Please see the "Number Reported" column in Table 2 for n-counts of all students in each grade and subject.

2. RECENT AND FORTHCOMING CHANGES TO THE TEST

The purpose of this section is to highlight and document any major issues affecting the test or test administration during the current year, and any major changes that have occurred to the test or test administration procedures over time.

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

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 2022 administration.

3.1 SPRING ADMINISTRATION PROCEDURES

Table 4 shows the schedule for the spring administration of the 2021–2022 Florida Statewide Assessments, broken down by testing window and subject area.

Table 4: Testing Windows by Subject Area

Assessment	Testing Window
Algebra 1 Retake	February 21 – March 11, 2022
ELA Retake Reading and Writing	September 13 – October 15, 2021 February 21 – March 11, 2022
Paper Grade 3 Reading	April 4 – April 15, 2022
Online Grade 4–10 Writing	April 4 – April 15, 2022
Paper Grade 4–6 Reading Paper Grade 3–6 Mathematics	May 2 – May 13, 2022
Paper Science 5 & 8	May 9 – May 20, 2022
Online Grade 7–10 Reading Online Grades 7–8 Mathematics	May 2 – May 27, 2022
Paper and Online Algebra 1 & Geometry Paper and Online Biology 1, Civics, and U.S. History	September 13 – October 15, 2021 November 29 – December 17, 2021 May 2 – May 27, 2022 July 11 – 22, 2022

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 Statewide Assessments were administered in sessions, with each session having a time limit. Once a session was started, a student was required to finish it before he or she was permitted to leave the school's campus. A student could not return to a session once he or she left campus.

The key personnel involved with the Florida Statewide Assessments 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 Statewide Assessments 2021–2022 Technical Report*.

A secure browser developed by CAI (CAI Secure Browser) was required to access the online Florida Statewide Assessments. 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 Florida Statewide Assessments 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 of the platforms on which Florida Statewide Assessments 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, American Psychological Association, & National Council on Measurement in Education, 1999, 2014, Standards 9.7 & 9.9), CAI conducted a device comparability study to provide evidence of comparability of the Florida Statewide Assessments scores across devices. 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 paper matches the content of the items as administered online (i.e., wording, graphics, paragraph breaks, and option order).

3.2 FLORIDA STATEWIDE ASSESSMENTS ACCOMMODATIONS

Florida assessments are 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 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).

The paper version is constructed to the exact same test specifications and, in many cases, the items on the online and paper forms are exactly the same. Some technology-enhanced items are replaced on the paper versions with items intended to render on paper. They are chosen to essentially mirror the online items they are replacing such that the paper form measures the same construct in a similar way.

Observed data collected from the test administrations provide evidence that the test forms are equally as reliable and that students using the paper 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 raw score summary (including mean score, standard deviation, minimum score, maximum score, and

cronbach's alpha) 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 paper-based (accommodated) version of the 2021–2022 Florida Statewide Assessments varied between 370 and 708 across grades and subjects, as shown in Table 5.

Table 5: Counts of Paper-Based Assessments by Grades and Subjects

Subject	Grade	Spring 2022
Mathematics	7	632
	8	600
ELA	7	652
	8	708
	9	570
	10	572
EOC	Algebra 1	634
	Geometry	498
	Biology 1	444
	Civics	577
	U.S. History	370

Table 6 shows the percentage of students in each performance level for grades and subjects that had paper accommodated forms in spring 2022. In general, online test takers tend to score at higher achievement levels, compared to paper-based test takers.

Table 6: Percentage of Students Taking Paper Forms by Performance Level

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	7	54.4	22.6	16.5	5.1	1.4
	8	58.2	18.8	15.7	5.2	2.2
ELA	7	52.6	21.9	16.0	7.2	2.3
	8	53.5	19.9	15.5	8.5	2.5
	9	49.8	23.5	14.6	9.6	2.5
	10	47.6	20.5	15.2	12.2	4.5
EOC	Algebra 1	55.4	14.4	18.8	7.4	4.1
	Geometry	52.6	18.9	19.3	6.0	3.2
	Biology 1	24.5	33.6	29.1	7.0	5.9
	Civics	31.7	23.1	23.7	11.6	9.9
	US History	27.8	25.1	24.3	10.3	12.4

The TA and the school assessment coordinator were responsible for ensuring that arrangements for accommodations were made before the test administration dates. For eligible students participating in paper-based assessments, a variety of accommodations were available, such as large print, contracted braille, uncontracted braille, and displaying only one item per page. 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 had the opportunity to use these accommodations only as dictated on their IEPs or Section 504 Plans. An accommodation summary for the Florida Statewide Assessments in school year 2021-2022 is provided in *Accommodation Analysis* (*Appendix H)*. The information includes the accommodations provided for test takers overall and the accommodations for test takers from two special subgroups: ELL and students with disabilities (SWD). Additional accommodations and further explanation of the 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 and number of forms administered in spring 2022.

4.1 OVERVIEW OF ITEM DEVELOPMENT

Complete details of the item development plan for Cambium Assessment, Inc. (CAI) and Pearson are provided in the *Florida Statewide Assessments 2021–2022 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 added to the operational item pool after being field tested. 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.

In spring 2022, field-test items were embedded on paper forms for Grades 3–6 Reading and Mathematics, Grade 5 and 8 Science, and on online forms for Grades 7–8 Mathematics, end-of course exams (EOCs), and Grades 7–10 Reading. All assessments are fixed-form tests with a predetermined number and location of field-test items. The paper 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 operational calibration, items are reviewed based on their performance during the spring administration. In spring 2022, no items were removed from scoring.

Prior to the spring administration, a *Calibration and Scoring Specifications* document is created by CAI, Pearson, 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 paper accommodated forms, and scoring. CAI and Pearson use the specifications to complete classical item analyses and IRT calibrations (see Sections 5 and 6 of this volume of the technical report) for each test and post results to a secure location for review. During the spring calibrations, daily calls are scheduled that include all parties, including CAI, Pearson, FDOE, Test Development Center (TDC), HumRRO, and Buros. 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 sections that follow. Psychometricians and content experts work together to review items and their statistics to determine if any items are to be removed from scoring.

4.3 FIELD TESTING

The Florida Statewide Assessments 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

Florida Statewide Assessments forms are pre-built with approximately 6–10 field-test items embedded onto each test form, and each form is assigned to students randomly, as described here. Some field-test items may appear on multiple forms.

Table 7 shows the number of Mathematics and EOC items by grade and item type that are included on spring 2022 forms for field testing. Table 8 shows the number of Reading items by grade and item type that were included on spring 2022 forms for field testing. Table *9* shows the number of items field tested on the spring 2022 forms for NGSSS Science and EOC. During calibrations, some items were dropped from the initial item pool due to poor performance. Appendix C, Field-Test Item Statistics, provides the number of field-test items remaining after removal of items during calibrations. The descriptions of item types are presented in Tables 22–24 of Volume 2 of the *Florida Statewide Assessments 2021–2022 Technical Report.*

Table 7: Mathematics and 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	45	48	57	65	118	84	82	90
ETC	13	12	13	20	15	25	19	33
GI	0	0	0	0	4	10	6	6
HT	0	0	0	0	0	0	0	1
MC	57	62	47	73	81	106	107	67
MI	10	6	6	8	5	5	5	3
MS	20	9	13	24	7	24	18	15
Multi	2	3	4	4	11	8	19	10
Total Number of Items	147	140	140	194	241	262	256	225

Table 8: 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	16	15	24	19	33	36	32	45
HT	0	0	1	0	3	2	4	6
MC	131	104	104	118	197	172	169	177
MI	9	3	7	9	19	16	12	8
MS	21	18	20	12	13	7	10	23
Two-Part HT	0	0	1	0	0	1	0	1
Total Number of Items	177	140	157	158	265	234	227	260

Table 9: NGSSS Science and EOC Field-Test Items by Item Type and Grade

Item Type	Grade 5	Grade 8	Biology 1	U.S. History	Civics
MC	77	121	206	125	119

With fixed-form assessments, it is known how many items are unique to a form. Thus, based on the number of students participating, as well as the number of forms, the expected number of responses per item can be calculated.

The form distribution algorithm employed by CAI ensures that forms are drawn and assigned to students according to a simple random sample. For example, suppose there are *J* total forms in the pool, items appear on only one form, and a total of *N* students are participating in the field test. The probability that any one of the *J* forms can be assigned to one student is *1/J*. Thus, the expected number of student responses for each form is

$$S = \frac{N}{J},$$

where *J* is the number of forms in the pool, *N* is the number of students who will be participating in the field test, and *S* is the sample size per item. If an item appears on more than one form, the expected sample size would be *S* times the number of forms on which the item appears.

The aim was to achieve a minimum sample size of 1,500 students per item. Hence, given a test length of *L* and fixing *S* at 1,500 (the expected sample size per item), we can determine the maximum number of forms that can exist in the pool as

$$J = \frac{N}{1500}.$$

From this, we see that

- a random sample of students receives each form; and
- for any given form, the students are sampled with equal probability.

It is important to note that even though 1,500 is the minimum requirement, many more responses than 1,500 (typically around 3,000 to 3,500) are always available given Florida's large student population.

Table 10, Table 11, and Table 12 show the total number of forms administered in spring 2022. In each grade, there is a single core or operational form. The same core form is replicated for each anchor or embedded field-test form, resulting in multiple forms for each grade and subject. For the EOCs, there are multiple core forms, each also replicated to create several embedded field-test forms.

Table 10: Reading Form Summary

Grade	Total Number of Forms
3	30
4	25
5	28
6	27
7	43
8	41
9	35
10	39

Table 11: Mathematics and EOC Form Summary

Grade	Total Number of Forms
3	19
4	18
5	19
6	24
7	33
8	32
Algebra 1	28*
Geometry	28*

*Note that Text-To-Speech (TTS) was not counted in the total number of forms in EOC.

Table 12: NGSSS Science and EOC Form Summary

Grade	Total Number of Forms
5	13
8	17
Biology 1	27*
U.S. History	16*
Civics	15*

*Note that Text-To-Speech (TTS) was not counted in the total number of forms in EOC.

A detailed overview of the development and review process for new items is given in the *Florida Statewide Assessments 2021–2022 Technical Report*, Volume 2, Test Development. Additional details on development and maintenance of the item pool are also given in the same volume.

5. ITEM ANALYSES OVERVIEW

This chapter summarizes the classical item analyses and differential item functioning (DIF) analyses and provides the results.

5.1 CLASSICAL ITEM ANALYSES

Item analyses examine whether test items function as intended. Overall, a minimum sample of 1,500 responses (Kolen & Brennan, 2004) per item is required for both classical item analysis and IRT analysis. However, many more responses than 1,500 are always available. For operational item calibrations, an early processing sample is used in the analyses; for field-test item calibrations, all students are used. 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), for integrity and appropriateness of the statistical characteristics of the items. The thresholds used to flag an item for further review based on classical item statistics are presented in Table 13.

Table 13: Thresholds for Flagging Items in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Point biserial or point polyserial correlation for the correct response is < 0.25.
Distractor Analysis	Point biserial correlation for any distractor response is > 0.
Item Difficulty (MC items)	The proportion of students (*p*-value) is < 0.20 or > 0.90.
Item Difficulty (non-MC items)	Relative mean is <0.15 or >0.95.

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 is able to differentiate between high- and low-achieving students. The discrimination index for MC items is calculated as the correlation between the item score and the ability estimate for students. Corrected Point biserial or corrected 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 most frequently selected option 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

Items that are either extremely difficult or extremely easy 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, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) document provides a guideline for 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 Statewide Assessments 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 Disability (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 ability estimate

on the operational items on a given test is used as the ability-matching variable. Specifically, raw scores on operational items are used during initial operational and anchor item calibrations. After operational scoring is complete, DIF analyses for these operational items are updated using IRT ability estimates 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 and 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 standardized mean difference (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 14. Items were also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American, Hispanic, female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., white, male). If the DIF statistics fell into the "C" category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged on the basis of DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance in field testing.

Table 14: DIF Classification Rules

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

DIF summary tables can be found in Appendix A, Operational Item Statistics, for operational items, Appendix B, Anchor Item Statistics, for anchor items, and Appendix C, Field-Test Item Statistics, for field-test items. Across all tested grades and DIF comparison groups, less than 1% of Mathematics, EOC, Science, and English Language Arts (ELA) items were classified as C DIF for operational items and anchor items. Items were reviewed by content specialists and psychometricians to ensure that they were free of bias.

For field-test items, less than 1% of Mathematics, EOC, Science, and ELA items were classified as C DIF across all tested grades and DIF comparison groups. All field-test items will be reviewed by content specialists and psychometricians before being placed on forms for operational use. More information about test construction and item review can be found in Volume 2 of this technical report.

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 Statewide Assessments 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 a large number of 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 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 2- and 3-parameter logistic (2PL, 3PL) models and the Generalized Partial Credit Model. 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 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 Florida Statewide Assessments items go through 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.

Operational item calibrations were completed on an Early Processing Sample (EPS) collected during the spring administration. The EPS was a representative, scientific sample of students across the state. The sampling of students was accomplished using a stratified random sample with explicit and implicit strata that were chosen to represent important characteristics of the tested student population. Region was used as explicit strata, whereas gender, ethnicity, school size, mean theta score, and curriculum group (Standard, Limited English Proficiency [LEP], Exceptional Student Education [ESE]) were used as implicit strata. The *region* variable is intended to capture the differences in student population across the state. Male and female are the subgroups under *gender*, whereas *ethnicity* is comprised of white, African American, Hispanic, and other subgroups. *Mean theta score* provides the measure of the student ability across the population based on the previous year's data. The *school size* variable is used in sampling to ensure that the sample is comprised of schools of various sizes. The *curriculum group* variable has three subgroups: Standard, LEP, and ESE. This variable shows that the representativeness of the ELL population is also evaluated as part of the sample evaluation. More information about the EPS can be found in Appendix D, EPS Sampling Plan, of this volume of the technical report.

The Florida Department of Education (FDOE) and Cambium Assessment, Inc. (CAI) collaborated through several rounds of review to ensure that the strata were appropriately defined and the student population was adequately represented; this EPS plan, which can be found in Appendix D, was also reviewed and affirmed by the Technical Advisory Committee (TAC). For Grade 8

Mathematics and EOC calibrations, the entire population was used instead of the EPS. Please note that students taking certain anchor forms *only* take certain anchor items (for example, 10 items in Mathematics), but when all anchor forms are randomly administered to the entire calibration sample, the resulting data include student responses across all anchor items.

No EPS was used to define the calibration sample for NGSSS tests. For Grades 5 and 8 Science, at least 65% of the total population, including about 90% of Miami-Dade County's population, was used for calibration. For Biology 1, Civics, and U.S. History EOC, at least 60% of the total population, including about 85% of Miami-Dade County's population, was used.

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 prior to the operational testing, whereas post-equating happens after the operational testing, and 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.

Both of these approaches are employed in Florida. For retake test administrations, test forms are pre-equated, and student responses are directly scored based on pre-equated statistics available in the bank. For spring non-retake administrations, post-equating is used, and all data regarding item responses are derived from the most recent group of students to be administered the test. Beginning in 2016, FSA test forms were equated to the IRT calibrated item pool, a step that was not necessary in the initial year. Grades 5 and 8 Science and Biology 1 forms are equated to the IRT scale established in 2012. Civic and U.S History forms are equated to the scale established in 2013 and 2014, respectively. This process is described in further detail in Section 6.2, Equating to the IRT-Calibrated Item Pool, of this volume of the technical report.

Field-test item calibrations were completed on the entire sample from the spring administration to ensure adequate sample sizes for all items. Field-test items were equated to the operational scale using the Stocking-Lord procedure.

6.1 ITEM RESPONSE THEORY METHODS

The generalized approach to item calibration was to use the 3-parameter logistic model (3PL; Lord & Novick, 1968) for MC items; to use the 2-parameter logistic model (2PL; Lord & Novick, 1968) for binary items that assume no guessing; and to use the generalized partial credit model (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, the content of the item 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

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

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

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

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

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

All item parameter estimates were obtained with IRTPRO version 5.0 (Cai, Thissen, & du Toit, 2011). The Marginal Maximum Likelihood Estimation (MMLE) procedure is employed by IRTPRO to estimate item parameters.

6.2 EQUATING TO THE IRT-CALIBRATED ITEM POOL

Equating is a statistical procedure in which test scores from different test instruments measuring the same or similar construct are placed onto a common scale so that scores from different test administrations can be directly compared. Equating in English Language Arts (ELA), Mathematics, and Mathematics End-of-Course (EOC) began in 2016 when item parameters in the 2016 forms were post-equated to the Florida Statewide Assessments scale established in 2015, by constructing a calibrated pool of items on the same scale via the common-item non-equivalent groups anchor design. In 2016, an "anchor" set was selected from the item pool such that these common items can be used for equating the 2016 form to the IRT calibrated item pool. This is the basis of the common-item equating design that enables items to be placed in a calibrated pool (Kolen & Brennan, 2004). The same equating procedure was also performed in 2017 and beyond. The anchor set is essentially a miniature version of a parallel test with respect to its content and statistical characteristics. That is, the items in the anchor set represent the blueprint percentages as well as having similar statistical properties as the full test.

The same equating process was followed for the NGSSS Science and EOC assessments. In 2013, Science (in grades 5 and 8) and Biology test forms were post-equated to the NGSSS scale established in 2012. U.S. History forms were post-equated in 2014 and placed on the 2013

operational scale. Civics forms were post-equated in 2015 to the operational scale established in 2014.

During test construction, items are selected and evaluated using their statistical properties collected from the item bank. These statistical characteristics are provisional, given that post-equating is used, but they are useful for guiding the construction of anchor sets, as well as the overall test form. The statistical characteristics typically include evaluations of an item's *p*-value, point-biserial correlations, and IRT-based characteristics (i.e., difficulty, guessing, slope) and differential item functioning (DIF). Items are selected such that forms meet the test blueprint, and classical and IRT summary statistics are also calculated and compared to the prior years. The process is iterative and continues to choose items with content and statistical properties, as well as professional judgment by content experts, to build a linking set that conforms to the blueprint and statistical characteristics of the prior year forms. Once finalized, a subset of items is labeled as *anchor* items to be used to complete equating during operational calibrations. Additional details about test construction are available in Volume 2, Test Construction, of the *Florida Statewide Assessments 2021–2022 Technical Report*.

6.2.1 Online Forms

Online operational and anchor items were jointly analyzed using the EPS in Grades 7–10 ELA and Grade 7 Mathematics and using the entire population in Grade 8 Mathematics and in the EOCs. The EPS is a scientific sample of students and is representative of Florida students. Prior to analyses, demographics of the EPS were compared to state values used to draw the samples to ensure representativeness. More information about the EPS can be found in Appendix D, EPS Sampling Plan, of this volume of the technical report. Grades 5 and 8 Science calibration used at least 65% of the total population, including about 90% of Miami-Dade County's population. For Biology 1, Civics, and U.S. History EOC tests, at least 60% of the total population, including about 85% of Miami-Dade County's population, was used. The Human Resources Research Organization (HumRRO) replicated all item calibrations and provided an independent list of flagged items. Buros provided additional commentary on calibrations and flagged items.

Classical item statistics, as described in Section 5, Item Analysis Overview, were computed first and reviewed to determine if any items should be removed from analyses prior to either IRT calibrations or equating. Content experts from CAI, Pearson, and TDC reviewed flagged items to ensure that they were being scored correctly. IRT calibrations were then performed, and item summaries, as described in Section 6.3, IRT Item Summaries, were calculated. Items with anomalous parameters or flagged for item fit were reviewed by psychometricians and content experts. Any item found to be performing poorly was dropped, though it was encountered very rarely, and the IRT calibration was then rerun. Once the IRT calibration was completed, all parties could proceed with the equating process.

Using the calibrated item statistics from IRTPRO, the complete set of anchor items (all internal and external anchor items) was used to calculate the equating constants to place the 2022 item parameters onto the IRT-calibrated item pool. Internal anchor items are operational and used to calculate student scores. External anchor items are located in embedded field-test slots and do not count toward student scores. The Stocking-Lord procedure was used to complete the equating.

The Stocking-Lord (Stocking & Lord, 1983) procedure is a method commonly used alongside the 3PL model and GPCM and establishes 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 symmetric approach evaluates the following integral, where the index *i* denotes a common item, and subscripts *I* and *J* denote the item parameters for the bank and item parameters to be rescaled:

$$\arg\min SL = \int \left[\sum_{i=1}^{K} E\left(z_{i,I}|\theta_1\right) - \sum_{i=1}^{K} E\left(z_{i,J}^{*}|\theta_1\right)\right]^2 f(\theta_1|\mu,\sigma^2)\, d\theta_1 + \int \left[\sum_{i=1}^{K} E\left(z_{i,I}^{*}|\theta_2\right) - \sum_{i=1}^{K} E\left(z_{i,J}|\theta_2\right)\right]^2 f(\theta_2|\mu,\sigma^2)\, d\theta_2$$

where $f(\theta_1|\mu,\sigma^2)$ is the normal population density associated with putting operational items onto the bank scale and $f(\theta_2|\mu,\sigma^2)$ is the density associated with putting bank items onto the operational scale. Without loss of generality to permit for compact notation, let $E\left(z_{i,I}|\theta\right)$ denote the expected value of response on the ith item from either the binary or partial credit model and let $E\left(z_{i,J}|\theta\right)$ be the same for the items to be rescaled.

Where for dichotomous items we have

$$p\left(z_{i,I}=1|\theta\right) = \mathrm{c}_{\mathrm{i},I} + \frac{1-\mathrm{c}_{\mathrm{i},I}}{1+\exp[-\mathrm{Da}_{\mathrm{i},I}\left(\theta-\mathrm{b}_{\mathrm{i},I}\right)]}$$

and for the polytomous IRT models

$$p\left(z_{i,I}|\theta\right) = \frac{\exp\left(\sum_{\mathrm{k}=0}^{\mathrm{z_i}} \mathrm{Da_i}\left(\theta-\delta_{\mathrm{ki},I}\right)\right)}{\sum_{\mathrm{h}=0}^{\mathrm{m_i}} \exp \sum_{\mathrm{k}=0}^{\mathrm{h}} \mathrm{Da}_{\mathrm{i},I}\left(\theta-\delta_{\mathrm{ki},I}\right)},$$

where z_i denotes score point $\mathrm{z_i} = \{0,1,\ldots,\mathrm{m}_i\}$ to item i. The expected score for the polytomous models is

$$E\left(z_{i,I}|\theta\right) = \sum_{z=1}^{m_i} zp(\mathrm{z}_i|\theta).$$

The symmetric approach uses the reverse transform for the bank items

$$p\left(z_{i,I}^{*}=1|\theta\right) = \mathrm{c}_{\mathrm{i},I} + \frac{1-\mathrm{c}_{\mathrm{i},I}}{1+\exp[-\mathrm{DAa}_{\mathrm{i},I}\left(\theta-\frac{(\mathrm{b}_{\mathrm{i},I}-B)}{A}\right)]}$$

and for the polytomous IRT models

$$p\left(z_{i,I}^{*}|\theta\right) = \frac{exp\left(\sum_{k=0}^{z_i} DAa_i\left(\theta-\frac{(\delta_{ki,I}-B)}{A}\right)\right)}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h} DAa_{i,I}\left(\theta-\frac{(\delta_{ki,I}-B)}{A}\right)},$$

Once the equating constants were estimated, they were applied to *maximum a posteriori* (MAP) ability estimates, which were derived using IRTPRO, to project the percentage of students, based on the calibration sample, who were likely to score in each performance category. This initial equating solution was referred to as the baseline solution for each grade and subject combination.

After the baseline solution was estimated, two iterative procedures were implemented. The first procedure dropped one item from the equating set per iteration, resulting in a new slope and intercept that was plotted for review. The second procedure started with the baseline solution and cumulatively dropped extreme items from the equating set. This second procedure was implemented via the following steps:

1) Rescaled the current year item parameters to be on the IRT-calibrated item pool using transformation constants based on the Stocking-Lord procedure
2) Computed the weighted area between the item characteristic curves (ICC), a method known as D^2 or the mean squared difference (MSD)
3) Computed the mean and standard deviation of the MSD and standardized the MSD to get SMSD (standardized MSD)
4) Ordered all equating items by |SMSD|
5) Identified "extreme" items as any item with |SMSD| > 2.5 and removed item(s) with max|SMSD| from equating set
6) Iteratively removed items in the linking set until $SMSD < 2.5 \; \forall \; i$

The D^2, or the MSD, is computed by integrating out θ as follows:

$$D^2 = \int \left(E(z_{i,J}|\theta) - E(z_{i,I}|\theta)\right)^2 f(\theta; \mu, \sigma^2) d\theta.$$

The D^2 integral does not have a closed form solution, and so its approximation is based on the weighted summation over q={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,

$$\mathrm{D}^2 = \sum_{q=1}^{30} w_q \left(E\left(z_{i,J}|\theta_q\right) - E\left(z_{i,I}|\theta_q\right)\right)^2.$$

The iterative nature of this process provided for a baseline solution with a projection of its impact on the population percentage of students scoring in each performance level, as well as additional solutions based on the number of extreme items removed. The number of additional solutions conducted depended on the number of extreme items in the equating set. For each additional solution, population impact statistics were provided.

The process described in this section was automated via CAI's equating software. After the initial equating solution and two iterative procedures were complete, CAI and Pearson produced an equating report to deliver to FDOE. A sample calibration report, a sample calibration summary report, and a sample equating report produced by CAI and Pearson can be found in Appendix I, Calibration, Anchor, and Equating Reports. Upon review of these solutions and discussion during a calibration and equating call including HumRRO and Buros, FDOE and TDC were also able to request removal of additional anchor items based on a variety of factors. These factors included, but were not limited to: (1) content review of any flagged item; (2) item position shift; (3) item fit plots; (4) results of anchor item stability checks (e.g., D^2); (5) individual and cumulative impact

of anchor items on the scale transformation coefficients; (6) scatterplots of old and new IRT parameters and their correlation coefficients; (7) scatterplots of the resulting transformation coefficients after removing one anchor item at a time; and (8) evaluation of pre- vs. post-equated item characteristic curves. Interested readers can refer to the 2022 Calibration and Scoring Specifications for details of all statistical analyses and evaluation protocols to be followed during operational work.

While the transition from online to paper-based forms occurred in spring 2019, three anchor items in Grade 4 ELA and two anchor items in Grade 6 Mathematics came from an online administration prior to spring 2019. Such items were reviewed for sensitivity to mode change, and listed in the 'watch list' of the spring 2021 calibration and scoring specifications. For example, transferring an item from online to a paper form might result in the following changes: (1) an item may have different physical appearance on paper versus online (e.g., font, layout, graphics); and (2) instead of providing answers on screen by typing or using available online tools, students need to fill in answers on bubble sheets or write on test booklets. CAI and Pearson accommodated requests by dropping items identified by FDOE and TDC following calibration and equating calls. The final determination of the linking sets was made by FDOE and TDC. It is important to note that this process resulted only in items being dropped from anchor sets solely for equating, not for scoring. The equating report was updated with new solutions after each request was completed.

Table 15 shows the final equating results. The number of items in the equating design is shown, as well as the number of dropped items and the number of items in the final equating solution. The last two columns show the slope and intercept from the final Stocking-Lord equating solution. A trend analysis study was conducted to investigate potential impacts of the modality change, the report of which can be found in Volume 7 of the *Florida Standards Assessments 2018–2019 Technical Report*.

In equating, there are two possible sources of error: sampling error and equating error (Phillips, 2010). Sampling error exists given that calibrations and equating methods are performed on a sample of students drawn from the population. Our sampling design minimizes the design effect that arises from the clustering of students within a group (Phillips, 2010) and uses a stratified random sample of students from across the state. This sampling is described in this section and in Appendix D, EPS Sampling Plan, of this volume of the technical report.

A second, and potentially larger, source of variance is due to the sampling of common items. The items chosen to link the test forms are a sample of only the items that could have been used to establish the linkage. That is, these items are not treated as fixed, but as a random draw from the universe of potential linking items. Had different items been chosen, a different equating solution would have been found and the degree to which this varies due to the common items can be a very large source of potential error variance (Michaelides and Haertel, 2004). The source of such error is explored during the equating work by dropping items one by one from the anchor set and recalculating the slope and intercept. The final distribution of slopes and intercepts was reviewed to see if any single item had a large impact. This process was included in the reports and can be seen on page 25 of Appendix I, Calibration, Anchor, and Equating Reports. Cohen, Johnson, and Angeles (2000) found that the error associated with the uncertainty of IRT parameter estimates resulted in a 25% to 100% increase in standard errors. Error due to the sampling of items is reduced as the number of linking items increases (Michaelides and Haertel, 2004). The uncertainty due to the sampling of items is unaffected by an increase in sample size.

Equating results in the table represent the final solutions used for FSA and NGSSS equating. The intercept and slope represent the first and second moments of the ability distribution, respectively. Hence, slope values greater than 1 indicate greater heterogeneity in the population relative to the baseline year, and values less than 1 indicate greater homogeneity than previously observed. Similarly, intercept values greater than 0 indicate an improvement in mean performance relative to the baseline group and values less than 0 denote the opposite. It is important to note that the column, *Number of Items in Design,* in Table 15 refers to the size of the equating set for a given test. A matrix sampling design is used to collect student responses, in which each student receives a portion of the equating set.

Table 15: Final Equating Results

Subject	Grade	Number of Items in Design	Number of Items Dropped	Number of Items in Final Solution	Slope	Intercept
ELA	3	37	4	33	1.00950	-0.01585
	4	40	1	39	1.11103	0.03508
	5	37	0	37	1.15421	-0.00605
	6	40	0	40	1.15159	0.00095
	7	40	0	40	1.11410	-0.09270
	8	36	0	36	1.17539	-0.15797
	9	36	0	36	1.11559	-0.02181
	10	36	0	36	1.13134	-0.03545
Mathematics	3	40	0	40	1.10144	-0.01445
	4	40	0	40	1.12953	0.04344
	5	40	0	40	1.19540	-0.10195
	6	40	0	40	1.11164	-0.11368
	7	40	2	38	1.06794	-0.12619
	8	40	0	40	1.11170	-0.34161
Science	5	32	0	32	1.15043	-0.14973
	8	32	0	32	1.07117	-0.10571
EOC	Algebra 1	33	0	33	1.12072	-0.11138
	Geometry	30	0	30	1.09369	-0.10585
	Biology 1	25	0	25	1.10678	0.07831
	U.S. History	23	0	23	1.08114	0.29094
	Civics	24	0	24	1.09536	0.28702

6.2.2 Paper Accommodated Forms

During spring 2022, the paper accommodated forms were scored using item parameters from the item's latest spring administration. The accommodated paper form contains the same items as the online core form for Grades 7, 8, 10 ELA, Biology 1, Civics, and U.S History. For Mathematics

and Algebra 1 and Geometry EOC, while most items overlapped between the core online form and the accommodated forms, some items had to be replaced due to the technology being used with online forms. The paper accommodated items that were common with the online forms used item parameters from the spring 2022 online calibrations, and all other items used item parameters from previous online administrations.

The paper accommodated forms were automatically equated to the IRT calibrated item pool since item parameters came from previously equated spring 2021, spring 2019, spring 2018, spring 2017, spring 2016, or spring 2015 item parameters.

6.2.3 Census Paper Form

During spring 2022, students in Grades 3–6 Reading and Mathematics were tested entirely on paper. The classical item analysis, IRT calibration, and equating were conducted using the EPS and followed the same procedure as described in Section 6.2.1, Online Forms. Several factors were considered during the paper form review process. These factors included, but were not limited to: (1) content review of any flagged item; (2) item position shift; (3) item fit plots; (4) results of anchor item stability checks (e.g., D^2); (5) individual and cumulative impact of anchor items on the scale transformation coefficients; (6) scatterplots of old and new IRT parameters and their correlation coefficients; (7) scatterplots of resulting transformation coefficients after removing one anchor item at a time; and (8) evaluation of pre-equated vs. post-equated item characteristic curves. Readers can refer to the 2022 Calibration and Scoring Specifications for details of all statistical analyses and evaluation protocols to be followed during operational work. The number of dropped items listed in Table 15 was partially reflective of the decisions from this process.

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, 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

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

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

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

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

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

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 ZQ1:

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

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 ZQ1 is greater than ZQ_{crit}.

The number of items flagged by Q1 can be found in Appendix A for operational items, Appendix B for anchor items, and Appendix C for field-test items.

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

Appendix B, Anchor Item Statistics, lists the number of anchor items flagged by Q1 after removal of items that were dropped from equating. These items were reviewed by psychometricians and content specialists before a final decision was made about their inclusion to calculate the equating constants.

Appendix C, Field-Test Item Statistics, lists the number of field-test items by grade and subject flagged by Q1. Before field-test items are placed onto forms for operational use in future administrations, they will be reviewed by content specialists and psychometricians. More information about test construction and item review can be found in Volume 2 of this technical report.

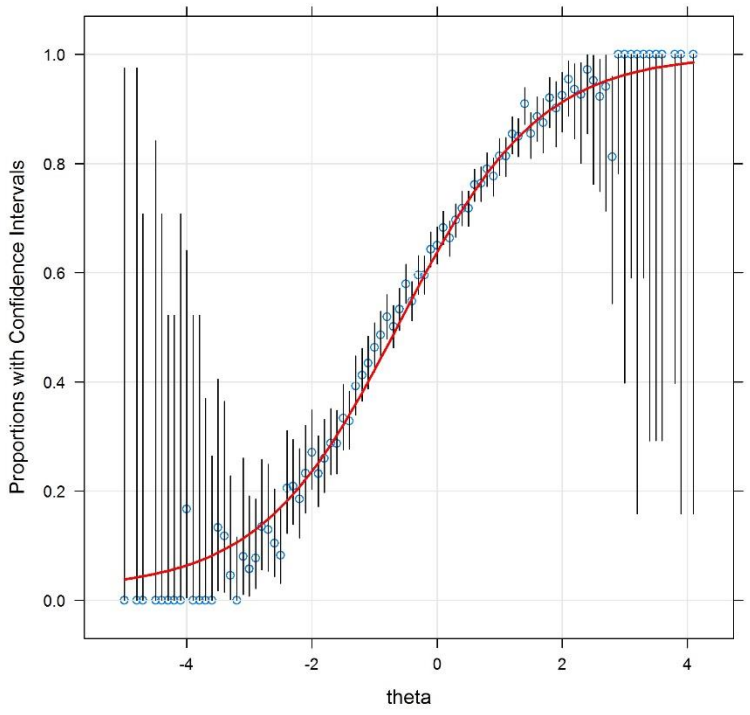
6.3.2 Item Fit Plots

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 upon request. Along with classical item statistics and Q1 flags, item fit plots were used to review items.

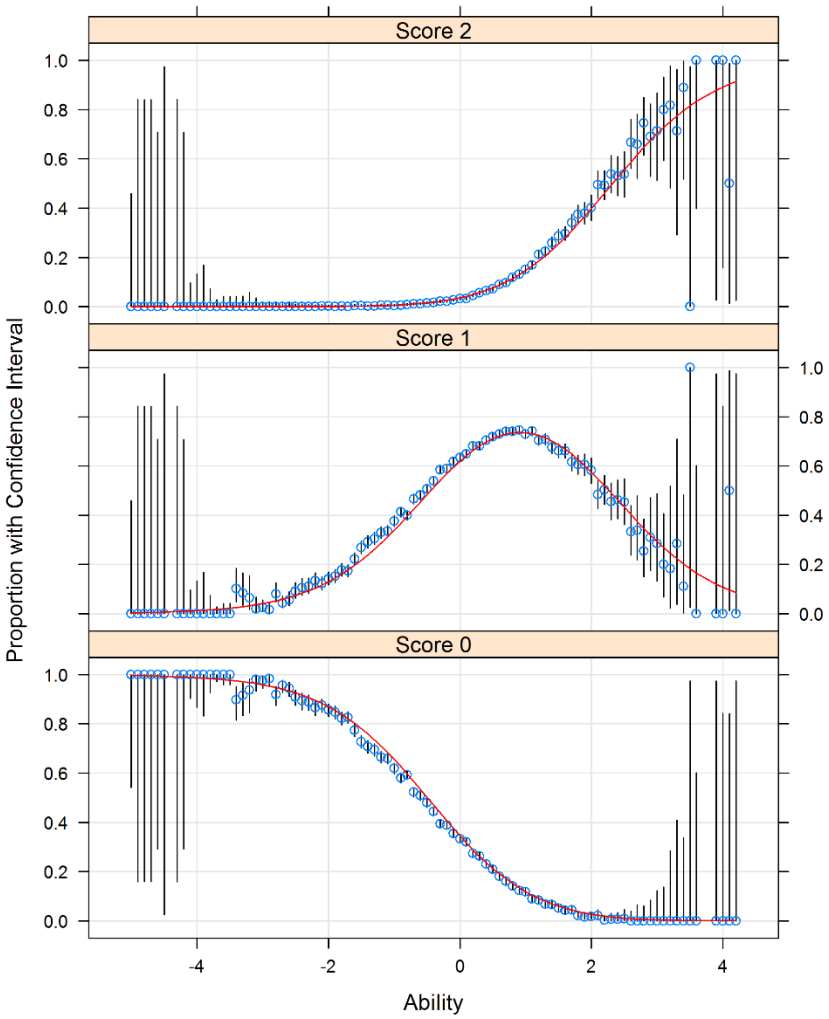
The fit plot in Figure 1 illustrates a one-point item that fits the item response model well. The blue dots represent the proportion of students within a score bin correctly answering the item. The red solid line is the IRT-based item characteristic curve. The black lines indicate the error bands associated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

Figure 1: Example Fit Plot—One-Point Item



The plot in Figure 2 is provided for items worth two points or more. Again, the red lines are 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 2: Example Fit Plot—Two-Point Item



6.4 RESULTS OF CALIBRATIONS

This section presents a summary of the results from the classical item analysis and IRT analysis described in Section 5, Item Analysis Overview, for the spring 2022 operational and field-test items. The summaries here are aggregates; item-specific details are found in the appendices.

Table *16* to Table 19 provide summaries of the *p*-values by percentile as well as the range by grade and subject for operational items. Note that the column *Total OP Items* shows the number of items that were used in the computation of the percentiles. As noted in Section 1.4 above, there were multiple operational forms for the EOC assessments.

Table 20 to Table 23 present the summary and range of the operational item parameters across all forms. The field-test item summaries after excluding the dropped items can be found in Appendix C, Field-Test Item Statistics.

Table 16: Operational Item p-*Value Five-Point Summary and Range, Mathematics*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	54	0.23	0.37	0.51	0.69	0.78	0.83	0.85
4	54	0.19	0.31	0.46	0.58	0.68	0.85	0.89
5	54	0.19	0.29	0.44	0.49	0.63	0.73	0.85
6	56	0.15	0.16	0.34	0.45	0.58	0.75	0.83
7	56	0.14	0.15	0.22	0.31	0.45	0.70	0.80
8	56	0.10	0.12	0.19	0.30	0.43	0.69	0.72

Table 17: Operational Item p-*Value Five-Point Summary and Range, ELA*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	50	0.24	0.31	0.45	0.54	0.66	0.74	0.80
4	53	0.29	0.34	0.49	0.60	0.72	0.82	0.84
5	53	0.27	0.36	0.49	0.64	0.72	0.88	0.91
6	55	0.25	0.31	0.42	0.57	0.67	0.81	0.92
7	55	0.22	0.33	0.45	0.55	0.62	0.76	0.86
8	55	0.16	0.32	0.48	0.57	0.69	0.81	0.90
9	57	0.17	0.33	0.42	0.54	0.64	0.76	0.84
10	57	0.30	0.37	0.47	0.56	0.67	0.80	0.88

Table 18: Operational Item p-*Value Five-Point Summary and Range, EOC*

Grade	Total OP Items*	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Algebra 1	136	0.09	0.14	0.26	0.37	0.50	0.67	0.72
Geometry	114	0.11	0.18	0.27	0.38	0.49	0.66	0.74
Biology 1	117	0.32	0.38	0.44	0.52	0.60	0.70	0.84
Civics	96	0.25	0.37	0.49	0.60	0.68	0.81	0.88
U.S. History	106	0.29	0.38	0.48	0.55	0.64	0.76	0.87

*Note that operational items across all forms were combined.

Table 19: Operational Item p*-Value Five-Point Summary and Range, Science*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	56	0.38	0.45	0.59	0.66	0.74	0.83	0.86
8	56	0.45	0.47	0.55	0.62	0.68	0.78	0.84

Table 20 to Table 23 give the 2PL, 3PL, and GPCM item parameter summaries for Mathematics, ELA, EOC, and Science. If fewer than 10 items existed in a model type for a given test, only the number of items, minimum, and maximum are given.

Table 20: Operational Item Parameter Five-Point Summary and Range, Mathematics

Grade	IRT Model	Parameter	Number of Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	2PL & 3PL	a	54	0.54	0.69	0.84	1.01	1.20	1.40	1.50
		b	54	-1.58	-1.50	-1.00	-0.48	0.21	0.76	1.14
		c	38	0	0.01	0.06	0.17	0.25	0.35	0.37
4	2PL & 3PL	a	54	0.56	0.67	0.87	1.03	1.18	1.45	1.84
		b	54	-2.07	-1.55	-0.45	-0.02	0.50	0.89	1.41
		c	38	0	0.01	0.03	0.14	0.22	0.30	0.42
5	2PL & 3PL	a	54	0.51	0.58	0.67	0.91	1.18	1.73	2.27
		b	54	-1.75	-1.25	-0.36	0.12	0.54	1.15	1.58
		c	38	0	0.01	0.09	0.17	0.21	0.34	0.40
6	2PL & 3PL	a	56	0.54	0.62	0.88	1.00	1.15	1.44	1.60
		b	56	-1.17	-0.74	-0.18	0.45	0.94	1.31	1.40
		c	46	0	0.01	0.07	0.18	0.22	0.32	0.38
7	2PL & 3PL	a	53	0.46	0.53	0.84	1.01	1.20	1.38	1.79
		b	53	-1.40	-0.81	0.52	0.84	1.06	1.51	2.28
		c	31	0	0.01	0.08	0.11	0.26	0.36	0.40
	GPCM	A	3	0.58	-	-	-	-	-	0.70
		D1	3	-0.17	-	-	-	-	-	1.44
		D2	3	0.76	-	-	-	-	-	2.08
8	2PL & 3PL	a	56	0.48	0.51	0.65	0.77	0.94	1.12	1.36
		b	56	-1.30	-0.94	0.36	0.99	1.27	1.76	1.95
		c	37	0	0	0.04	0.11	0.20	0.32	0.35

Note: IRT scaling constant D=1.7

Table 21: Operational Item Parameter Five-Point Summary and Range, ELA

Grade	IRT Model	Parameter	Number of Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	2PL & 3PL	a	50	0.42	0.51	0.73	0.93	1.22	1.55	1.65
		b	50	-1.15	-0.84	-0.27	0.25	0.72	1.08	1.70
		c	50	0.01	0.01	0.06	0.15	0.21	0.26	0.34
4	2PL & 3PL	a	50	0.19	0.38	0.60	0.81	0.93	1.17	1.39
		b	50	-1.85	-1.31	-0.59	0.03	0.33	0.94	1.88
		c	50	0	0.01	0.04	0.13	0.21	0.26	0.34
	GPCM	a	3	0.74	-	-	-	-	-	0.85
		D1	3	-2.02	-	-	-	-	-	-0.95
		D2	3	-0.59	-	-	-	-	-	1.42
5	2PL & 3PL	a	50	0.24	0.49	0.70	0.78	0.96	1.15	1.23
		b	50	-2.58	-1.87	-0.82	-0.28	0.28	1.31	2.17
		c	50	0.01	0.02	0.05	0.14	0.23	0.27	0.35
	GPCM	a	3	0.76	-	-	-	-	-	0.96
		D1	3	-2.08	-	-	-	-	-	-1.34
		D2	3	-0.71	-	-	-	-	-	1.23
6	2PL & 3PL	a	52	0.17	0.48	0.61	0.73	0.90	1.24	1.39
		b	52	-2.23	-1.54	-0.49	0.11	0.53	1.98	2.40
		c	52	0.01	0.02	0.08	0.17	0.22	0.33	0.39
	GPCM	a	3	0.75	-	-	-	-	-	0.91
		D1	3	-2.41	-	-	-	-	-	-1.86
		D2	3	-1.65	-	-	-	-	-	0.92
7	2PL & 3PL	a	52	0.24	0.39	0.64	0.80	1.01	1.36	1.55
		b	52	-1.81	-1.21	-0.50	0.14	0.73	1.39	2.08
		c	52	0	0.01	0.06	0.17	0.27	0.31	0.39
	GPCM	a	3	0.86	-	-	-	-	-	0.96
		D1	3	-2.02	-	-	-	-	-	-1.29
		D2	3	-0.81	-	-	-	-	-	0.77
8	2PL & 3PL	a	52	0.38	0.41	0.60	0.77	1.00	1.31	1.49
		b	52	-2.50	-1.34	-0.63	-0.14	0.37	0.93	1.78
		c	52	0.02	0.02	0.08	0.19	0.24	0.29	0.54
	GPCM	a	3	0.82	-	-	-	-	-	0.82
		D1	3	-2.62	-	-	-	-	-	-1.99
		D2	3	-0.99	-	-	-	-	-	0.44

Grade	IRT Model	Parameter	Number of Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
9	2PL & 3PL	a	54	0.43	0.50	0.65	0.80	0.97	1.35	1.44
		b	54	-1.25	-0.98	-0.34	0.28	0.67	1.19	2.30
		c	53	0	0.01	0.05	0.12	0.23	0.35	0.41
	GPCM	a	3	0.96	-	-	-	-	-	0.98
		D1	3	-2.12	-	-	-	-	-	-1.65
		D2	3	-1.03	-	-	-	-	-	0.45
10	2PL & 3PL	a	54	0.25	0.39	0.58	0.73	0.86	1.09	1.27
		b	54	-2.29	-1.33	-0.54	-0.08	0.48	0.99	1.47
		c	54	0.01	0.01	0.05	0.15	0.22	0.32	0.35
	GPCM	a	3	1.02	-	-	-	-	-	1.06
		D1	3	-2.47	-	-	-	-	-	-1.71
		D2	3	-1.28	-	-	-	-	-	0.84

Note: IRT scaling constant D=1.7

Table 22: Operational Item Parameter and Five-Point Summary and Range, EOC

Grade	IRT Model	Parameter	Number of Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Algebra 1	2PL & 3PL	a	136	0.40	0.61	0.80	1.02	1.21	1.48	1.88
		b	136	-1.05	-0.55	0.37	1.05	1.41	1.89	2.09
		c	112	0.00	0.00	0.06	0.16	0.23	0.32	0.41
	GPCM	a	-	-	-	-	-	-	-	-
		D1	-	-	-	-	-	-	-	-
		D2	-	-	-	-	-	-	-	-
Geometry	2PL & 3PL	a	113	0.50	0.63	0.85	1.12	1.38	1.74	2.10
		b	113	-0.85	-0.46	0.54	0.92	1.20	1.65	1.98
		c	100	0.00	0.01	0.09	0.15	0.25	0.37	0.48
	GPCM	a	1	0.66	-	-	-	-	-	0.66
		D1	1	0.10	-	-	-	-	-	0.10
		D2	1	1.37	-	-	-	-	-	1.37
Biology 1	3PL	a	117	0.38	0.52	0.66	0.85	1.08	1.29	1.56
		b	117	-1.20	-0.75	0.03	0.62	0.97	1.41	1.66
		c	117	0.01	0.02	0.13	0.20	0.26	0.33	0.42
Civics	3PL	a	96	0.42	0.56	0.75	0.94	1.16	1.45	1.52
		b	96	-1.54	-0.81	-0.23	0.42	0.96	1.60	1.96
		c	96	0.00	0.02	0.13	0.20	0.26	0.33	0.35
U.S. History	3PL	a	106	0.42	0.51	0.68	0.77	1.01	1.20	1.99
		b	106	-1.38	-0.64	-0.08	0.55	1.01	1.40	1.81
		c	106	0.01	0.02	0.12	0.19	0.24	0.33	0.36

Note: IRT scaling constant D=1.7

Table 23: Operational Item Parameter and Five-Point Summary and Range, Science

Grade	IRT Model	Parameter	Number of Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	3PL	a	56	0.43	0.50	0.67	0.77	0.93	1.10	1.27
		b	56	-1.96	-1.73	-1.00	-0.48	-0.07	0.90	1.12
		c	56	0.01	0.02	0.13	0.20	0.22	0.26	0.35
8	3PL	a	56	0.54	0.55	0.80	0.92	1.08	1.26	1.45
		b	56	-1.50	-1.19	-0.74	-0.12	0.17	0.62	0.87
		c	56	0.01	0.03	0.15	0.18	0.23	0.29	0.36

Note: IRT scaling constant D=1.7

7. SUMMARY OF ADMINISTRATION

This chapter provides the summary of Florida Statewide Assessments tests administered in spring 2022. It covers item and test characteristic curves, estimates of classification accuracy and consistency, and reporting scales.

7.1 ITEM AND TEST CHARACTERISTIC CURVES

An item characteristic curve (ICC) shows the probability of a correct response as a function of ability given an item's parameters. Test characteristic curves (TCCs) can be constructed as the sum of ICCs for the items included on the test. The TCC can be used to determine test taker raw scores or percentage-correct scores that are expected at given ability levels. When two tests are developed to measure the same ability, their scores can be equated using TCCs. As such, it is useful to use TCCs during test construction. Items are selected for a new form so that the new form's TCC matches the target form's TCC as closely as possible.

The figures in Appendix E, Test Characteristic Curves, show the TCCs by grade and subject, based on the final operational item parameters from the spring 2022 calibrations.

7.2 ESTIMATES OF CLASSIFICATION ACCURACY AND CONSISTENCY

See Classification Accuracy and Consistency results in Section 3.4 of this technical report, Volume 4, Evidence of Reliability and Validity.

7.3 REPORTING SCALES

For spring 2022, the Mathematics, ELA, Science, and EOC tests report scale scores for each student. The score is based on the operational items presented to the student.

Appendix F, Distribution of Scale Scores and Standard Errors, provides a summary of the scale scores.

8. SCORING

This chapter provides the scoring procedure used in Florida Statewide Assessments tests administered in the 2021–2022 school year. It covers the computational details of the maximum likelihood estimation (MLE), standard error of estimate, scale scores, performance level, and subscores reported in Florida Statewide Assessments tests.

8.1 FLORIDA STATEWIDE ASSESSMENTS SCORING

8.1.1 Maximum Likelihood Estimation

The Florida Statewide Assessments tests were based on the three-parameter logistic (3PL) model and generalized partial credit model (GPCM) of item response theory models, with the two-parameter logistic (2PL) model treated as a special case of the 3PL. 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 items 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 kth step for item *i* with m total categories, and $D = 1.7$.

A student's theta based on 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 of the 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 model, and N_{3PL} is the number of items scored using the 3PL or 2 PL 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

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

where N_{CR} is the number of items that are scored using the GPCM model, and N_{3PL} is the number of items scored using the 3PL or 2 PL 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 Florida Statewide Assessments 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 Statewide Assessments tests included items that were scored using the 3PL, 2PL, and GPCM from IRT. The 2PL 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 model, and N_{3PL} is the number of items scored using the 3PL or 2 PL model.

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

8.1.2 Scale Scores

There are two scale types created for the Florida Statewide Assessments:

- A vertical scale score for Grades 3–10 ELA and Grades 3–8 Mathematics
- A within-test scaled score for Science Grades 5 and 8, and all EOC tests

Table 24 shows the theta to scaled score transformation equations.

Table 24: Theta to Scale Score Transformation Equations

Subject	Grade	Theta to Scale Score Transformation
ELA	3	Scale Score = round(theta *20.000000 + 300.000000)
	4	Scale Score = round(theta *20.237420 + 311.416960)
	5	Scale Score = round(theta *21.230040 + 320.961420)
	6	Scale Score = round(theta *21.861120 + 325.061500)
	7	Scale Score = round(theta *21.581900 + 332.124320)
	8	Scale Score = round(theta *21.531360 + 338.432720)
	9	Scale Score = round(theta *21.751840 + 341.749740)
	10	Scale Score = round(theta *21.284300 + 348.328540)
Mathematics	3	Scale Score = round(theta *20.000000 + 300.000000)
	4	Scale Score = round(theta *20.899320 + 313.617800)
	5	Scale Score = round(theta *22.050760 + 321.802560)
	6	Scale Score = round(theta *21.684500+ 325.299220)
	7	Scale Score = round(theta *20.379620 + 330.157540)
	8	Scale Score = round(theta *19.952780 + 332.946420)
Algebra 1		Scale Score= round(theta *25.000000 + 500.000000)
Geometry		Scale Score = round(theta *25.000000 + 500.000000)
Science	5	Scale Score = round(theta *20.000000 + 200.000000)
	8	Scale Score = round(theta *20.000000 + 200.000000)

Subject	Grade	Theta to Scale Score Transformation
Biology 1		Scale Score = round(theta *25.000000 + 400.000000)
U.S. History		Scale Score = round(theta *25.000000 + 400.000000)
Civics		Scale Score = round(theta *25.000000 + 400.000000)

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

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

The standard error 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 24

8.1.3 Performance Levels

Each student is assigned a performance category according to his or her accountability scale score. Table 25 to Table 28 provide the cut scores for performance levels for Mathematics, ELA, Science, and EOC.

Table 25: 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	285	297	311	327
4	299	310	325	340
5	306	320	334	350
6	310	325	339	356
7	316	330	346	360
8	322	337	353	365

Table 26: Cut Scores for ELA 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	285	300	315	330
4	297	311	325	340
5	304	321	336	352
6	309	326	339	356

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
7	318	333	346	360
8	322	337	352	366
9	328	343	355	370
10	334	350	362	378

Table 27: Cut Scores for 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	487	497	518	532
Geometry	486	499	521	533
Biology 1	369	395	421	431
U.S. History	378	397	417	432
Civics	376	394	413	428

Table 28: Cut Scores for Science 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
5	185	200	215	225
8	185	203	215	225

8.1.4 Alternate Passing Score

The alternate passing score (APS) is the FCAT 2.0-equivalent score reported as an FSA scaled score. When Grade 10 ELA and EOC cut scores were reported in 2015, there was no approved FSA reporting scale, and so cut scores were reported as an FCAT 2.0 equivalent. The FSA scale transformation constants are now known, and so the passing scores can be reported on the FSA scale. Since the cuts recommended from the summer 2015 standard setting process have been approved, it is important to note that these APS cuts are used with only students who are retaking the test.

Equipercentile linking was used to find the FCAT 2.0 linked score, and this methodology relied on using an FCAT-looking score. The FCAT-looking score is the student's MLE transformed to be on a scale that uses the same transformation constants as the FCAT 2.0. Let $\hat{\theta}_s$ denote the FCAT-looking score for test *s* from the 2015 linking score conversion table. The APS is then found as

$$APS_{algebra} = \left[\frac{\hat{\theta}_{alg} - 400}{25}\right] * 25 + 500$$

$$APS_{geometry} = \left[\frac{\hat{\theta}_{geo} - 400}{25}\right] * 25 + 500$$

$$APS_{ela} = \left[\frac{\hat{\theta}_{e} - 244.870126}{18.822290}\right] * 21.284300 + 348.328540$$

The FSA score that corresponds to the cut score used for passing in 2015 is then found. These scores are shown in Table 29.

Table 29: Alternate Passing Score Cut Points

Test	APS	FCAT-Linked Score	FCAT 2.0 and NGSSS EOC Looking Scales
Grade 10 ELA	349	245	245
Algebra 1	489	399	389
Geometry	492	396	392

Note that 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.

In Grade 10 ELA, the APS is 349 and the scale score cut point for Level 3 is 350. If a Grade 10 ELA student scores 349, he or she receives a passing status of Y and a performance level of 2.

More information can be found in Section 6.3 of Volume 1 of the *Florida Standards Assessments 2014–2015 Technical Report*.

8.1.5 Reporting Category Scores

In addition to overall scores, students also receive scores on reporting categories. Let b_{sq} represent the subset of operational items presented to student *s* in reporting category q. Students will receive a raw score for each reporting category, with these scores being derived using only b_{sq}. That is, the raw score is calculated as the sum of the scores on the subset of operational items measuring reporting category q. The number of raw score points for each test and reporting category is provided in Appendix G, Distribution of Reporting Category Scores, along with summaries of scores.

9. STATISTICAL SUMMARY OF TEST ADMINISTRATION

This chapter provides the demographics of the tested population in English Language Arts (ELA), Mathematics, EOC, and Science tests administered in spring 2022.

9.1 DEMOGRAPHICS OF TESTED POPULATION

Table 30 to Table 33 present the distribution of students, in counts and in percentages, who participated in the spring administration of the 2021–2022 Florida Statewide Assessments by grade and subject. The numbers presented here are based on the reported status in the approved spring State Student Results (SSR) files. The subgroups reported are gender, ethnicity, students with disabilities (SWD), and English language learners (ELL). Section 1.2 of Volume 5 of this technical report provides explicit definitions for the two major subgroups to which accommodations are available: ELL and SWD.

Table 30: Distribution of Demographic Characteristics of Tested Population, Mathematics

Grade	Group	All Students	Female	Male	African–American	Hispanic	White	SWD	ELL
3	*N*	207,531	101,004	106,527	43,757	74,721	73,442	24,367	31,900
	%	100	48.67	51.33	21.08	36.00	35.39	11.74	15.37
4	*N*	195,047	96,284	98,763	39,988	70,070	70,161	23,828	24,676
	%	100	49.36	50.64	20.50	35.92	35.97	12.22	12.65
5	*N*	210,709	102,750	107,959	44,737	77,500	73,172	30,283	24,555
	%	100	48.76	51.24	21.23	36.78	34.73	14.37	11.65
6	*N*	185,275	91,531	93,744	38,524	67,857	65,870	25,857	16,763
	%	100	49.40	50.60	20.79	36.63	35.55	13.96	9.05
7	*N*	171,011	83,655	87,356	38,461	63,684	58,095	25,132	15,646
	%	100	48.92	51.08	22.49	37.24	33.97	14.70	9.15
8	*N*	150,778	73,025	77,753	36,223	58,379	47,071	24,398	15,353
	%	100	48.43	51.57	24.02	38.72	31.22	16.18	10.18

Table 31: Distribution of Demographic Characteristics of Tested Population, ELA

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
3	*N*	210,396	102,302	108,094	44,533	75,967	74,135	24,624	32,173
	%	100	48.62	51.38	21.17	36.11	35.24	11.70	15.29
4	*N*	198,594	97,972	100,622	40,615	71,399	71,438	24,111	24,764
	%	100	49.33	50.67	20.45	35.95	35.97	12.14	12.47
5	*N*	212,492	103,658	108,834	45,196	78,367	73,574	30,604	24,544
	%	100	48.78	51.22	21.27	36.88	34.62	14.40	11.55
6	*N*	197,122	97,207	99,915	40,625	71,682	70,455	26,420	16,913
	%	100	49.31	50.69	20.61	36.36	35.74	13.40	8.58
7	*N*	207,191	101,897	105,294	43,234	75,940	73,158	26,097	16,129
	%	100	49.18	50.82	20.87	36.65	35.31	12.60	7.78
8	*N*	213,464	104,609	108,855	44,251	78,460	75,668	26,271	15,445
	%	100	49.01	50.99	20.73	36.76	35.45	12.31	7.24
9	*N*	209,276	104,269	105,007	42,029	75,803	76,742	22,342	14,032
	%	100	49.82	50.18	20.08	36.22	36.67	10.68	6.71
10	*N*	203,493	101,673	101,820	41,744	73,080	74,402	22,140	12,773
	%	100	49.96	50.04	20.51	35.91	36.56	10.88	6.28

Table 32: Distribution of Demographic Characteristics of Tested Population, EOC

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
Algebra 1	*N*	217,541	106,429	111,112	44,240	79,184	78,954	24,604	15,815
	%	100	48.92	51.08	20.34	36.40	36.29	11.31	7.27
Geometry	*N*	191,298	95,206	96,092	38,927	68,915	69,911	20,171	11,735
	%	100	49.77	50.23	20.35	36.02	36.55	10.54	6.13
Biology 1	*N*	208,677	103,215	105,462	42,205	75,441	76,200	22,334	13,142
	%	100	49.46	50.54	20.23	36.15	36.52	10.70	6.30
Civics	*N*	211,533	103,966	107,567	43,811	77,242	75,107	26,414	16,618
	%	100	49.15	50.85	20.71	36.52	35.51	12.49	7.86
U.S. History	*N*	180,243	89,634	90,609	37,324	64,583	65,903	18,611	12,002
	%	100	49.73	50.27	20.71	35.83	36.56	10.33	6.66

Table 33: Distribution of Demographic Characteristics of Tested Population, Science

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
5	*N*	211,831	103,297	108,534	44,878	78,351	73,292	30,482	24,735
	%	100	48.76	51.24	21.19	36.99	34.60	14.39	11.68
8	*N*	199,034	97,217	101,817	41,071	71,807	72,363	25,369	15,173
	%	100	48.84	51.16	20.64	36.08	36.36	12.75	7.62

10. 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.

10.1 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. Pearson follows similar quality assurance procedures.

Prior to 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 file was finalized, it was passed to two psychometricians who used the files for all analyses independently. Each psychometrician independently implemented the 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. During calibrations, daily calls were held with CAI, Pearson, FDOE, Test Development Center (TDC), HumRRO, and Buros to discuss classical statistics and IRT analyses. Content experts from CAI, Pearson, and TDC also reviewed classical statistics and gave input to the discussion. Results were approved by FDOE only when there was replication and verification from all parties.

The daily calibration calls were an important source for quality control and proceeded in an iterative fashion. Typically, two to three tests were evaluated during the calls, reviewing all the evidence on item quality, including classical analyses, IRT-based statistics and fit statistics, fit plots, and in many cases, reviewing the content of the item in a web-based setting.

During these calls, the teams discussed any observed issues or concerns with flagged items and determined if the item suffered from any content or statistical issues that warranted removing it from the set of core items used for scoring.

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

10.2 SCORING QUALITY CHECK

Prior to the operational testing window, CAI's scoring engine was tested to ensure that the maximum likelihood estimations (MLEs) produced by the engine were accurate. This is a process referred to as *mock data*. During mock data, CAI established all systems and simulated item response data as if real students responded to the test items. CAI and Pearson then tested all

programs and verified all results before implementing the operational test. Simulated data were posted to the SFTP site for FDOE, Pearson, HumRRO, and Buros to allow all parties to test their systems.

Once final operational item calibrations were complete and approved by 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, independent score checks were performed by CAI, Pearson, FDOE, and HumRRO. Scores were approved by FDOE only when there was three-way replication and verification.

10.3 SCORE REPORT QUALITY CHECK

PearsonAccess Next Reporting System (PANext Reporting) provides access to Florida assessment results in two main formats. The first format is PDF or Microsoft Excel reports, which provides score data for each of the Florida assessments. Users can compare score data of individual students with the school, district, or overall state average scores. The second format is downloadable into pipe-delimited text data files; this format allows users to download zipped data files containing aggregate data for their district and the state.

Before deploying the reports in PANext Reporting, test cases are designed to verify that scoring and reporting of testing records are performing as intended. All software and interfaces are utilized and executed in the same manner as used for live data. All scoring and reporting outputs (including reports and data files) are validated against expected results to verify scoring and reporting are accurate.

11. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

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

Cohen, J., Johnson, E., & Angeles, J. (2000). *Variance estimation when sampling in two dimensions via the jackknife with application to the National Assessment of Educational Progress.* Washington DC: American Institutes for Research.

Council of Chief State School Officers (2020). *Releases Statement on Assessments in the Next Academic Year*. Washington, DC: CCSSO.

CTB Document A. (1998). Technical Report: Florida Comprehensive Assessment Test (FCAT): 1998, p.13. Monterey, CA: CTB McGraw-Hill.

CTB Document B. (1998). Fit Measurement: A Generalization of Q1. Technical Report: Florida Comprehensive Assessment Test (FCAT): 1998, p.107, Appendix A. Monterey, CA: CTB McGraw-Hill.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service.

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). Hillsdale, NJ: Lawrence Erlbaum.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. *(2nd ed.)* New York, NY: Springer.

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

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

Michaelides, M.P., & Haertel, E.H. (2004). *Sampling of Common Items: An Unrecognized Source of Error in Test Equating* (CSE Report 636). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*(2)*,* 159–176.

Phillips, G.W. (2010). *Score drift: Why district and state achievement results unexpectedly bounce up and down from year to year.* Training session at the National Council for Measurement in Education, Denver, CO.

Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tong, Y., Wu, S-S., & Xu, M. 2008. *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.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory.* New York: Springer-Verlag.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262.

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). Princeton, NJ: Educational Testing Service.