Florida
Standards Assessments

# Florida Standards Assessments

# 2014–2015

# Volume 1
# Annual Technical Report

FLORIDA DEPARTMENT OF
EDUCATION
fldoe.org

# ACKNOWLEDGEMENTS

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

Major contributors to this technical report include the following staff from the American Institutes for Research (AIR): Dr. Harold Doran, Dr. Elizabeth Ayers-Wright, Dr. Dipendra Subedi, Dr. MinJeong Shin, Dr. AhYoung Shin, Danielle Peterson, and Patrick Kozak. The major contributors from the Florida Department of Education are as follows: Dr. Salih Binici, Dr. Molly Hand, Dr. Qian Liu, Vince Verges, Victoria Ash, Susie Lee, Mengyao Cui, Steve Ash, Renn Edenfield, and Chris Harvey.

## TABLE OF CONTENTS

APPENDICES

# LIST OF TABLES

## LIST OF FIGURES

# 1. INTRODUCTION

The Florida Standards Assessments (FSA) technical report is provided to document all methods used in test construction, psychometric properties of the tests, summaries of student results, and evidence and support for its intended uses and interpretations of the test scores. The technical reports are reported as seven separate, self-contained volumes:

1) *Annual Technical Report*. This volume is updated each year and provides a global overview of the tests administered to students each year.
2) *Test Development*. This volume summarizes the procedures used to construct test forms and provides summaries of the item development procedures.
3) *Standard Setting*. This volume documents the methods and results of the FSA standard setting process.
4) *Evidence of Reliability and Validity*. This volume 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*. This volume describes the methods used to administer all forms, security protocols, and modifications or accommodations available.
6) *Score Interpretation Guide*. This volume describes the score types reported and describes the appropriate inferences that can be drawn from each score reported.
7) *Special Studies*. During the course of the year, the Florida Department of Education may request technical studies to investigate issues surrounding the test. This volume is a set of reports provided to the department in support of any requests to further investigate test quality, validity, or other issues as identified.

## 1.1 PURPOSE AND INTENDED USES OF THE FLORIDA STANDARDS ASSESSMENTS

The primary purpose of Florida's K–12 assessment system is to measure students' achievement of Florida's education standards. Assessment supports instruction and student learning, and the results help Florida's educational leadership and stakeholders determine whether the goals of the education system are being met. Assessments help Florida determine whether it has equipped its students with the knowledge and skills they need to be ready for careers and college-level coursework.

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 student learning forward. Florida's assessment and accountability efforts have had a significant positive impact on student achievement over time.

The tests are constructed to meet rigorous technical criteria (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 principles of universal design and appropriate accommodations. The FSA yields test scores that are useful for understanding whether individual students have a firm grasp of the Florida standards and also 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 in program evaluation methods.

Table 1 outlines required uses of the FSA.

*Table 1: Required Uses and Citations for the FSA*

| Assessment | Assessment Citation | Required Use | Required Use Citation |
|---|---|---|---|
| Statewide Assessment Program | s. 1008.22, F.S.<br>Rule 1.09422, F.A.C.<br>Rule 1.0943, F.A.C<br>Rule 1.09432, F.A.C. | Third Grade Retention;<br>Student Progression;<br>Remedial Instruction;<br>Reporting Requirements | s. 1008.25, F.S.<br>Rule 6A-1.094221, F.A.C.<br>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.<br>Rule 6A-1.09981, F.A.C. |
| | | School Improvement Rating | s. 1008.341, F.S.<br>Rule 6A-1.099822, F.A.C. |
| | | District Grades | s. 1008.34, F.S. |
| | | Differentiated Accountability | s. 1008.33, F.S.<br>Rule 6A-1.099811, F.A.C. |
| | | Opportunity Scholarship | s. 1002.38, F.S. |

## 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 the spring of 2015, replacing the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0). It was primarily delivered as an online, fixed-form assessment. Paper forms were administered to students in grades 3 and 4, and paper accommodated versions were available to students whose Individual Education Plans (IEP) or Section 504 Plans indicated such a need.

Within the current Florida statewide assessments program, students in grade 3 must score a Level 2 or higher on the FSA ELA grade 3 assessment in order to be promoted to grade 4. Grade 3 students who score in 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 Level 3 or above on the grade 10 ELA and Algebra 1 EOC assessments in order to meet the assessment graduation requirements set in statute. Students who do not score Level 3 or higher on these assessments have the opportunities for multiple retakes of the assessments, and may also use concordant scores on the ACT or SAT to meet the Grade 10 ELA requirement, or earn a comparative passing score on the Postsecondary Education Readiness Test (PERT) for Algebra 1. Also, students' scores on EOC assessments must count for 30% of a student's final course grade for those courses for which a statewide EOC is administered.

In the rest of this chapter, developments in Florida statewide assessments since 2008 will be highlighted, including the transition from FCAT to FCAT 2.0, the introduction of the EOC assessments, and finally the transition to the FSA. This brief background should establish the legislative and curricular framework for the technical analyses described in the remaining chapters of this volume and other volumes of the technical report.

### Developments in 2008

The 2008 legislation (Senate Bill [SB] 1908) authorizing the extension of the FCAT program and the introduction of EOC assessments was passed based on the experience gathered over four decades of statewide assessment implemented within Florida. One major goal of that legislation was to authorize the State Board of Education (SBE) to establish the Next Generation Sunshine State Standards (NGSSS) to replace the Sunshine State Standards following a routine review of the state's academic standards. SB 1908 also removed the requirement that the statewide assessment program include norm-referenced tests.

The most transformative aspect of SB 1908 was that it allowed for the development and administration of EOC assessments, which were to be administered "within the last 2 weeks of a course."

SB 1908 mandated revisions to what had been known as the FCAT Writing+ program, eliminating multiple-choice writing items beginning in 2009.

Additionally, the SBE accepted a major revision of the 1996 Sunshine State Standards for Social Studies and for Science.

Beginning with the 2010–11 school year, SB 1908 required that the Writing assessment be administered no earlier than the week of March 1 and that the comprehensive statewide assessments of any other subject be administered no earlier than the week of April 15.

### Developments in 2009

In 2009, the revisions of the Sunshine State Standards approved by the SBE in 2007 and 2008 started to be referred to as the 2007 NGSSS and 2008 NGSSS, respectively.

Item development and review by Florida educators began for the FCAT 2.0 assessments in Reading/Language Arts and Mathematics, based on the 2007 NGSSS in those subject areas.

A concordance study was conducted to ensure that the concordant scores used for graduation continued to be equivalent measures of the level of performance expected on the Grade 10 FCAT. As a result of this study, the required SAT Reading score was increased from 410 to 420, the required ACT Reading score was increased from 15 to 18, the required SAT Mathematics score was decreased from 370 to 340, and the required ACT Mathematics score of 15 remained the same. The new concordance score requirements were in effect for students scheduled to graduate in 2011 who had not already earned the previous passing scores by November 30, 2009.

### Developments in 2010

Embedded field testing began for the FCAT 2.0 assessments in Reading/Language Arts and in Mathematics. A sample of students was drawn to take stand-alone field-test forms for the Algebra 1 EOC Assessment.

SB 4 amended the assessment window for the EOC assessments by striking the language “within the last 2 weeks” of the course and adding “during a 3-week period at the end” of the course. This increased flexibility for Florida’s educators, schools, and districts in scheduling assessments.

Item development and review began for the FCAT 2.0 Science Assessment at grades 5 and 8, based upon the 2008 NGSSS for Science. Item development and review also began for the Geometry and Biology 1 EOC Assessments.

SB 4 also amended Section 1008.22 F.S. to require the implementation of EOC assessments at the high school level, which replace the FCAT Mathematics and Science assessments administered in grades 9 and 10 Mathematics and 11 Science, and to require the implementation of an EOC assessment in civics education at the middle school level. SB 4 mandated that the Algebra 1 EOC Assessment be administered beginning in the 2010–11 school year, the Geometry and Biology 1 EOC Assessments be administered beginning in the 2011–12 school year, the U.S. History EOC Assessment be administered beginning in the 2012–13 school year, and the Civics EOC Assessment be administered beginning in the 2013–14 school year. For the first year, these EOC assessments are administered, the EOC results shall constitute 30% of a student’s course grade. With the exception of U.S. History, once standards are established for these EOC assessments, students must pass the assessment to earn course credit. SB 4 also authorized the Commissioner to establish an implementation schedule of EOC assessments in other subject areas, if feasible.

**Developments in 2011**

The first operational administration of FCAT 2.0 Reading and Mathematics and the Algebra 1 EOC Assessment occurred during the spring administration window. Standard-setting meetings for these grade/subject/course combinations occurred with educators in September 2011.

A sample of students was drawn to take stand-alone field-test forms for the Geometry and Biology 1 EOC Assessments. FCAT 2.0 Science field-test items were embedded in grades 5 and 8 FCAT Science forms in preparation for the first operational administration of FCAT 2.0 Science in those grades in 2011.

FDOE implemented a data forensics program beginning with the spring 2011 administration. The purpose of the program is to analyze data to identify highly unusual results. For the first year of implementation, student tests with extremely similar responses and schools with extraordinarily high levels of erasures on paper-based assessments were held by FDOE pending further investigation and appeals by school districts.

**Developments in 2012**

The first operational administration of FCAT 2.0 Science assessments and the Geometry and Biology 1 EOC Assessments occurred during the spring administration window. Standard-setting meetings for these grade/subject/course combinations occurred with educators in September 2012. In addition, grades 6 and 10 FCAT 2.0 Reading were administered online the first time.

A sample of students taking U.S. History or U.S. History Honors was drawn to take stand-alone field-test forms for the U.S. History EOC Assessment.

For FCAT 2.0 Writing, in addition to the elements of focus, organization, support, and conventions described in the rubrics, the scoring decisions included expanded expectations regarding the following: (1) increased attention to the correct use of standard English conventions and (2) increased attention to the quality of details, requiring use of relevant, logical, and plausible support, rather than contrived statistical claims or unsubstantiated generalities.

### Developments in 2013

The first operational administration of the U.S. History EOC Assessment occurred during the spring administration window. Standard-setting meetings for this course occurred with educators in August 2013. In addition, grades 7 and 9 FCAT 2.0 Reading and Grade 5 Mathematics were administered online for the first time.

A sample of students taking Civics was drawn to take stand-alone field-test forms for the Civics EOC Assessment. Also a sample of students representative of the state's demographic was administered the Writing prompt field test.

### 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 normal competitive bid process, a contract was awarded to the American Institutes for Research (AIR) to develop the new Florida Standards Assessments. 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 AIR, the Florida Department of Education, and the Department's Test Development Center met to build 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 AIR reviewed each item and its associated statistics to determine alignment to Florida's academic standards and to judge the suitability of the statistical qualities of each item. Only those 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 onto test forms administered online in order to build the Florida-specific FSA 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 the 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 are scheduled to be used beginning in the spring of 2016.

**Developments in 2015**

The first operational administration of the FSA occurred in spring 2015. Grades 3 and 4 ELA and mathematics 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 per an IEP or 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 FCAT 2.0 Reading, Grade 10 FCAT 2.0, and the 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 FSA Algebra 1), which are also statutory requirements.  This was accomplished using equipercentile linking for Grade 10 ELA and for 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.

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 FSA 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 through 10 ELA and grades 3 through 8 Mathematics was established using statistics from the spring 2015 administration. Standard-setting meetings for grades 3 through 10 ELA, grades 3 through 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 took into account the feedback from more than 400 educators from across the state, as well as members of the community, businesses and district-level education leaders. Additionally, the Commissioner took into account 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 twelve weeks.

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

FDOE manages the Florida statewide assessment program with the assistance of several participants, including multiple offices within FDOE, Florida educators, a Technical Advisory Committee (TAC), and vendors. 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 Assessment.* The Office of Assessment oversees all aspects of Florida's statewide assessment program, including coordination with other FDOE offices, Florida public schools, and vendors.

*Test Development Center.* Funded by FDOE via a grant to the local school district, the Test Development Center (TDC) works with Florida educators and vendors to develop test specifications and test content and to build test forms.

### Florida Educators

Florida educators participate in most aspects of the conceptualization and development of Florida assessments. Educators participate in the development of the academic standards, the clarification of how these standards will be assessed, the test design, and the review of test questions and passages.

### Technical Advisory Committee

FDOE typically convenes a panel biannually to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Florida testing. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple Florida school districts.

### American Institutes for Research

American Institutes for Research (AIR) is the vendor that was selected through the state-mandated competitive procurement process. AIR was responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the Florida assessments described in this report. All activities were conducted under the close direction of FDOE staff experts. Beginning in summer 2014, AIR became the primary party responsible for executing psychometric operations for the Florida statewide assessments.

### Human Resources Research Organization

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 FDOE. HumRRO also provides consultative services to FDOE on psychometric matters.

### Buros Institute of Mental Measurements

Buros Institute of Mental Measurements (Buros) provides professional assistance, expertise, and information to users of commercially published tests. For the 2015 Florida statewide assessments, Buros provided independent operational checks on the equating procedures of the FSA, on-site monitoring of Writing hand-scoring activities, and the scanning and editing services provided by AIR.

### Caveon Test Security

Caveon Test Security analyzes data for the FSA using Caveon Data Forensics$^{TM}$ 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

The FSA was administered primarily as an online, fixed-form assessment, making use of several technology-enhanced item types. Students in grades 3 and 4 were administered paper forms in 2015, and students in grades 5 and higher were provided with access to an accommodated paper form only if such a need was indicated on their IEP or Section 504 Plan.

Administered test forms contained operational items and embedded field test (EFT) items in pre-determined slots across each form. Operational items were items used to calculate student scores. The EFT items were non-scored items and are used either to populate the FSA test bank for future operational use or to establish a new vertical scale. While there is only one operational form in grades 3 through 8 Mathematics and 3 through 10 Reading, there are multiple test forms in order to vary the number of EFT items on each form and build a large test bank.

Students in grades 4 through 10 responded to a single text-based Writing prompt, with grades 4 through 7 Writing administered on paper and grades 8 through 10 Writing administered online. Writing and Reading item responses were combined such that the data could be calibrated concurrently and subsequently to form an overall English Language Arts (ELA) score. Scale scores for the separate components were 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.

End-of-Course (EOC) assessments were administered as online, fixed forms to students enrolled in Algebra 1, Algebra 2, and Geometry. These tests had multiple operational forms and also contained EFT items to build future test forms.

## 1.5 Student Participation

By statute, all Florida public school students are required to participate in the statewide assessments. Students take the FSA Mathematics, Reading, Writing, or EOC tests in the spring. Retake administrations for EOC assessments occur in the summer, fall, and winter, and grade 10 ELA retake administrations only occur 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 2015 FSA by grade and subject area. The participation count by subgroup, including gender, ethnicity, special education, and ELL, is presented in Section 9 of this volume.

Table 2: Number of Students Participating in FSA 2014–2015

| Mathematics | | | ELA | | |
|---|---|---|---|---|---|
| Grade | Number Tested | Number Reported | Grade | Number Tested | Number Reported |
| 3 | 216,703 | 215,473 | 3 | 216,321 | 215,317 |
| 4 | 200,437 | 199,351 | 4 | 201,086 | 197,681 |
| 5 | 199,721 | 199,010 | 5 | 201,002 | 196,812 |
| 6 | 192,833 | 191,091 | 6 | 199,790 | 192,614 |
| 7 | 181,619 | 179,194 | 7 | 200,159 | 192,024 |
| 8 | 126,482 | 123,928 | 8 | 206,206 | 198,412 |
| Algebra 1 | 207,387 | 203,235 | 9 | 213,134 | 201,252 |
| Algebra 2 | 162,241 | 158,254 | 10 | 203,028 | 191,080 |
| Geometry | 199,205 | 195,113 | | | |

## 2. Recent and Forthcoming Changes to the Test

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

In the spring of 2015, online administration of the test was affected by distributed denial of service (DDoS) attacks and other test administration issues potentially impacting students taking the tests online. During the spring of 2015, Florida House Bill 7069 was enacted, requiring, among other things, an independent audit of the entire FSA system  before test scores could be released. Alpine Testing Solutions was selected by a three-member panel selected by the Executive Office of the Governor, the President of the Florida Senate, and the Speaker of the Florida House of Representatives. Alpine was awarded the contract to serve as an independent evaluator of the test. Alpine's full report can be found here: http://www.fldoe.org/core/fileparse.php/5306/urlt/FSA-Final-Report_08312015.pdf.

In addition to the work Alpine performed, AIR also investigated the degree to which test scores and item parameters were potentially affected by the test administration issues. This special study, The Impact of Test Administration on FSA Test Scores, is a stand-alone technical report and is included in Volume 7 of the 2015 technical reports.

The report found that students experiencing test administration issues did not score differently than other students. One notable difference is that students completing all test sessions within a single day did in fact tend to score lower than other students who completed the test on separate days, as was intended. The report confirmed that the assessment was an accurate measure of student mastery of the Florida Standards, and that the results can be used for group-level decisions.

A new vertical scale was created for grades 3 through 10 ELA and grades 3 through 8 Mathematics based on the spring 2015 operational administration. Section 6.4 provides an overview of the work and a summary of the final scale. A complete report is included in Volume 7 of the 2015 technical reports.

For 2016 testing and beyond, additional grades and subjects will be phased in as online assessments, with the intent to administer all FSA tests online by spring of 2018, with the exception of paper-based accommodations for students who require them per an IEP or Section 504 Plan. The 2016 FSA tests will be comprised mostly of items developed specifically for Florida, with all Utah SAGE items to be phased out as soon as possible.

In addition to the Alpine report, two other independent studies have been completed that have findings relevant to the FSA in support of its validity.

An independent study was recently completed by Dr. Gary Phillips comparing all states' performance standards to NAEP performance levels. This report concludes that Florida's achievement level 4 aligns strongly with NAEP Proficient for both ELA and Mathematics.

Additionally, the FSA makes use of technology-enhanced items that are similar to those used by the Smarter Balanced Assessment Consortium. These items are administered to students using the same test delivery system developed by AIR. The Smarter Balanced Assessment Consortium completed a cognitive lab study providing validity evidence in support of the technology-enhanced items. Because the same item types are used in the FSA, many of the study's findings regarding those item types can be applied to the FSA.

# 3. SUMMARY OF OPERATIONAL PROCEDURES

## 3.1 ONLINE ADMINISTRATION PROCEDURES

Table 3 shows the schedule for the 2014–2015 FSA administration by test window.

*Table 3: Test Windows by Subject Area*

| Assessment | Testing Window |
|---|---|
| Grades 4–7 paper Writing | March 2–13, 2015 |
| Grades 8–10 online and paper Writing | March 2–13, 2015 |
| Grades 3–4 paper Reading and Mathematics | March 23–April 10, 2015 |
| Grades 5–10 online Reading, Grades 5–8 online Mathematics | April 13–May 8, 2015 |
| Grades 5–10 paper Reading, Grades 5–8 paper Mathematics | April 13–24, 2015 |
| Algebra 1, Algebra 2, and Geometry online | April 20–May 15, 2015 |
| Algebra 1, Algebra 2, and Geometry paper | April 20–May 1, 2015 |

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 FSA tests were administered in sessions, with each session having a time limit. For the online tests, students could begin a session and complete it at another time. However, 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 was not able to return to a session once he or she left campus.

The key personnel involved with the FSA administration included the district test coordinators (DTCs), school administrators, and the test administrators (TAs) who proctored the test. An online test administrator training course was available to TAs. More detailed information about the roles and responsibilities of various testing staff can be found in Volume 5 of the 2015 FSA Annual Technical Report.

A secure browser developed by AIR was required to access the online FSA tests. The browser provided a secure environment for student testing by disabling the hot keys, copy and screenshot capabilities, and access to desktop functionalities, such as the Internet and e-mail. Other measures that protected the integrity and security of the online test are presented in Volume 5 of the 2015 FSA Technical Report.

## 3.2 ACCOMMODATIONS FOR FSA

Florida assessments are inclusive for all students, which serves as one of the evidences for test validity. To maximize the accessibility of the assessments, various accommodations were provided to students with special needs, as indicated by documentation such as Individualized Educational Plans (IEP) or 504 plans. Such accommodations improve the access to state assessments and help students with special needs demonstrate what they know and are able to 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 number of students who took the paper-and-pencil version of the 2014–2015 FSA varies between 562 and 2,369 across grades and subjects, as shown in Table 4.

*Table 4: Counts of Paper-and-Pencil Assessments by Grades and Subjects*

| Subject | Grade | Spring 2015 |
|---|---|---|
| Mathematics | 5 | 2369 |
| | 6 | 1149 |
| | 7 | 1227 |
| | 8 | 958 |
| EOC | Algebra 1 | 1058 |
| | Algebra 2 | 562 |
| | Geometry | 902 |
| Reading | 5 | 2342 |
| | 6 | 1157 |
| | 7 | 1245 |
| | 8 | 1092 |
| | 9 | 1309 |
| | 10 | 1230 |

The test administrator and the school assessment coordinator were responsible for ensuring that arrangements for accommodations were made prior to 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, masking, text-to-speech, and regular or large print passage booklets were made available. Students had the opportunity to utilize these accommodations only as dictated on their IEP or 504 Plans. Additional accommodations and further explanation of the guidelines can be found in the 2014–2015 Annual Technical Report, Volume 5, Summary of Test Administration Procedures.

# 4. Maintenance of the Item Bank

## 4.1 Overview of Item Development

Complete details of AIR's item development plan are provided in the 2014–2015 Annual Technical Report, Volume 2, Test Development. The test development phase included a variety of activities designed to produce high quality assessments that accurately measure skills and abilities of students 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 2015, field test items were embedded on online forms. Future FSA items were not being field tested on paper, so there were no field test items in grades 3 and 4. All assessments were fixed-form with a predetermined number and location of field test items. The paper accommodated versions of online assessments contained filler items in the field test slots to ensure equal length assessments. These items were not analyzed as part of field test calibrations.

## 4.2 Review of Operational Items

During operational calibration, items were reviewed based on their performance during the spring administration. In some instances, operational items were removed from scoring based on content or statistical anomalies that were not apparent during form building.

Prior to the spring administration, a Calibration Specifications document was created by AIR, FDOE, and HumRRO and reviewed by the TAC. The specifications document outlined all details of item calibration, flagging rules for items, linking between paper and online forms, and scoring. AIR used the specifications to complete classical item analyses and IRT calibrations (see Chapters 5 and 6 of this volume) for each test and posted results to a secure location for review. During the spring calibrations, daily calls were scheduled that included all parties: AIR, FDOE, TDC, HumRRO, and Buros. Items were reviewed, with special attention being paid to items flagged based on the statistical rules described in the Calibration Specifications. These flagging rules are outlined in the chapters below. Psychometricians and content experts worked together reviewing items and their statistics to determine if any items were to be removed from scoring.

## 4.3 Field Testing

The Florida Standards Assessments item pool grows each year by field testing new items. Any item used on an assessment is field tested before it is used as an operational item. There are two primary ways to field test items: through either an independent field test (IFT) or an embedded field test (EFT).

IFTs are useful when there are many items to field test. However, student motivation is often a factor impacting the results, as students tend to have less concern over their responses to an IFT. EFTs are commonly viewed as more useful, since field test items can be placed in an operational test, and students are unaware which items are operational or field test. Hence, this tends to mitigate the motivation effect.

### 4.3.1 Independent Field Test

A Writing Independent Field Test (IFT) was administered to a statistically representative sample of students in grades 4 through 10 in the state from December 2014 to early February 2015. The IFT model was used in order to minimize the testing time needed for the operational ELA assessments. 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 test student population. The Writing IFT sampling plan was created collaboratively between AIR and FDOE and vetted through the TAC; it can be found in Appendix A. In grades 4 through 7, students were administered paper-based tests (PBT), while students in grades 8 through 10 were administered computer-based tests (CBT). Approximately 15 prompts per grade were administered, with each student answering two prompts.

The objectives for the IFT were:

- to obtain item statistics on the newly developed Writing prompts for grades 4 through 10; and
- to review the item statistics and choose Writing prompts that will be used as operational items beginning in the spring 2016 administration.

Writing items were analyzed and used during form building in summer 2015.

### 4.3.2 Embedded Field Test

FSA forms were pre-built with ten field test items embedded into each test form, and each form was assigned to students randomly as described below. Some field test items appeared on multiple forms.

Table 5 shows the number of Mathematics and EOC items by grade and item type that were included on forms for field testing.

Table 6 shows the number of Reading items by grade and item type that were included on forms for field testing.

*Table 5: Mathematics and EOC Field Test Items by Item Type and Grade*

| Item Type | 5 | 6 | 7 | 8 | Algebra 1 | Algebra 2 | Geometry |
|---|---|---|---|---|---|---|---|
| MC4 | 35 | 38 | 48 | 64 | 61 | 35 | 39 |
| MS5 | 27 | 12 | 6 | 15 | 4 | 4 | 6 |
| MS6 | 6 | 9 | 5 | 4 | 3 | 3 | 4 |
| GRID | 23 | 26 | 20 | 34 | 12 | 17 | 14 |
| HT |  |  |  |  |  | 2 | 10 |
| EQ | 123 | 121 | 129 | 92 | 55 | 74 | 59 |
| NL |  | 6 | 3 | 4 |  | 1 | 2 |
| Match | 4 | 4 | 5 | 2 | 2 | 3 |  |
| Table | 6 | 8 | 4 | 8 | 3 | 1 | 1 |

*Table 6: Reading Field Test Items by Item Type and Grade*

| Item type | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| MC | 89 | 110 | 116 | 99 | 107 | 94 |
| MS | 18 | 18 | 35 | 12 | 12 | 9 |
| Editing Task Choice | 52 | 51 | 48 | 50 | 53 | 51 |
| Hot Text | 37 | 36 | 24 | 27 | 26 | 18 |
| GRID | 1 |  |  |  |  |  |
| EBSR | 9 | 14 | 18 | 19 | 18 | 14 |
| NL | 3 |  | 3 | 1 | 1 |  |

With fixed forms, 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 AIR 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 there are a total of *N* students participating in the field test. The probability that any one of the *J* forms can be assigned to one student is *1/J*. So, 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 1500 students per item. Hence, given a test length of $L$ and fixing $S$ at 1500 (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.

Table 7 and Table 8 show the total number of forms administered in spring 2015. In each grade, there was a single core or operational form. The same core form was replicated for each vertical linking or embedded field test form, resulting in multiple forms for each grade and subject. For the EOCs, there were multiple core forms, each also replicated to create a number of embedded field test forms. The EOCs were not used in vertical linking, so no vertical linking forms were created for these.

*Table 7: ELA Form Summary*

| Grade | Total Number of Forms |
|---|---|
| 3 | 9 |
| 4 | 14 |
| 5 | 43 |
| 6 | 42 |
| 7 | 46 |
| 8 | 39 |
| 9 | 43 |
| 10 | 34 |

*Table 8: Mathematics and EOC Form Summary*

| Grade | Total Number of Forms |
|---|---|
| 3 | 6 |
| 4 | 9 |
| 5 | 32 |
| 6 | 32 |
| 7 | 32 |
| 8 | 29 |
| Algebra 1 | 19 |
| Algebra 2 | 15 |
| Geometry | 18 |

A detailed overview of the development and review process for new items is given in the 2014–2015 FSA 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

## 5.1 Classical Item Analyses

Item analyses examine whether test items function as intended. Overall, a minimum sample of 1500 responses (Kolen & Brennan, 2004) per item was required for both classical analysis and for the Item Response Theory (IRT) analysis. However, many more responses than 1500 were always available. For operational item calibrations, an early processing sample was used in the analyses; for field test item calibrations, all students were used. Similarly, a minimum sample of 200 responses (Zwick, 2012) per item in each subgroup was applied for differential item functioning (DIF) analyses.

Several item statistics were used to evaluate multiple-choice (MC) and non-multiple choice 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 9.

*Table 9: Thresholds for Flagging Items in Classical Item Analysis*

| Analysis Type | Flagging Criteria |
|---|---|
| Item Discrimination | Point biserial correlation for the correct response is < 0.25 |
| Distractor Analysis | Point biserial correlation for any distractor response is > 0 |
| Item Difficulty (1pt items) | The proportion of students (p-value) is < 0.20 or > 0.90 |
| Item Difficulty (>1pt 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 examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the ability estimate for students. Point biserial correlations for operational items can be found in Appendix B.

### Distractor Analysis

Distractor analysis for multiple-choice items was used to identify items that may have had marginal distractors or ambiguous correct responses, the wrong key, or more than one correct answer which attracted high-scoring students. For multiple-choice 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 were either extremely difficult or extremely easy were flagged for review but were not necessarily deleted if they were grade-level appropriate and aligned with the test specifications.

For multiple-choice items, the proportion of students in the sample selecting the correct answer (the *p*-value) was computed in addition to the proportion of students selecting incorrect responses. For constructed response items, item difficulty was 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 6.5. P-values for operational items can be found in Appendix B.

## 5.2 DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSIS

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) provides a guideline of 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, FSA items were evaluated in terms of DIF statistics.

DIF analysis was conducted for all items to detect potential item bias across major ethnic and gender 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
- Special with disability (SWD)/Not SWD
- English Language Learner (ELL)/Not ELL

*Differential item functioning* 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 reexamine each flagged item to make a decision about whether the item should have been excluded from the 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 raw score on the operational items on a given test is used as the ability-matching variable. That score is divided into ten intervals to compute the MH$\chi^2$ 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 $\mathrm{k} = \{1, 2, \ldots \mathrm{K}\}$ for the strata, $\mathrm{n_{R1k}}$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 $\mathrm{n_{+1k}}$ is the total number of correct responses, $\mathrm{n_{R+k}}$is the number of students in the reference group, and $\mathrm{n_{++k}}$ is the number of students, in stratum $k$, and the variance is calculated as

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

$\mathrm{n_{F+k}}$ is the number of students in the focal group, $\mathrm{n_{+1k}}$ is the number of students with correct responses, and $\mathrm{n_{+0k}}$ 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 ($\Delta_{MH}$, 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 were 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 10. 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, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., white or 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 10: DIF Classification Rules*

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

DIF summary tables can be found in Appendix B for operational items and in Appendix C for field test items. Across all operational items and DIF comparison groups, less than 1% of Mathematics and EOC items were classified as C DIF and 1.9% of ELA items were classified as

C DIF. Items were reviewed by content specialists and psychometricians to ensure they were free of bias.

Across all field test items and DIF comparison groups, less than 1% of Mathematics and EOC items were classified as C DIF, and 2.9% of ELA items were classified as C DIF. All field test items will be reviewed by content specialists and psychometricians prior to being placed on forms for operational use. More information about test construction and item review can be found in Volume 2.

In addition to the classical item summaries described in this chapter, two IRT-based statistics were used during item review. These methods are described in Section 6.2.

# 6. ITEM CALIBRATION AND SCALING

Item Response Theory (IRT; van der Linden & Hambleton, 1997) was used to calibrate all items and to derive scores for all FSA 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 person 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 the Master's Partial Credit Model. The Rasch family is distinguished in that models do not incorporate a pseudo-guessing parameter and assumes that all items have the same discrimination.

Extensions to the Rasch model include the 2- and 3-parameter logistic 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 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 items, but it can be used with any item where there is a possibility of 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. FDOE and AIR collaborated through several rounds of review to ensure that the strata were appropriately defined and the student population was adequately represented; this Early Processing Sample Plan, which can be found in Appendix D, was also reviewed and affirmed by TAC. For grade 8 Mathematics and EOC calibrations, the entire population was used instead of the EPS.

There are two general approaches used in IRT to calibrate items and score students based on the estimated item difficulties. In pre-equating, item responses are collected from a student group, the statistical characteristics of the items are estimated from that group, and then these statistics are used to score all future groups of students. This approach assumes that the characteristics of the items remain constant over time. A second approach is post-equating. In this approach, item responses are collected from a student group, and the statistical characteristics of the items are estimated from those responses. However, these statistical characteristics are assumed to apply only to this student group. New item statistics are collected each year when items are used, thus assuming the statistical characteristics of the item may be changing as students change.

In Florida, this second approach of post-equating was used, and all data regarding item responses were derived from the most recent group of students to be administered the test. In future years, items will be equated back to the spring 2015 FSA scale, a step that was not necessary in this initial year.

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 spring 2015 operational scale using Stocking-Lord.

## 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 multiple choice items, to use the 2-parameter logistic (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 multiple choice 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, then the 2PL model was used; however, the 3PL model was used if guessing was in fact possible.

The 3-parameter model is typically expressed as

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

where $\mathrm{P_i}(\theta_\mathrm{j})$ is the probability of examinee $j$ answering item $i$ correct, $\mathrm{c_i}$ is the lower asymptote of the item response curve (i.e., the pseudo-guessing parameter), $\mathrm{b_i}$ is the location parameter, $\mathrm{a_i}$ is the slope parameter (i.e., 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 $\theta_\mathrm{j}$. For the 2PL the pseudo-guessing parameter ($\mathrm{c_i}$) is set to 0.

The Generalized Partial Credit Model is typically expressed as the probability for individual $j$ of scoring in the *x^th^* category to the *i^th^* item as:

$$\mathrm{P}(x|\theta_\mathrm{j}) = \frac{\exp\sum_{\mathrm{k}=1}^{x}\mathrm{Da_i}\left(\theta_\mathrm{j} - \delta_\mathrm{ki}\right)}{1 + \sum_{\mathrm{h}=1}^{\mathrm{m_i}}\exp\sum_{\mathrm{k}=1}^{\mathrm{h}}\mathrm{Da_i}\left(\theta_\mathrm{j} - \delta_\mathrm{ki}\right)}$$

where $\delta_{ki}$ is the ith step value, $\mathrm{x} = 0,1,..,\mathrm{m_i}$, $\mathrm{m_i}$ is the maximum possible score of the item.

All item parameter estimates were obtained with IRTPRO version 2.1 (Cai, Thissen, & du Toit, 2011). IRTPRO uses marginal maximum likelihood estimation (MLE).

## 6.2 IRT Item Summaries

### 6.2.1 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. In order to calculate fit statistics prior to scores being available from AIR's scoring engine, MAP estimates from IRTPRO were used for student ability estimates in the calculations. IRTPRO does not calculate the MLE; however, the mean and variance for the MAP were set to 0 and 100, respectively, so that the resulting MAP estimates approximate the MLE.

Q1 is calculated as

$$\mathrm{Q}_{1\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \frac{\mathrm{N}_{\mathrm{ij}}(\mathrm{O}_{\mathrm{ij}} - \mathrm{E}_{\mathrm{ij}})^2}{\mathrm{E}_{\mathrm{ij}}\left(1 - \mathrm{E}_{\mathrm{ij}}\right)}$$

where $\mathrm{N}_{\mathrm{ij}}$ is the number of examinees in cell j for item i, $\mathrm{O}_{\mathrm{ij}}$and $\mathrm{E}_{\mathrm{ij}}$ are the observed and predicted proportions of examinees in cell j for item i. The expected or predicted proportion is calculated as

$$\mathrm{E}_{\mathrm{ij}} = \frac{1}{\mathrm{N}_{\mathrm{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N}_{\mathrm{ij}}} \mathrm{P}_{\mathrm{i}}\left(\hat{\theta}_{\mathrm{a}}\right)$$

where $\mathrm{P}_{\mathrm{i}}\left(\hat{\theta}_{\mathrm{a}}\right)$ is the item characteristic function for item i and examinee a. The summation is taken over examinees in cell j. The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$$gen\ \mathrm{Q}_{1\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{k}=1}^{\mathrm{m}_{\mathrm{i}}} \frac{\mathrm{N}_{\mathrm{ij}}(\mathrm{O}_{\mathrm{ikj}} - \mathrm{E}_{\mathrm{ikj}})^2}{\mathrm{E}_{\mathrm{ikj}}}$$

with

$$\mathrm{E}_{\mathrm{ikj}} = \frac{1}{\mathrm{N}_{\mathrm{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N}_{\mathrm{ij}}} \mathrm{P}_{\mathrm{ik}}\left(\hat{\theta}_{\mathrm{a}}\right).$$

Both the Q1 and Generalized Q1 results are transformed into the statistic ZQ1, and are compared to a criterion, $\mathrm{ZQ}_{\mathrm{crit}}$, to determine acceptable fit.

$$\mathrm{ZQ}_1 = \frac{\mathrm{Q}_1 - \mathrm{df}}{\sqrt{2\mathrm{df}}}$$

and

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

where Q is either Q1 or Generalized Q1 and df is the degrees of freedom for the statistic. The degrees of freedom is calculated as 10 – number of parameters estimated. For example, multiple choice items have df = 7. Poor fit is indicated where ZQ1 is greater than $\mathrm{ZQ}_{\mathrm{crit}}$.

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

Overall, few operational items were flagged by Q1. Algebra 1 had the most items flagged, with a total of eight flags; however, these flagged items appeared across the four different core forms. Items flagged by Q1 were reviewed by psychometricians and content specialists before a final decision was made about their inclusion for student score calculation.

Comparatively, a higher percentage of field test items were 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.

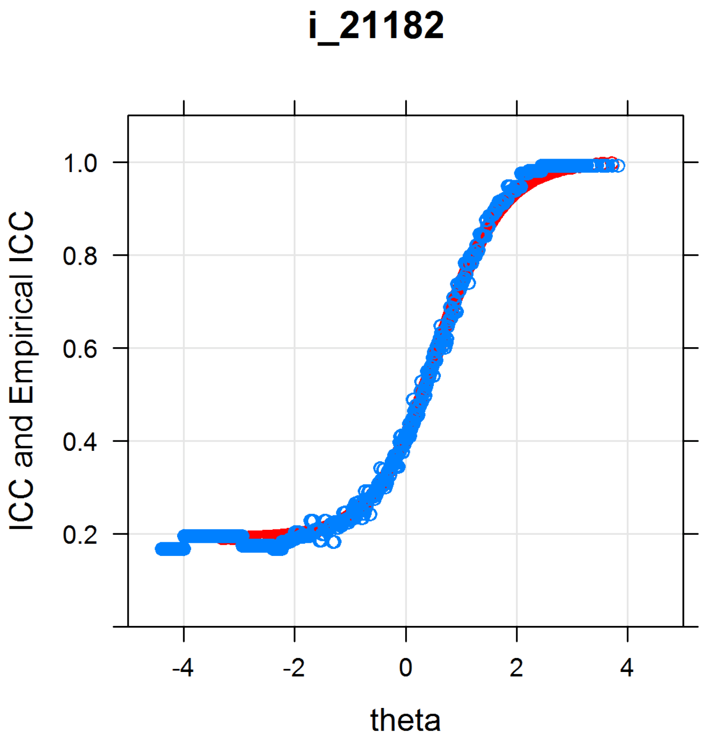
### 6.2.2 Item Fit Plots

Another way to evaluate item fit is to examine empirical fit plots for each item. The plots below 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.

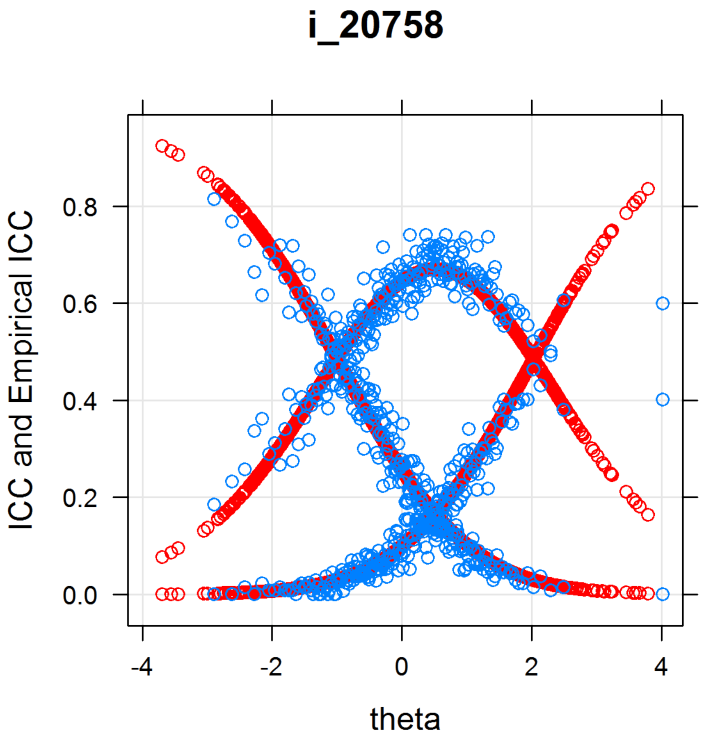
The fit plot in Figure 1 illustrates a one-point item that fits the item response model well. The dots represent the proportion of students within a score bin correctly answering the item. The solid line is the IRT-based item characteristic curve. A "good" item is one in which the dots are essentially superimposed over the line across the range of ability. In fact, the solid line is almost not visible underneath the dots for the first plot.

*Figure 1: Example Fit Plot – Good Fitting 1-pt Item*



The plot in Figure 2 is provided for items worth two or more points. 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. Similar to the first plot, a "good" item is one in which the dots follow the solid lines across the range of ability.

*Figure 2: Example Fit Plot – Good Fitting 2-pt Item*



## 6.3 Equipercentile Linking

Per state statute, there was a requirement to link student performance on 2015 FSA assessments in Grade 3 ELA, grade 10 ELA, and EOC Algebra 1 to 2014 performance on FCAT 2.0 Grades 3 Reading, Grade 10 Reading, and the NGSSS Algebra 1 EOC, respectively, in order to make student-level promotion and graduation decisions prior to standard setting. This was accomplished via equipercentile linking, as was done previously in Florida when transitioning to new standards and assessments. As discussed previously, there was further legislation in 2015 that required that students scoring in the bottom quintile of the FSA Grade 3 ELA test be identified and reported to districts in order for the scores to be considered in the decision to promote students to Grade 4.

In grade 3 ELA, each student's status was determined based on the percentile distribution of the T score. Students who were below the 20th percentile (i.e., bottom quintile) of the T score were identified for further district consideration for promotion ("not passing"), whereas those at or above the 20th percentile of the T score were categorized as "pass." Although the pass/not pass labels were used, it was left to the school districts to use this and any other information districts deemed appropriate in order to determine whether students were promoted to Grade 4. More information about T scores can be found in Section 8.1.2. The following steps were followed in order to determine the percentile distribution and create the *Pass/NotPass* flag.

1. Rescore the entire population, including calculation of T score.
2. Subset to retain only valid records after applying all necessary exclusion rules.

3. Sort the data based on the T score and using empirical cumulative distribution find the T score cut score that corresponds to the 20th percentile.
4. Assign a PASS score of *NotPass* if the student falls below the cut score; otherwise assign *Pass*. Using the PASS score variable, confirm that no more than 20% of the students were given a score of *NotPass*.

In grade 10 ELA and EOC Algebra 1, RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2005) was used to conduct randomly equivalent groups equipercentile linking. The objective of this equipercentile linking was to find the ability estimate for the $i^{th}$ student at percentile rank *p* on the FSA that corresponded to an FCAT 2.0 or EOC score at the same percentile rank given the observed distribution of scores from 2014. Once student MLE ability estimates based on spring 2015 item calibration were calculated, the theta to scale score transformation equations in Table 11 were used to get an FSA score on an FCAT 2.0 or EOC equivalent scale. Frequencies were computed at each score point for both the 2015 FSA scale and the 2014 FCAT 2.0 and EOC scales.

*Table 11: Transformation Equations for Grade 10 ELA and Algebra 1 Equipercentile Linking*

| Subject and Grade | Transformation Equation |
|---|---|
| Grade 10 ELA | $\hat{\theta}_i^* = \text{round}(\hat{\theta}_i \times 18.822290 + 244.870126)$ |
| Algebra 1 | $\hat{\theta}_i^* = \text{round}(\hat{\theta}_i \times 25 + 400)$ |

RAGE-RGEQUATE was then used to apply the equipercentile linking function defined as

$$e_{FCAT2.0}(FSA) = G^{-1}[F(FSA)]$$

where $F$ is the cumulative distribution function of the FSA and $G^{-1}$ is the inverse cumulative distribution function of the FCAT 2.0.

Interpolation at the extremes was then conducted. Kolen (1984) recommended using 0.5 percentile as the cut-off value, which includes score points between 0.5 and 99.5 percentile ranks. The steps for the cut-off value were:

1. Obtain the percent of students scoring at the two extreme scores in 2014 and 2015
2. Add 0.5 to the obtained percent values
3. Choose the larger percent values between the two years at LOSS and HOSS

Based on the results, a complete scale-score-to-scale-score concordance table was created, showing an FCAT 2.0 and EOC equivalent scale score for each FSA scale score. Performance levels were assigned using the previous FCAT 2.0 and EOC cut scores shown in Table 12.

Table 13 and Table 14 show the proportion of students scoring in each performance level on the original and linked scores for Algebra 1 and Grade 10 ELA respectively.

*Table 12: Cut Scores for Grade 10 ELA and Algebra 1 Equipercentile Linking*

| Subject | Grade | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|---|
| ELA | 10 | 188–227 | 228–244 | 245–255 | 256–270 | 271–302 |
| EOC | ALG I | 325–374 | 375–398 | 399–424 | 425–436 | 437–475 |

*Table 13: Proportion of Students by Performance Level and Score Type for Algebra 1*

| | Algebra1 EOC Equivalent | Algebra1 EOC |
|---|---|---|
| Level 1 | 0.113514 | 0.112636 |
| Level 2 | 0.218023 | 0.222985 |
| Level 3 | 0.389956 | 0.382485 |
| Level 4 | 0.134932 | 0.138777 |
| Level 5 | 0.143575 | 0.143117 |
| Proficient* | 0.668463 | 0.664379 |

* Students at or above level 3

*Table 14: Proportion of Students by Performance Level and Score Type for Grade 10 ELA*

| | Grade 10 ELA FCAT 2.0 Equivalent | Grade 10 ELA FCAT 2.0 |
|---|---|---|
| Level 1 | 0.176627 | 0.174273 |
| Level 2 | 0.279092 | 0.275399 |
| Level 3 | 0.214942 | 0.222921 |
| Level 4 | 0.220818 | 0.217663 |
| Level 5 | 0.108521 | 0.109744 |
| Proficient* | 0.544281 | 0.550328 |

* Students at or higher than level 3

Figure 3 and Figure 4 below show the CDFs for the FCAT 2.0 or EOC score and the equivalent scores for Algebra 1 and Grade 10 ELA.

*Figure 3: CDFs by Score Type for Algebra 1*

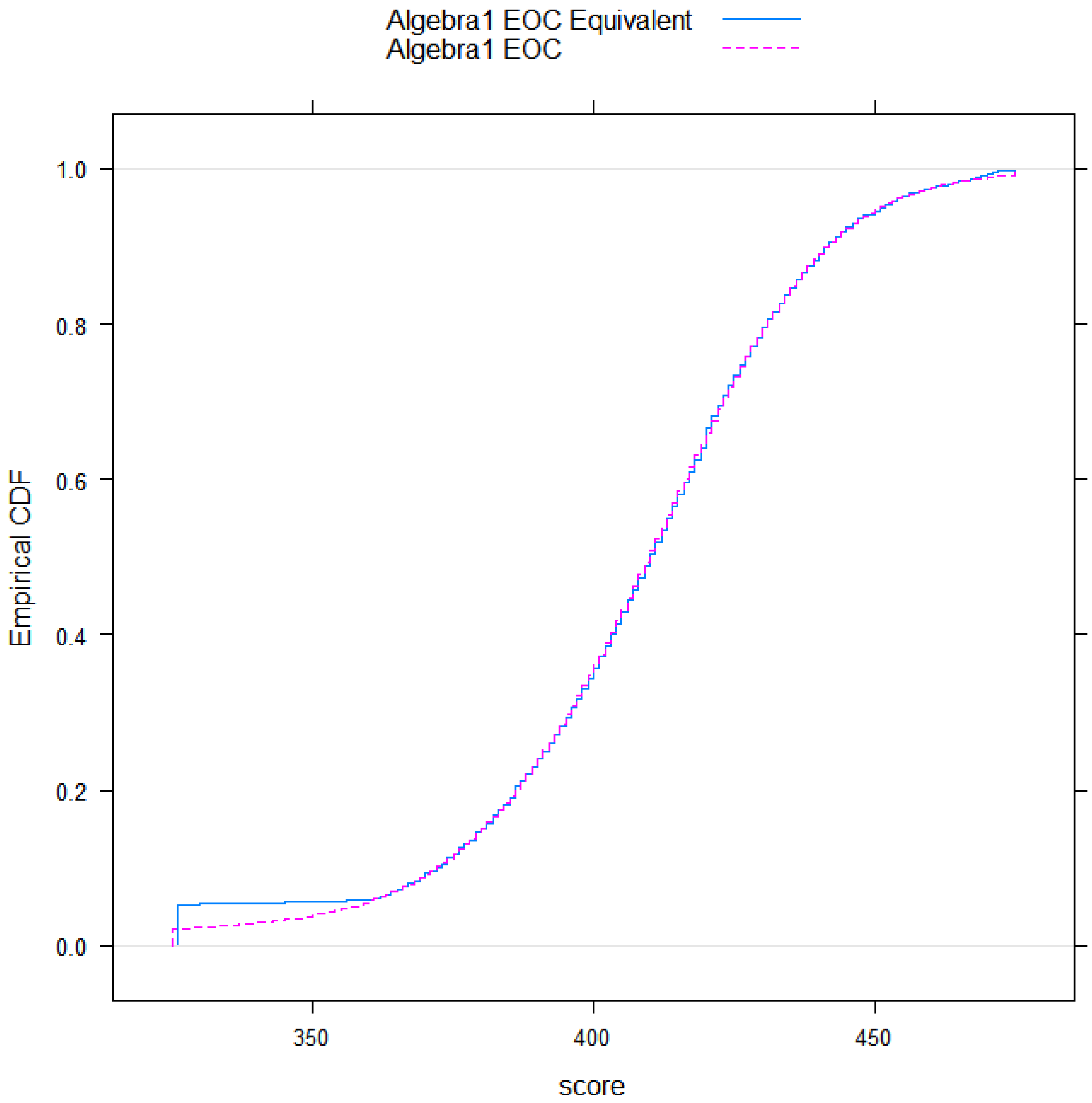
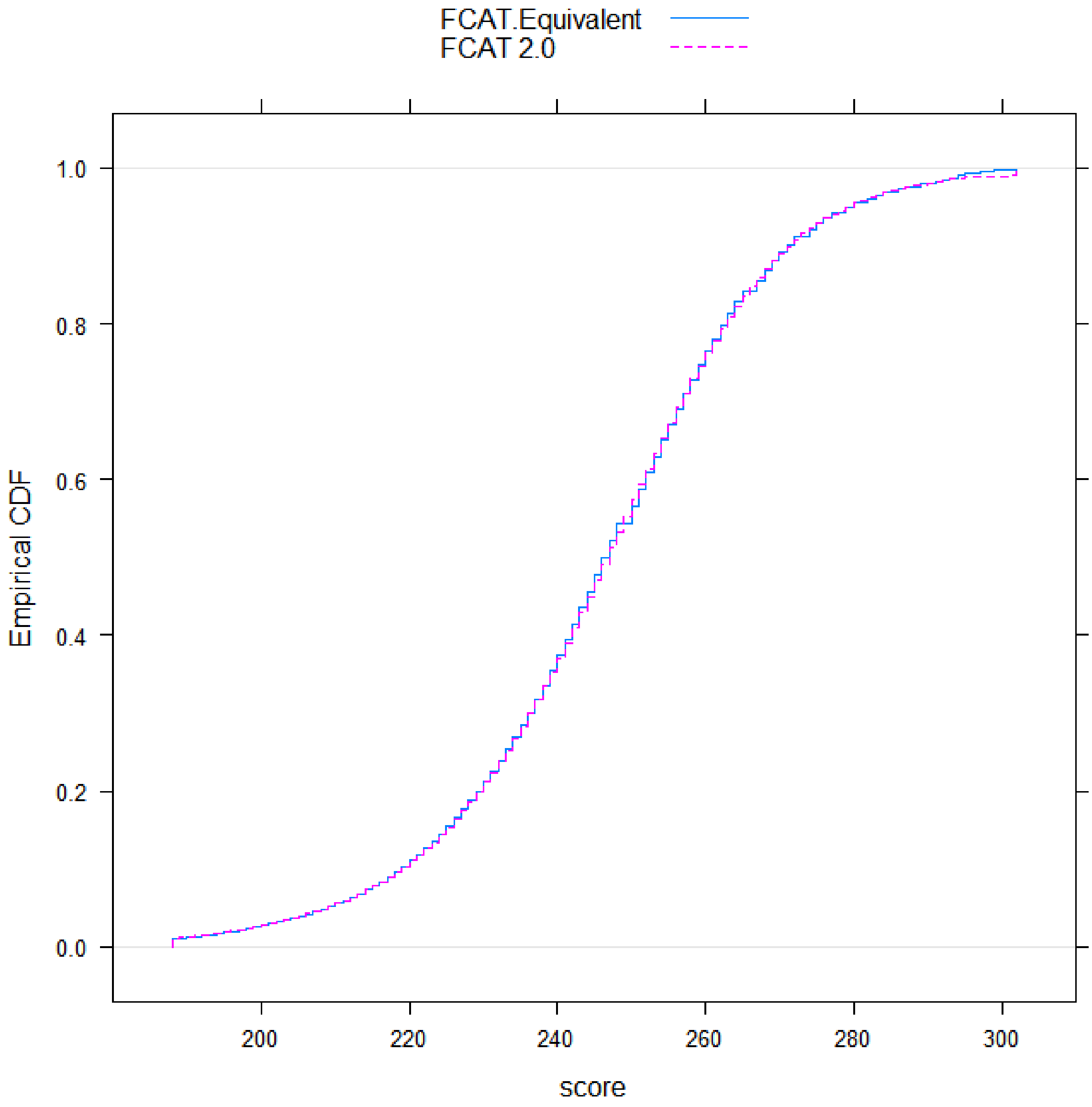


*Figure 4: CDFs by Score Type for Grade 10 ELA*



## 6.4 Vertical Scaling for ELA and Mathematics

### 6.4.1 Methodology

A new vertical scale for grades 3 through 10 ELA and grades 3 through 8 Mathematics was created using a common-item, non-equivalent group design procedure (Kolen & Brennan, 2004) based on the first operational administration of the FSA in spring 2015.

For both ELA and Mathematics, grade 3 served as the base grade.

Two types of vertical linking items, *forward*- and *backward*-linking items, were placed on operational forms for data collection. The forward-linking items are items measuring content in grade *g* and placed on the test forms in grade *g+1*. Backward-linking items are those measuring content in grade *g* and placed on the test forms in grade *g–1*. Forward linking occurs when only forward linking items are used, and backward linking occurs when only backward linking items are used. Mixed linking occurs when both forward and backward linking methods are combined to create a vertical scale.

### Calibration of Vertical Linking Items

To complete the vertical linking, four IRT calibrations were performed per grade: (1) the operational or core items only, (2) operational items with backward vertical linking items, (3) operational items with forward vertical linking items, and (4) operational items with on-grade vertical linking items. Note that the grades at the end of the vertical scale required one less calibration. For example, in grade 5 the on-grade calibration included the vertical linking items that came from grade 5, the forward calibration included the vertical linking items that came from grade 4, and the backward calibration included the vertical linking items that came from grade 6. In each calibration, the operational or core items and the vertical linking items were freely calibrated.

It was necessary to link the item parameters from the operational plus vertical linking item calibrations to the operational item parameters from the original operational only calibration. Stocking-Lord equating was implemented, using the operational items as the common items. The resulting item parameters for the vertical linking items were used to build the vertical scale.

### Variants

Multiple variants of the vertical scale were implemented to explore the effects of various methods and provide options for FDOE. Versions A–G are outlined below.

1) **Version A:** Backward linking: Use items that are on-grade in grade *g* and vertical linking in grade *g–1*.

    a. The following is an example of how this approach was implemented. Items measuring on-grade content in grade 4 were placed on to the grade 3 test forms. These served as the backward linked items. The two sets of parameters from the calibrations were used to find the linking constants between the tests.

2) **Version B:** Forward linking:  Use items that are on-grade in grade *g* and vertical linking in grade *g+1*.

    a. The following is an example of how this approach was implemented. Items measuring on-grade content in grade 3 were placed on to the grade 4 test forms. These served as the forward linked items. The two sets of parameters from the calibrations were used to find the linking constants between the tests.

3) **Version C:** Mixed linking: Use items that are on-grade in grade *g* and vertical linking in grade *g–1* and items that are on-grade in grade *g–1* and vertical linking in grade *g*.

    a. As an example, grade 4 contains on-grade items placed onto the grade 3 forms and also includes items from grade 3 placed onto the grade 4 forms. Both sets of these items were used to find the linking constants between grades 3 and 4.

4) **Version D:** Mixed Linking: Use the calibrated item parameters from Version C, but drop items with poor model fit. The Q1 statistic was used as the measure of model fit.
5) **Version E:** Mixed Linking: Use the calibrated item parameters from Version C, but remove items if p-values are reversed across grades.
6) **Version F:** Mixed Linking: Use the calibrated item parameters from Version C, but remove items if poor model fit or if p-values are reversed.
7) **Version G:** Mixed Linking: Use the calibrated item parameters from Version C, but remove items with poor classical and/or $D^2$statistics (see flagging criteria below).

### Flagging Criteria

After performing the Stocking-Lord in Option C above, the equated parameters were compared by rescaling items to be on the same scale. $D^2$, the sum of the squared differences between ICCs, was calculated as

$$D^2 = \sum_j w_j \left(P_{ai}(\theta_j) - P_{bi}(\theta_j)\right)^2$$

where $P_{ai}(\theta_j)$ is the probability of a correct response on item $i$ in grade $a$ given an ability of $\theta_j$ and $P_{bi}(\theta_j)$ is the probability of a correct response on item $i$ in grade $b$ given an ability of $\theta_j$, and $w_j$ is the quadrature weight at node $j$. Grades $a$ and $b$ are the adjacent grades being compared, with $a$ being the higher grade. For example, if comparing the ICCs between a given item in grades 3 and 4, grade $a$ is grade 4 and grade $b$ is grade 3. $D^2$ was calculated and ICCs were plotted. Items with $D^2$ values more than 3 standard deviations were reviewed for possible removal. Table 15 outlines the flagging criteria used in dropping items.

*Table 15: Flagging Criteria for Vertical Linking*

| Rule | Flagging Criteria | Rationale |
|---|---|---|
| p-values | For multiple choice items, flag if p < .25 or p > .95 | Items are too difficult and p-value is less than expected from random chance or item is too easy for population |
| Relative mean | For polytomous items, flag if relative mean is < .15 or > .95 | Item difficulty is too difficult or too easy |
| Biserial/polyserial | Flag if < .15 | Non-discriminating item |
| Distractor p-value | Flag if p-value for distractor is larger than p-value for key | Potentially problematic item |
| Distractor biserial | Flag if biserial for any distractor is larger than biserial for key | Distractor is more discriminating than the keyed response |
| Item Position shift | Flag if item shifts more than 5 positions | Item position can affect item performance |
| $D^2$ and ICCs | Flag if $D^2$ greater than 3 standard deviations | Too much difference between grades |
| Convergence Issues | Flag IRT statistics if IRTPRO does not converge | The number of iterations and convergence should be noted in a table. |

The flagging rules outlined in Table 15 did not require that items be removed, but they indicated items that required further review before a final decision was made.

**Implementation**

The following list outlines the step-by-step procedures used when constructing the linkages.

1) Prepare data file for grade *g*. Data will include only the operational items and the vertical linking items.
2) Conduct a separate calibration of the core/operational items administered to each grade using IRTPRO.
3) Perform a second calibration including the core/operational and vertical linking items.
4) Link operational plus vertical linking item calibration to the operational item only calibration using the Stocking-Lord procedure to put vertical linking items on the scale of a given grade level
5) For each version, implement chain linking via the Stocking-Lord procedure, removing flagged items if necessary, according to the following plan:

   i. Link grade 3 to grade 4 to find linking constants $\mathbf{A_{34}}$ and $\mathbf{B_{34}}$

   ii. Link grade 4 to grade 5 to find linking constants $\mathbf{A_{45}}$ and $\mathbf{B_{45}}$

   iii. Link grade 5 to grade 6 and identify linking constants $\mathbf{A_{56}}$ and $\mathbf{B_{56}}$

   iv. Link grade 6 to grade 7 and identify linking constants $\mathbf{A_{67}}$ and $\mathbf{B_{67}}$

   v. Link grade 7 to grade 8 to find linking constants $\mathbf{A_{78}}$ and $\mathbf{B_{78}}$

   vi. Link grade 8 to grade 9 to find linking constants $\mathbf{A_{89}}$ and $\mathbf{B_{89}}$

   vii. Link grade 9 to grade 10 to find linking constants $\mathbf{A_{9,10}}$ and $\mathbf{B_{9,10}}$

6) Update linking constants via the following transformations:

   i. $\boldsymbol{A'_{34}} = \boldsymbol{A_{34}}$ and $\boldsymbol{B'_{34}} = \boldsymbol{B_{34}}$

   ii. $\boldsymbol{A'_{45}} = \boldsymbol{A'_{34}}\, \boldsymbol{A_{45}}$ and $\boldsymbol{B'_{45}} = \boldsymbol{B'_{34}} + \boldsymbol{A'_{34}}\boldsymbol{B_{45}}$

   iii. $\boldsymbol{A'_{56}} = \boldsymbol{A'_{45}}\, \boldsymbol{A_{56}}$ and $\boldsymbol{B'_{56}} = \boldsymbol{B'_{45}} + \boldsymbol{A'_{45}}\boldsymbol{B_{56}}$

   iv. $\boldsymbol{A'_{67}} = \boldsymbol{A'_{56}}\, \boldsymbol{A_{67}}$ and $\boldsymbol{B'_{67}} = \boldsymbol{B'_{56}} + \boldsymbol{A'_{56}}\boldsymbol{B_{67}}$

   v. $\boldsymbol{A'_{78}} = \boldsymbol{A'_{67}}\, \boldsymbol{A_{78}}$ and $\boldsymbol{B'_{78}} = \boldsymbol{B'_{67}} + \boldsymbol{A'_{67}}\boldsymbol{B_{78}}$

   vi. $\boldsymbol{A'_{89}} = \boldsymbol{A'_{78}}\, \boldsymbol{A_{89}}$ and $\boldsymbol{B'_{89}} = \boldsymbol{B'_{78}} + \boldsymbol{A'_{78}}\boldsymbol{B_{89}}$

   vii. $\boldsymbol{A'_{9,10}} = \boldsymbol{A'_{89}}\, \boldsymbol{A_{9,10}}$ and $\boldsymbol{B'_{9,10}} = \boldsymbol{B'_{89}} + \boldsymbol{A'_{89}}\boldsymbol{B_{9,10}}$

## Results

Steps 1–6 above were completed for each method. The results based on these steps were presented to the FDOE's technical advisory committee (TAC). In addition, TDC reviewed the vertical linking items, their statistics, and content coverage. Using this input from TDC and the TAC members, FDOE requested that AIR compare the results of the vertical linking study from versions C through G described above in addition to a new version H. This version was similar to version G and differed only in the number of items dropped from the linking set. This version is labeled as *Final* throughout this section, as this was selected by FDOE to produce the FSA vertical scales.

Table 16 and Table 17 show the number of items by version for Mathematics and ELA.

*Table 16: Number of Items by Version – Mathematics*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| G34 | 60 | 58 | 57 | 55 | 48 | 54 |
| G45 | 60 | 56 | 50 | 46 | 45 | 52 |
| G56 | 60 | 55 | 36 | 34 | 39 | 53 |
| G67 | 60 | 51 | 48 | 40 | 41 | 50 |
| G78 | 60 | 54 | 29 | 27 | 48 | 38 |

*Table 17: Number of Items by Version – ELA*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| G34 | 57 | 52 | 57 | 52 | 53 | 52 |
| G45 | 57 | 57 | 52 | 52 | 53 | 52 |
| G56 | 54 | 51 | 45 | 42 | 49 | 45 |
| G67 | 51 | 40 | 46 | 38 | 46 | 40 |
| G78 | 53 | 42 | 47 | 36 | 48 | 38 |
| G89 | 54 | 47 | 47 | 40 | 49 | 43 |
| G910 | 52 | 48 | 49 | 46 | 47 | 48 |

Raw growth and effect sizes were calculated based on the slopes and intercepts from each version and are shown in Table 18 through Table 21.

*Table 18: Raw Growth by Version – Mathematics*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| 3to4 | 0.82 | 0.80 | 0.89 | 0.88 | 0.73 | 0.68 |
| 4to5 | 0.58 | 0.57 | 0.71 | 0.73 | 0.40 | 0.41 |
| 5to6 | 0.22 | 0.20 | 0.43 | 0.43 | 0.25 | 0.17 |
| 6to7 | 0.32 | 0.30 | 0.43 | 0.42 | 0.20 | 0.24 |
| 7to8 | 0.12 | 0.12 | 0.64 | 0.61 | 0.09 | 0.14 |

*Table 19: Effect Size by Version – Mathematics*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| 3to4 | 0.82 | 0.80 | 0.89 | 0.88 | 0.73 | 0.68 |
| 4to5 | 0.52 | 0.51 | 0.63 | 0.64 | 0.37 | 0.39 |
| 5to6 | 0.17 | 0.16 | 0.33 | 0.32 | 0.22 | 0.16 |
| 6to7 | 0.25 | 0.24 | 0.33 | 0.33 | 0.17 | 0.22 |
| 7to8 | 0.09 | 0.10 | 0.52 | 0.50 | 0.08 | 0.14 |

*Table 20: Raw Growth by Version – ELA*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| 3to4 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 |
| 4to5 | 0.49 | 0.49 | 0.53 | 0.53 | 0.44 | 0.48 |
| 5to6 | 0.23 | 0.23 | 0.26 | 0.27 | 0.20 | 0.21 |
| 6to7 | 0.35 | 0.36 | 0.37 | 0.38 | 0.35 | 0.35 |
| 7to8 | 0.33 | 0.33 | 0.33 | 0.33 | 0.32 | 0.32 |
| 8to9 | 0.19 | 0.17 | 0.19 | 0.19 | 0.17 | 0.17 |
| 9to10 | 0.34 | 0.33 | 0.34 | 0.34 | 0.33 | 0.33 |

*Table 21: Effect Size by Version – ELA*

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| 3to4 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 |
| 4to5 | 0.48 | 0.48 | 0.52 | 0.52 | 0.44 | 0.47 |
| 5to6 | 0.22 | 0.22 | 0.25 | 0.25 | 0.19 | 0.19 |

| Grade | Mixed (Ver. C) | Mixed (Q1; Ver. D) | Mixed (p-value; Ver. E) | Mixed (Q1, p-value; Ver. F) | Mixed (Ver. G) | Final |
|---|---|---|---|---|---|---|
| 6to7 | 0.31 | 0.32 | 0.33 | 0.34 | 0.31 | 0.32 |
| 7to8 | 0.30 | 0.30 | 0.30 | 0.31 | 0.29 | 0.29 |
| 8to9 | 0.17 | 0.16 | 0.18 | 0.17 | 0.15 | 0.15 |
| 9to10 | 0.30 | 0.30 | 0.32 | 0.31 | 0.30 | 0.30 |

Figure 5 and Figure 6 are graphical representations of the growth over grades by version for Mathematics and ELA.

*Figure 5: Growth over Grades by Version – Mathematics*

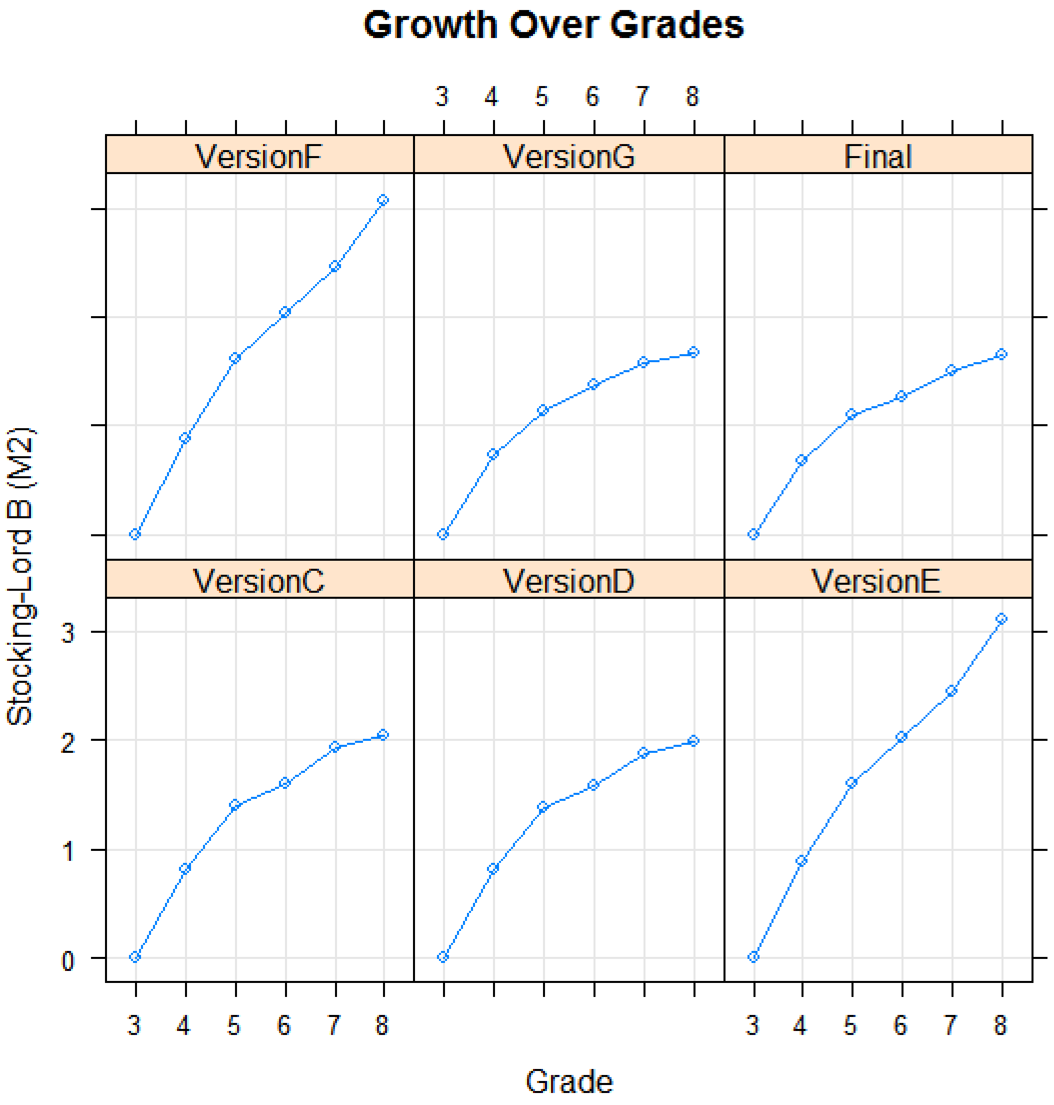
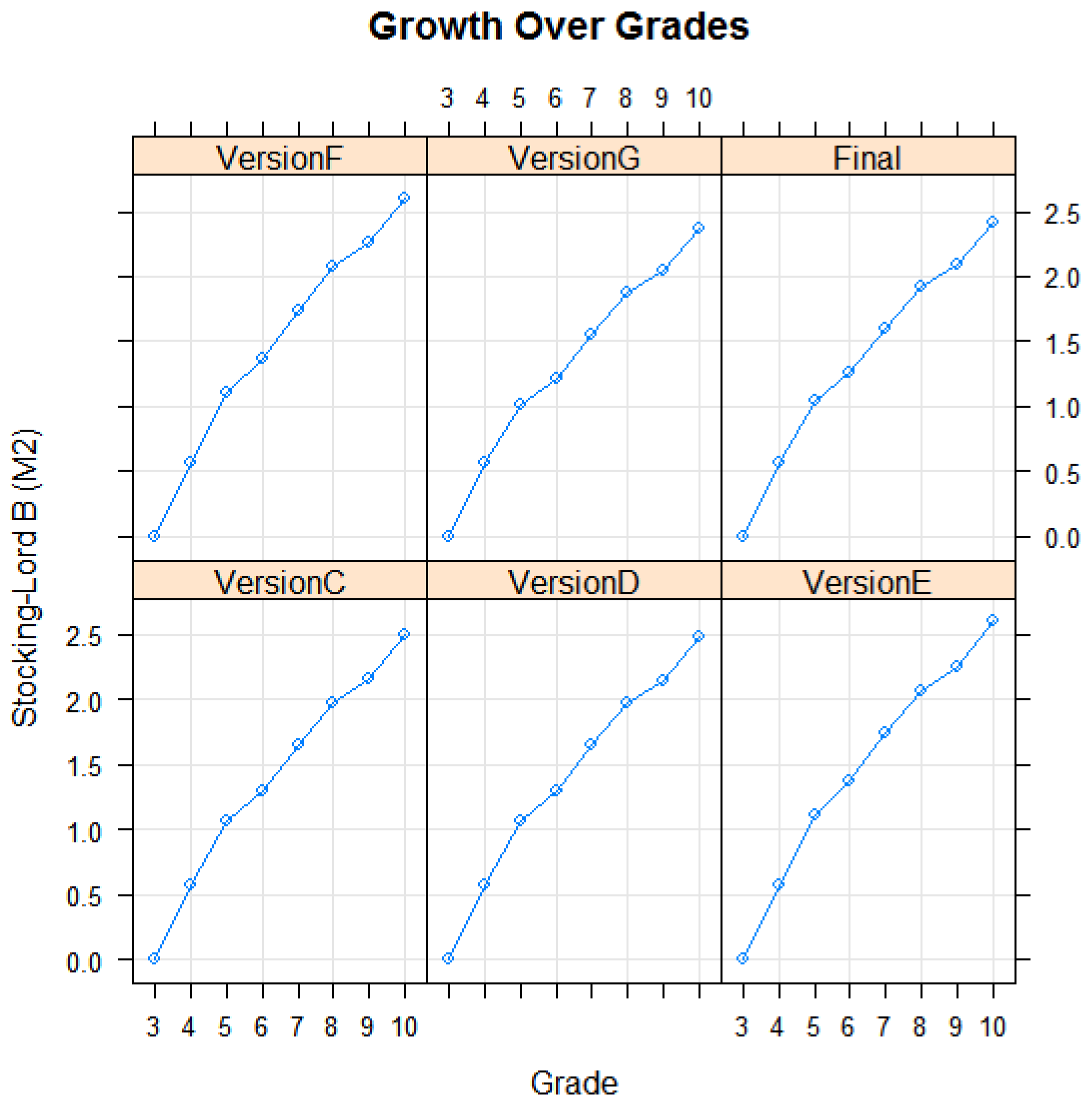


*Figure 6: Growth Over Grades by Version – ELA*



There is a noticeable drop in growth observed in Mathematics in the final grade 8 version; this is related to a policy requiring students in grade 8 to take the test for the course in which they were enrolled. Approximately 80,000 students took the Algebra 1 test while in grade 8, and therefore these students did not take the grade 8 Mathematics test. In all other grades and subjects, the tests included in this vertical scale were estimated on the basis of the entire tested population. Grade 8 Mathematics is the only grade based on a subset.

**Final Scale**

Once versions C–G and the additional version H were complete, AIR presented the results to FDOE. Based on the feedback from TAC and TDC's review of the vertical linking sets, FDOE selected a final version that created a smooth transition from one grade level to the next with an increasing intercept, and the vertical scale was established. Final vertical scaling constants are given in Table 22 and Table 23.

*Table 22: Vertical Scaling Constants for FSA Mathematics*

| Grade | Slope (a) | Intercept (b) |
|---|---|---|
| 3 | 1.000000 | 0.000000 |
| 4 | 1.044966 | 0.680890 |
| 5 | 1.102538 | 1.090128 |
| 6 | 1.084225 | 1.264961 |
| 7 | 1.018981 | 1.507877 |
| 8 | 0.997639 | 1.647321 |

*Table 23: Vertical Scaling Constants for FSA ELA*

| Grade | Slope (a) | Intercept (b) |
|---|---|---|
| 3 | 1.000000 | 0.000000 |
| 4 | 1.011871 | 0.570848 |
| 5 | 1.061502 | 1.048071 |
| 6 | 1.093056 | 1.253075 |
| 7 | 1.079095 | 1.606216 |
| 8 | 1.076568 | 1.921636 |
| 9 | 1.087592 | 2.087487 |
| 10 | 1.064215 | 2.416427 |

## 6.4.2 Calculation of Scores

On-grade MLE estimates, described in Section 8.1.1, are converted to a vertically scaled theta as follows:

$$\theta_{VS} = a * \theta_G + b$$

where $\theta_{VS}$ is the vertical scale theta value, $\theta_G$ is the on-grade MLE estimate of theta, and a and b are the vertical scaling constants given in Table 22 and Table 23.

For a given grade and subject in ELA and Mathematics, the on-grade theta to on-grade scale score transformation equation is

$$SS_G = A * \theta_G + B$$

where $A = 20$ and $B = 300$ for all grades. Replacing the on-grade theta with the vertically scaled theta will yield the vertical scale score. The vertical theta can be replaced using the equation above to find the final on-grade theta to vertical scale score transformation equation.

$$SS_{VS} = A * \theta_{VS} + B$$

$$SS_{VS} = A * (a * \theta_G + b) + B$$

$$SS_{VS} = A * a * \theta_G + (A * b + B)$$

$$SS_{VS} = 20 * a * \theta_G + (20 * b + 300)$$

$$SS_{VS} = a' * \theta_G + b'$$

Applying the vertical scaling constants, the final intercept and slope are provided in Table 24 and Table 25.

*Table 24: Intercept and Slope Values for FSA Mathematics*

| Grade | Slope $(a')$ | Intercept $(b')$ |
|---|---|---|
| 3 | 20.000000 | 300.000000 |
| 4 | 20.899320 | 313.617800 |
| 5 | 22.050760 | 321.802560 |
| 6 | 21.684500 | 325.299220 |
| 7 | 20.379620 | 330.157540 |
| 8 | 19.952780 | 332.946420 |

*Table 25: Intercept and Slope Values for FSA ELA*

| Grade | Slope $(a')$ | Intercept$(b')$ |
|---|---|---|
| 3 | 20.000000 | 300.000000 |
| 4 | 20.237420 | 311.416960 |
| 5 | 21.230040 | 320.961420 |
| 6 | 21.861120 | 325.061500 |
| 7 | 21.581900 | 332.124320 |
| 8 | 21.531360 | 338.432720 |
| 9 | 21.751840 | 341.749740 |
| 10 | 21.284300 | 348.328540 |

## 6.5 RESULTS OF CALIBRATIONS

This section presents a summary of the results from the classical item analysis and IRT analysis described in Chapter 5 for the 2015 spring operational and field test items. The summaries here are aggregates; item-specific details are found in the appendices.

Table 26, Table 27, and Table 28 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 after excluding the dropped items. As noted in Section 1.4 above, there were multiple operational forms for EOC assessments. The summaries in table Table 27 combine operational items across all forms. The field test item summaries can be found in Appendix C; note that grades 3 and 4 Mathematics and Reading did not have any field test items.

*Table 26: 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 | 53 | 0.23 | 0.36 | 0.59 | 0.74 | 0.84 | 0.91 | 0.94 |
| 4 | 54 | 0.26 | 0.34 | 0.57 | 0.68 | 0.76 | 0.89 | 0.94 |
| 5 | 53 | 0.21 | 0.42 | 0.53 | 0.63 | 0.75 | 0.89 | 0.92 |
| 6 | 54 | 0.08 | 0.20 | 0.40 | 0.56 | 0.70 | 0.80 | 0.86 |
| 7 | 55 | 0.05 | 0.17 | 0.31 | 0.46 | 0.65 | 0.79 | 0.87 |
| 8 | 51 | 0.11 | 0.14 | 0.20 | 0.39 | 0.55 | 0.82 | 0.86 |

*Table 27: 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 | 122 | 0.02 | 0.05 | 0.21 | 0.38 | 0.53 | 0.68 | 0.88 |
| Algebra 2 | 64 | 0.01 | 0.03 | 0.12 | 0.22 | 0.45 | 0.63 | 0.68 |
| Geometry | 58 | 0.07 | 0.09 | 0.20 | 0.42 | 0.62 | 0.79 | 0.84 |

*Note that operational items across all forms were combined.

*Table 28: 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 | 47 | 0.18 | 0.30 | 0.41 | 0.52 | 0.67 | 0.90 | 0.94 |
| 4 | 51 | 0.32 | 0.34 | 0.52 | 0.65 | 0.80 | 0.93 | 0.95 |
| 5 | 53 | 0.21 | 0.31 | 0.44 | 0.62 | 0.75 | 0.89 | 0.96 |
| 6 | 55 | 0.26 | 0.32 | 0.44 | 0.56 | 0.69 | 0.88 | 0.93 |
| 7 | 51 | 0.30 | 0.39 | 0.50 | 0.62 | 0.75 | 0.83 | 0.89 |
| 8 | 52 | 0.28 | 0.37 | 0.49 | 0.60 | 0.70 | 0.89 | 0.98 |
| 9 | 54 | 0.16 | 0.29 | 0.48 | 0.62 | 0.74 | 0.89 | 0.95 |
| 10 | 55 | 0.35 | 0.41 | 0.50 | 0.65 | 0.76 | 0.84 | 0.91 |

Table 29 through Table 37 give the 3PL, 2PL, and GPCM item parameter summaries for Mathematics, EOC, and Reading by IRT model and item role. The step parameters for a given GPCM item were averaged to find an overall item difficulty, and this overall value was summarized across items. If less than 10 items existed in a model type for a given test, only the minimum and maximum are displayed below.

### Table 29: 3PL Operational Item Parameter Five-Point Summary and Range, Mathematics

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | a | 0.54 | 0.58 | 0.80 | 0.93 | 1.16 | 1.38 | 1.54 |
| | b | -2.90 | -2.28 | -1.37 | -0.66 | -0.21 | 0.78 | 1.35 |
| | c | 0.02 | 0.03 | 0.08 | 0.14 | 0.22 | 0.32 | 0.42 |
| 4 | a | 0.44 | 0.61 | 0.86 | 1.03 | 1.31 | 1.77 | 2.17 |
| | b | -2.19 | -1.69 | -0.71 | -0.32 | 0.09 | 0.75 | 1.08 |
| | c | 0.03 | 0.06 | 0.12 | 0.16 | 0.24 | 0.48 | 0.64 |
| 5 | a | 0.34 | 0.65 | 0.87 | 1.07 | 1.41 | 1.72 | 1.99 |
| | b | -2.43 | -1.81 | -0.86 | -0.30 | 0.23 | 0.54 | 1.11 |
| | c | 0.03 | 0.03 | 0.09 | 0.18 | 0.27 | 0.33 | 0.46 |
| 6 | a | 0.34 | 0.53 | 0.74 | 0.95 | 1.16 | 1.36 | 1.47 |
| | b | -1.50 | -1.23 | -0.66 | 0.15 | 0.66 | 1.25 | 1.80 |
| | c | 0.04 | 0.07 | 0.14 | 0.22 | 0.29 | 0.41 | 0.44 |
| 7 | a | 0.34 | 0.55 | 0.71 | 0.94 | 1.24 | 1.43 | 1.61 |
| | b | -1.45 | -1.21 | -0.74 | 0.25 | 0.70 | 1.46 | 1.65 |
| | c | 0.01 | 0.03 | 0.11 | 0.20 | 0.25 | 0.36 | 0.37 |
| 8 | a | 0.36 | 0.50 | 0.69 | 0.84 | 0.97 | 1.22 | 1.63 |
| | b | -2.13 | -1.62 | -0.69 | 0.79 | 1.75 | 2.07 | 2.42 |
| | c | 0.01 | 0.02 | 0.12 | 0.17 | 0.22 | 0.34 | 0.40 |

### Table 30: 2PL Operational Item Parameter Five-Point Summary and Range, Mathematics

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3<br><10 items | a | 0.91 | | | | | | 1.15 |
| | b | -0.32 | | | | | | 1.07 |
| 4<br><10 items | a | 0.68 | | | | | | 0.84 |
| | b | -0.79 | | | | | | 1.00 |
| 5<br><10 items | a | 0.43 | | | | | | 0.99 |
| | b | -0.04 | | | | | | 1.20 |
| 6 | a | 0.54 | 0.57 | 0.81 | 1.13 | 1.21 | 1.29 | 1.29 |
| | b | -0.58 | -0.58 | 0.27 | 0.59 | 1.26 | 1.61 | 1.83 |
| 7 | a | 0.58 | 0.63 | 0.87 | 1.04 | 1.16 | 1.40 | 1.43 |
| | b | -0.20 | -0.09 | 0.75 | 1.02 | 1.20 | 1.61 | 2.27 |
| 8 | a | 0.31 | 0.34 | 0.59 | 0.66 | 0.82 | 0.89 | 1.00 |
| | b | -0.13 | 0.07 | 1.37 | 1.53 | 1.86 | 3.13 | 3.68 |

### Table 31: GPCM Operational Item Parameter Range, Mathematics

| Grade* | Parameter | Min | Max |
|---|---|---|---|
| 5 | a | 0.51 | 0.69 |
| | b | -0.52 | 0.35 |
| 7 | a | 0.37 | 0.69 |
| | b | 0.36 | 2.21 |
| 8 | a | 0.39 | 0.67 |
| | b | 0.69 | 2.78 |

*Grades 3, 4, and 6 had no GPCM Items; all other grades had less than 10 GPCM items

### Table 32: 3PL Operational Item Parameter and Five-Point Summary and Range, EOC

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| Algebra 1 | a | 0.45 | 0.57 | 0.95 | 1.08 | 1.31 | 1.77 | 2.37 |
| | b | -2.19 | -1.14 | 0.19 | 0.66 | 1.28 | 1.78 | 2.80 |
| | c | 0.02 | 0.04 | 0.11 | 0.21 | 0.27 | 0.38 | 0.48 |
| Algebra 2 | a | 0.43 | 0.62 | 0.97 | 1.21 | 1.59 | 1.98 | 2.02 |
| | b | -0.20 | -0.09 | 0.40 | 0.65 | 1.16 | 1.39 | 1.93 |
| | c | 0.13 | 0.14 | 0.18 | 0.28 | 0.34 | 0.44 | 0.49 |
| Geometry | a | 0.61 | 0.90 | 1.05 | 1.18 | 1.37 | 1.76 | 2.07 |
| | b | -1.12 | -0.88 | -0.18 | 0.32 | 0.75 | 1.33 | 1.57 |
| | c | 0.08 | 0.10 | 0.18 | 0.23 | 0.27 | 0.41 | 0.58 |

### Table 33: 2PL Operational Item Parameter Five-Point Summary and Range, EOC

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| Algebra 1 | a | 0.37 | 0.50 | 0.77 | 1.11 | 1.34 | 1.71 | 1.95 |
| | b | -0.14 | -0.01 | 0.90 | 1.52 | 2.09 | 2.54 | 3.00 |
| Algebra 2 | a | 0.37 | 0.71 | 0.97 | 1.20 | 1.56 | 1.71 | 1.85 |
| | b | -0.11 | 0.63 | 1.12 | 1.53 | 2.09 | 2.42 | 2.58 |
| Geometry | a | 0.68 | 0.78 | 1.00 | 1.13 | 1.54 | 1.78 | 1.95 |
| | b | -1.23 | -0.85 | -0.02 | 1.03 | 1.43 | 1.71 | 1.86 |

## Table 34: GPCM Operational Item Parameter Five-Point Summary and Range, EOC

| Course* | Parameter | Min | Max |
|---|---|---|---|
| Algebra 1 | **a** | 0.49 | 0.92 |
| | **b** | 1.84 | 2.10 |
| Algebra 2 | **a** | 0.80 | 0.81 |
| | **b** | 1.50 | 1.57 |
| Geometry | **a** | 0.72 | 0.72 |
| | **b** | 1.93 | 1.93 |

*All subjects had less than 10 GPCM items

## Table 35: 3PL Operational Item Parameter Five-Point Summary and Range, ELA

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | **a** | 0.25 | 0.51 | 0.76 | 0.93 | 1.08 | 1.29 | 1.54 |
| | **b** | -1.18 | -1.06 | -0.15 | 0.45 | 0.90 | 1.39 | 1.68 |
| | **c** | 0.02 | 0.02 | 0.13 | 0.19 | 0.23 | 0.27 | 0.32 |
| 4 | **a** | 0.30 | 0.38 | 0.56 | 0.73 | 0.95 | 1.15 | 1.20 |
| | **b** | -2.26 | -2.11 | -1.19 | -0.61 | 0.23 | 0.91 | 1.40 |
| | **c** | 0.02 | 0.03 | 0.07 | 0.12 | 0.20 | 0.36 | 0.43 |
| 5 | **a** | 0.35 | 0.41 | 0.56 | 0.73 | 0.88 | 1.17 | 1.26 |
| | **b** | -2.47 | -2.05 | -1.12 | -0.27 | 0.35 | 1.33 | 1.55 |
| | **c** | 0.02 | 0.04 | 0.12 | 0.20 | 0.24 | 0.36 | 0.45 |
| 6 | **a** | 0.34 | 0.47 | 0.58 | 0.79 | 1.03 | 1.29 | 1.49 |
| | **b** | -2.27 | -1.84 | -0.61 | 0.07 | 0.62 | 1.30 | 1.70 |
| | **c** | 0.02 | 0.03 | 0.07 | 0.16 | 0.23 | 0.40 | 0.45 |
| 7 | **a** | 0.31 | 0.34 | 0.55 | 0.75 | 0.92 | 1.10 | 1.31 |
| | **b** | -2.59 | -1.47 | -0.69 | -0.08 | 0.36 | 0.95 | 1.58 |
| | **c** | 0.02 | 0.03 | 0.08 | 0.20 | 0.26 | 0.40 | 0.56 |
| 8 | **a** | 0.31 | 0.40 | 0.55 | 0.78 | 1.08 | 1.29 | 1.47 |
| | **b** | -2.56 | -1.50 | -0.69 | -0.05 | 0.30 | 1.16 | 1.34 |
| | **c** | 0.01 | 0.02 | 0.07 | 0.14 | 0.23 | 0.28 | 0.31 |
| 9 | **a** | 0.43 | 0.56 | 0.68 | 0.86 | 1.07 | 1.38 | 1.52 |
| | **b** | -2.49 | -1.70 | -0.58 | -0.18 | 0.61 | 1.35 | 2.63 |
| | **c** | 0.002 | 0.01 | 0.12 | 0.23 | 0.28 | 0.41 | 0.60 |
| 10 | **a** | 0.34 | 0.38 | 0.52 | 0.71 | 0.87 | 1.15 | 1.37 |
| | **b** | -2.12 | -1.34 | -0.83 | -0.38 | 0.45 | 1.09 | 1.52 |
| | **c** | 0.01 | 0.02 | 0.05 | 0.19 | 0.26 | 0.46 | 0.76 |

Table 36: 2PL Operational Item Parameter Five-Point Summary and Range, ELA

| Grade | Parameter | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3<br><10 items | **a** | 0.30 | | | | | | 1.23 |
| | **b** | -1.30 | | | | | | 1.58 |
| 4<br><10 items | **a** | 0.53 | | | | | | 1.05 |
| | **b** | -2.20 | | | | | | 1.28 |
| 5 | **a** | 0.44 | 0.45 | 0.50 | 0.55 | 0.64 | 1.06 | 1.07 |
| | **b** | -2.16 | -1.45 | -0.34 | -0.08 | 1.02 | 1.70 | 1.81 |
| 6 | **a** | 0.51 | 0.57 | 0.69 | 0.81 | 0.91 | 1.07 | 1.10 |
| | **b** | -2.68 | -1.90 | -0.38 | 0.07 | 0.48 | 1.17 | 1.34 |
| 7<br><10 items | **a** | 0.41 | | | | | | 1.30 |
| | **b** | -2.01 | | | | | | 0.26 |
| 8 | **a** | 0.42 | 0.43 | 0.54 | 0.62 | 0.75 | 0.86 | 0.88 |
| | **b** | -1.88 | -1.79 | -1.40 | -0.81 | 0.26 | 0.63 | 0.65 |
| 9<br><10 items | **a** | 0.31 | | | | | | 0.82 |
| | **b** | -2.02 | | | | | | 1.09 |
| 10<br><10 items | **a** | 0.32 | | | | | | 1.32 |
| | **b** | -2.32 | | | | | | 1.02 |

Table 37: GPCM Operational Item Parameter Five-Point Summary and Range, ELA

| Grade* | Parameter | Min | Max |
|---|---|---|---|
| 3 | **a** | 0.33 | 0.70 |
| | **b** | -2.28 | -0.29 |
| 4 | **a** | 0.36 | 1.12 |
| | **b** | -1.39 | 1.10 |
| 5 | **a** | 0.54 | 0.88 |
| | **b** | -1.94 | 0.58 |
| 6 | **a** | 0.39 | 1.08 |
| | **b** | -1.5 | 0.57 |
| 7 | **a** | 0.63 | 1.22 |
| | **b** | -1.51 | 0.51 |
| 8 | **a** | 0.57 | 1.15 |
| | **b** | -1.69 | 0.76 |
| 9 | **a** | 1.26 | 1.35 |
| | **b** | -1.41 | 0.13 |
| 10 | **a** | 0.36 | 1.27 |
| | **b** | -1.63 | 0.56 |

* All grades had less than 10 GPCM items

# 7. Summary of Form Development/Administration Algorithms

## 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 examinee raw scores or percent-correct scores that are expected at given ability levels. When two tests are developed to measure the same ability, their scores can be equated through the use of 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 show the TCCs by grade and subject based on the final operational item parameters from the spring 2015 calibrations.

## 7.2 Estimates of Classification Consistency

See Classification Accuracy report in Volume 7.

## 7.3 Reporting Scales

For spring 2015 only, the FSA ELA, Mathematics, and EOC tests report T scores and percentile ranks for each student. The score is based on the operational items presented to the student. Section 8.1 describes exactly how scores were computed.

Appendix F provides a summary of T scores and scale scores.

# 8. Scoring

## 8.1 FSA Scoring

### 8.1.1 Maximum Likelihood Estimation

The FSA tests were based on the 3-parameter logistic model (3PL) and Generalized Partial Credit Models (GPCM) of Item Response Theory models, with the 2PL treated as a special case of the 3PL. Theta scores were generated using "pattern scoring," a method which scores students differently depending on how they answer individual items.

**Likelihood Function**

The likelihood function for generating the maximum likelihood estimates (MLEs) is based on a mixture of items types and can therefore be expressed as:

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

where:

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

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

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

$$Q_i = 1 - P_i$$

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

A student's theta (i.e., MLE) 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 $\theta_t$ denotes the estimated $\theta$ 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 3PL or 2 PL model.

### Standard Errors of Estimate

Whenever the MLE is available, the standard error of the MLE is estimated by

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

where

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

where $N_{CR}$ is the number of items that are scored using the GPCM model and $N_{3PL}$ is the number of items scored using 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 FSA 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 standard error (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 FSA tests included items that were scored using the 3PL, 2PL, and GPCM from item response theory. 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:

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

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 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 T Scores and Percentile Rank

For spring 2015 only, both T scores and percentile ranks were reported because standard setting was not completed by the time scores were required to be reported. T scores are standardized scores with the mean of 50 and standard deviation of 10. The percentile rank of a score is the percentage of scores (T scores) in its frequency distribution that are at or below it. Note that the distribution for percentile ranks was based on either the entire population or early processing sample depending on the grade and subject.

T scores were computed using the theta scores as follows:

$$\mathrm{T_i} = \mathrm{round}\left(\hat{\theta}_{\mathrm{i}} * 10 + 50\right)$$

where $\hat{\theta}_{\mathrm{i}}$ is an individual student's ability estimate obtained from maximum likelihood estimation in AIR's scoring engine. T scores were rounded to the nearest whole number for reporting. Since all theta values were between -3 and 3, T scores fall between 20 and 80. After converting the theta scores to T scores, percentiles were found using

$$P_i = \frac{[.5 * freq(T) + C(T)]}{N} * 100$$

where $freq(T)$ is the number of students at the given T score, C(T) is the number below that T score, and N is the total number of students.

Reported scores were from 1 to 99, with the ends constrained as follows:

a. If $P_i < 1$, then set $P_i = 1$
b. If $P_i > 99$, then set $P_i = 99$

The standard error of T score was computed by transformation of standard error of MLE as:

$$se(T) = se(\bar{\theta}) * 10$$

where $se(\bar{\theta})$ is the average standard error of MLE for all examinees at a given T score. This ensures that all examinees at a given T score have the same standard error of T score.

Appendix F provides a summary of T scores.

### 8.1.3 Scale Scores

There are two scale types created for the FSA:

- A vertical scale score for ELA grades 3 through 10 and Mathematics grades 3 through 8
- A within-test scaled score for EOC tests

Table 38 shows the theta to scaled score transformation equations.

*Table 38: Theta to Scale Score Transformation Equations*

| Subject | Grade | Theta to Scale Score Transformation |
|---|---|---|
| ELA | 3 | Scale Score= round(theta *20.000000 + 300.000000) |
| ELA | 4 | Scale Score = round(theta *20.237420 + 311.416960) |
| ELA | 5 | Scale Score = round(theta *21.230040 + 320.961420) |
| ELA | 6 | Scale Score = round(theta *21.861120 + 325.061500) |
| ELA | 7 | Scale Score = round(theta *21.581900 + 332.124320) |
| ELA | 8 | Scale Score = round(theta *21.531360 + 338.432720) |
| ELA | 9 | Scale Score = round(theta *21.751840 + 341.749740) |
| ELA | 10 | Scale Score = round(theta *21.284300 + 348.328540) |
| Mathematics | 3 | Scale Score= round(theta *20.000000 + 300.000000) |
| Mathematics | 4 | Scale Score = round(theta *20.899320 + 313.617800) |
| Mathematics | 5 | Scale Score = round(theta *22.050760 + 321.802560) |
| Mathematics | 6 | Scale Score = round(theta *21.684500+ 325.299220) |
| Mathematics | 7 | Scale Score = round(theta *20.379620 + 330.157540) |
| Mathematics | 8 | Scale Score = round(theta *19.952780 + 332.946420) |
| Algebra 1 |  | Scale Score= round(theta *25.000000 + 500.000000) |
| Algebra 2 |  | Scale Score= round(theta *25.000000 + 500.000000) |
| Geometry |  | Scale Score= round(theta *25.000000 + 500.000000) |

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

1. The same linear transformation was used for all students within a grade.
2. Rounded to the nearest integer (e.g., 302.4 becomes 302; 302.5 becomes 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 38.

### 8.1.4 Performance Levels

Each student is assigned a performance category according to his or her accountability scale score. Table 39, Table 40, and Table 41 provide the cut scores for performance standards for ELA, Mathematics, and EOC.

*Table 39: 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 |
| 7 | 318 | 333 | 346 | 360 |
| 8 | 322 | 337 | 352 | 366 |
| 9 | 328 | 343 | 355 | 370 |
| 10 | 334 | 350 | 362 | 378 |

*Table 40: 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 41: 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 |
| Algebra 2 | 497 | 511 | 529 | 537 |
| Geometry | 486 | 499 | 521 | 533 |

### 8.1.5 Reporting Category Scores

In addition to overall scores, students also receive scores on reporting categories. Let $\mathrm{b}_{\mathrm{sq}}$ 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 $\mathrm{b}_{\mathrm{sq}}$. 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, along with summaries of scores from spring 2015.

# 9. STATISTICAL SUMMARY OF TEST ADMINISTRATION

## 9.1 DEMOGRAPHICS OF TESTED POPULATION, BY ADMINISTRATION

Table 42 through Table 44 present the distribution of students, in counts and in percentages, who participated in the spring administration of 2014–2015 FSA by grade and subject. The subgroups reported here are gender, ethnicity, students with disabilities (SWD), and English language learners (ELL).

It should be noted that the numbers presented here are based on the Reported Status in the final spring SSR files and may vary slightly from the numbers reported in the Online Reporting System (ORS) for two reasons related to reporting. First, ORS is designed to only report aggregations for students who are eligible to test throughout the test administration. ORS counts do not include students whose enrolled grade was end-dated before the end of the administration (i.e., these students probably left the school system), though their individual student records are still there for users to see. Second, there may be some students whose enrolled grade is not in the reported range of ORS. These two reporting criteria may contribute to the discrepancies between ORS and all the records in SSR files.

Table 42: Distribution of Demographic Characteristics of Tested Population, FSA Mathematics

| Grade | Group | All Students | Female | Male | African-American | Hispanic | White | SWD | ELL |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *N* | 215473 | 104667 | 110806 | 49061 | 70002 | 78486 | 20991 | 28370 |
| | % | | 48.6 | 51.4 | 22.8 | 32.5 | 36.4 | 10.9 | 13.2 |
| 4 | *N* | 199351 | 97812 | 101539 | 43478 | 62862 | 75916 | 20960 | 22265 |
| | % | | 49.1 | 50.9 | 21.8 | 31.5 | 38.1 | 11.8 | 11.2 |
| 5 | *N* | 199010 | 97980 | 101030 | 42531 | 62965 | 76093 | 21697 | 19455 |
| | % | | 49.2 | 50.8 | 21.4 | 31.6 | 38.2 | 12.3 | 9.8 |
| 6 | *N* | 191091 | 93427 | 97664 | 42240 | 60345 | 72709 | 20996 | 14470 |
| | % | | 48.9 | 51.1 | 22.1 | 31.6 | 38.0 | 12.1 | 7.6 |
| 7 | *N* | 179194 | 87628 | 91566 | 40662 | 56159 | 68292 | 19355 | 13677 |
| | % | | 48.9 | 51.1 | 22.7 | 31.3 | 38.1 | 11.7 | 7.6 |
| 8 | *N* | 123928 | 58643 | 65285 | 33337 | 40316 | 42948 | 18726 | 11655 |
| | % | | 47.3 | 52.7 | 26.9 | 32.5 | 34.7 | 15.5 | 9.4 |

Table 43: Distribution of Demographic Characteristics of Tested Population, EOC

| Grade | Group | All Students | Female | Male | African-American | Hispanic | White | SWD | ELL |
|---|---|---|---|---|---|---|---|---|---|
| Algebra 1 | *N* | 203235 | 101197 | 102038 | 44036 | 61690 | 80422 | 18877 | 12119 |
| | % | | 49.8 | 50.2 | 21.7 | 30.4 | 39.6 | 10.2 | 6.0 |
| Geometry | *N* | 195113 | 98300 | 96813 | 41059 | 60070 | 77614 | 16275 | 9503 |
| | % | | 50.4 | 49.6 | 21.0 | 30.8 | 39.8 | 9.1 | 4.9 |
| Algebra 2 | *N* | 158254 | 81643 | 76611 | 31959 | 48221 | 62989 | 9645 | 5593 |
| | % | | 51.6 | 48.4 | 20.2 | 30.5 | 39.8 | 6.7 | 3.5 |

Table 44: Distribution of Demographic Characteristics of Tested Population, FSA ELA

| Grade | Group | All Students | Female | Male | African-American | Hispanic | White | SWD | ELL |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *N* | 215317 | 104613 | 110704 | 49098 | 69748 | 78572 | 21018 | 27956 |
| | % | | 48.6 | 51.4 | 22.8 | 32.4 | 36.5 | 10.9 | 13.0 |
| 4 | *N* | 197681 | 97182 | 100499 | 43146 | 62092 | 75448 | 20632 | 21569 |
| | % | | 49.2 | 50.8 | 21.8 | 31.4 | 38.2 | 11.8 | 10.9 |
| 5 | *N* | 196812 | 97053 | 99759 | 42065 | 62038 | 75431 | 21399 | 18753 |
| | % | | 49.3 | 50.7 | 21.4 | 31.5 | 38.3 | 12.3 | 9.5 |
| 6 | *N* | 192614 | 94326 | 98288 | 41890 | 60399 | 73492 | 20564 | 13849 |
| | % | | 49.0 | 51.0 | 21.7 | 31.4 | 38.2 | 11.9 | 7.2 |
| 7 | *N* | 192024 | 94575 | 97449 | 41975 | 59519 | 74280 | 19152 | 13150 |
| | % | | 49.3 | 50.7 | 21.9 | 31.0 | 38.7 | 11.1 | 6.8 |
| 8 | *N* | 198412 | 97629 | 100783 | 43766 | 61762 | 76467 | 20251 | 12281 |
| | % | | 49.2 | 50.8 | 22.1 | 31.1 | 38.5 | 11.2 | 6.2 |
| 9 | *N* | 201252 | 100385 | 100867 | 43615 | 60497 | 80344 | 19345 | 11025 |
| | % | | 49.9 | 50.1 | 21.7 | 30.1 | 39.9 | 10.4 | 5.5 |
| 10 | *N* | 191080 | 95224 | 95856 | 41592 | 57201 | 76737 | 17928 | 10404 |
| | % | | 49.8 | 50.2 | 21.8 | 29.9 | 40.2 | 10.1 | 5.4 |

# 10. Quality Control for Data, Analyses, Scoring, and Score Reports

## 10.1 Data Preparation and Quality Check

AIR's 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 AIR.

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 subsequently passed to two psychometricians who then proceeded to use the files for all analyses independently. Each psychometrician independently implemented the classical and IRT analyses. The results from the two psychometricians (e.g., the IRTPRO output files) were formally compared. Any discrepancies were identified and resolved.

When all classical and IRT results matched from the independent analysts, the results were uploaded to the secure file transfer protocol (SFTP) for review. FDOE psychometricians and HumRRO, a third party independent contractor, also completed independent replications. During calibrations, daily calls were held with all parties to discuss classical statistics and IRT parameters. Content experts from AIR and TDC also reviewed classical statistics and gave input to the discussion. Results were approved by FDOE only when there was 3-way replication and verification.

The daily calibration calls were an important source for quality control and typically proceeded in an iterative fashion. Typically, one to two 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 team 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.

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

## 10.2 Scoring Quality Check

Prior to the operational testing window, AIR's scoring engine was tested to ensure that the MLEs produced by the engine were accurate. This is a process referred to as mock data. During mock data, AIR established all systems and simulated item response data as if real students responded to the test items. We then tested all programs and verified all results before implementing the operational test.

Once final operational item calibrations were complete and approved by FDOE, item parameters were uploaded to AIR's item tracking system (ITS), and student scores, including MLEs, Tscores, percentiles ranks, and scale scores, were generated via the scoring engine.

Similar to the verification process with calibrations, independent score checks were performed by AIR, FDOE, and HumRRO. Scores were only approved by FDOE when there was a 3-way replication and verification.

## 10.3 Score Report Quality Check

Two types of score reports were produced for the 2014–2015 FSA: online reports and printed reports. The FSA online reporting system (ORS) provided the information on student performance and the aggregated summary at various levels (e.g., the district). The paper individual student reports (i.e., family reports) were provided to families of students who took the FSA tests.

Before deploying the 2014–2015 ORS, various test cases were produced. The test cases were generated based on users' role, functionality, and jurisdiction in the ORS. Each test case described a scenario and the expected result of the scenario. After all the applicable test cases were executed successfully on the trial site without any issues, the codes were then deployed to the live site.

AIR also implemented a series of quality control steps to ensure error-free production of paper family-score reports. To begin, using several types of dummy data, members from the AIR score reporting team compared proofs with mock-ups and communicated with the programmers to ensure that the reports were printing as they should appear. These dummy data were created to test the accurate placement of all variables on the score reports and to review graphic alignment. After thoroughly testing the code using the dummy data, AIR then reviewed thousands of reports with live data to ensure full accuracy of the data. The last quality assurance phase occurred at the print site. AIR provided training to print vendors on processes and procedures in order to ensure that the correct numbers of reports were printed, packaged, and shipped. Several AIR staff members also checked the reports as they were printed and packaged to ensure that they looked as they should and were packaged and shipped to the correct locations.

# 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.

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.

Education Bill of 2008. SB 1908, Florida 2008 Legislative Session. (2008).

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. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, *9*(1), 25–44.

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.

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

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

van der Linden, W. J. and 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.

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2005). RAGE-RGEQUATE Manual [Computer software and manual]. Iowa City, IA: University of Iowa.

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.