Florida
Standards Assessments

Florida Standards Assessments

2018–2019

Volume 2
Test Development

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

ACKNOWLEDGMENTS

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 American Institutes for Research (AIR): Dr. Harold Doran, Dr. Dipendra Subedi, Dr. MinJeong Shin, Dr. Shuqin Tao, Dr. Yu Meng, Patrick Kozak, Gloria Buame, and Katherine Kane. Major contributors from the Florida Department of Education include: Dr. Salih Binici, Dr. Qian Liu, Vince Verges, Susie Lee, Jenny Black, Racquel Harrell, Sally Rhodes, Travis Barton, Jiajing Huang, and Yachen Luo.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The Florida Standards Assessments (FSA) were first administered to students during spring 2015, replacing the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0) in English language arts (ELA) and Mathematics. In spring 2019, students in grades 3–6 Reading and Mathematics were administered fixed operational forms on paper. Students in grades 7–8 Mathematics and grades 7–10 Reading were administered fixed operational forms online. Online operational end-of-course (EOC) assessments were given to students taking Algebra 1 and Geometry. The online versions included the use of several technology-enhanced item types. For all online assessments, paper accommodated versions were available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicated such a need. For the ELA Writing component, the forms were administered on paper for students in grades 4–6 and online for students in grades 7–10, with paper-based accommodations offered to students whose IEPs or Section 504 Plans stipulated the need. Additional details on the implementation of the assessments can be found in Volume 1 of this technical report.

The interpretation, usage, and validity of test scores rely heavily upon the process of developing the test itself. This volume provides details on the test development process of the FSA that contributes to the validity of the test scores. Specifically, this volume provides evidence to support the following:

- The Test Design Summary/Blueprint stipulated the range of operational items from each reporting category that were required on each form. This document guided item selection and test construction for Mathematics and ELA. The test design summaries for both Mathematics and ELA were updated during the 2015–2016 school year in order to add clarifying language. The most substantial update to the test design summaries was a clarification added to the ELA Test Design Summary to better explain the scoring of the ELA assessment; the design summary now specifically states that the ELA Reading and ELA Writing components are combined to generate one single ELA scale score.
- The Test Item Specifications provided detailed guidance for item writers and reviewers to ensure that FSA items were aligned to the standards they were intended to measure. The Test Item Specifications for both ELA and Mathematics were updated during the 2015–2016 school year in order to add clarifying language. Additional updates were made in 2018–2019 to remove Computer Based Testing (CBT) language in grades administered in paper. A description of the specific changes made can be found on the last page of each document.
- The item development procedures employed for FSA tests were consistent with industry standards.
- The development and maintenance of the FSA item pool plan established an item bank, in which test items cover the range of measured standards, grade-level difficulties, and cognitive complexity (e.g., Depth of Knowledge [DOK]) through the use of both selected-response (SR) keyed items and constructed-response (CR) machine-scored or handscored item types.

- The thorough test development process contributed to the comparability of the online tests and the paper-pencil tests.

2. TEST SPECIFICATIONS

Following the adoption and integration of the Florida Standards into the school curriculum, items and test item specifications were developed to ensure that the tests and their items were aligned to the Standards and grade-level expectations they were intended to measure. Test item specifications were developed by the Florida Department of Education and content specialists.

The FSA test item specifications are based on the Florida Standards and the Florida course descriptions. They are a resource that defines the content and format for the test and test items for item writers and reviewers. Each grade-level and course specifications document indicates the alignment of items with the Florida Standards and also serves to provide all stakeholders with information about the scope and function of the FSA. In addition to these general guidelines, specifications for FSA ELA Reading and Writing components also include guidelines for developing reading and writing passages and prompts, such as length, type, and complexity.

2.1 BLUEPRINT DEVELOPMENT PROCESS

A test design summary/blueprint for each assessment identifies the number of items, item types, item distribution across Depth of Knowledge, and reporting categories.

The construction of the blueprints for the Florida Standards Assessments (FSA) in ELA and Mathematics is evidenced by the ELA and Mathematics Test Design Summary documents found at http://www.fsassessments.org/about-the-fsas/. These documents were created using Florida's course descriptions as the basis for the design. The course descriptions can be found on the CPALMS website at: http://www.cpalms.org/Public/search/Course.

The ELA and Mathematics content experts at the Test Development Center (TDC) conferred with content experts in the Florida Department of Education's Bureau of Standards and Instructional Support and Just Read, Florida! office to solidify the content of the blueprints. These meetings and calls occurred in May and June 2014.

The reporting categories for the ELA Reading component were derived from the applicable "Cluster" naming convention in the Florida Standards, and the percentages of the reporting categories within the tests were derived from considering the number, complexity, and breadth of the Standards to be assessed. Speaking and listening standards were folded into the Integration of Knowledge and Ideas reporting category; and applicable language standards were folded into the Craft and Structure reporting category. Guidelines for the weight of each reporting category for the FSA ELA Reading component were determined by Florida's Technical Advisory Committee (TAC). TAC advised that to avoid "statistical noise" generated from the items scored in a small reporting category, a minimum of 15% of the total raw score points should be derived from each reporting category.

The reporting categories for Mathematics were also derived from the "domain" naming convention in the Florida Standards. Like ELA, if a Mathematics domain had too few standards, two or more domains might be combined to make the reporting category 15% of the raw score points of that grade's assessment.

Detailed descriptions for the construct of reporting categories are presented in Appendix A for ELA and Appendix B for Mathematics and EOCs.

2.2 TARGET BLUEPRINTS

Test blueprints provided the following guidelines:

- Length of the test (duration and number of items)
- Content areas to be covered and the acceptable range of items within each content area or reporting category
- Acceptable range of item difficulty for the specified grade level
- Approximate number of field-test items, if applicable
- Descriptions of test item types

This section provides only a summary of the blueprints. Detailed blueprints for each content level are presented in Appendix C for ELA and Appendix D for Mathematics and EOCs.

In all grades and subjects, the assessments were administered as fixed-form tests. The grades 3–6 Reading and Mathematics tests were administered on paper, while the grades 7–10 ELA Reading, grades 7–8 Mathematics, and End-of-Course (EOC) assessments (Algebra 1 and Geometry) were administered online. Additionally, ELA Writing was administered on paper for grades 4–6, and online for grades 7–10. For grades and subjects testing online, paper-pencil-based accommodations were provided if indicated by a student's IEP or Section 504 Plan.

In grades 4–10, the FSA ELA test includes two components, which are combined to provide a whole-test FSA ELA scale score:

1. A text-based writing component in which students respond to one writing task
2. A reading, language, and listening component in which students respond to texts and multimedia content

Writing and Reading component item responses were combined such that the data were calibrated concurrently and subsequently to form an overall ELA score. In this document, the term ELA is used when referring to the combined Reading and Writing assessments. Reading is used when referring only to the Reading test form or items; and Writing is used when referring only to the text-based writing task.

Table 1 displays the blueprint for total test length by grade and subject or course. Each year, approximately six to 10 items on all tests are field-test items and are not used to calculate a student's score. Table 2 displays the number of operational and field-test items on the spring 2019 forms. Writing items are not included in the item counts listed for ELA tests.

Table 1: Blueprint Test Length by Grade and Subject or Course

Subject/Course	Grade	Total Number of Items
Reading	3	56–60
Reading	4	56–60
Reading	5	56–60
Reading	6	58–62
Reading	7	58–62
Reading	8	58–62
Reading	9	60–64
Reading	10	60–64
Mathematics	3	60–64
Mathematics	4	60–64
Mathematics	5	60–64
Mathematics	6	62–66
Mathematics	7	62–66
Mathematics	8	62–66
Algebra 1		64–68
Geometry		64–68

Table 2: Observed Spring 2019 Test Length by Grade and Subject or Course

Subject/Course	Grade	Number of Operational Items	Number of Field-Test Items	Total Items
Reading	3	50	10	60
Reading	4	50	10	60
Reading	5	50	10	60
Reading	6	52	10	62
Reading	7	52	10	62
Reading	8	52	10	62
Reading	9	54	10	64
Reading	10	54	10	64
Mathematics	3	54	10	64
Mathematics	4	54	10	64
Mathematics	5	54	10	64
Mathematics	6	56	10	66
Mathematics	7	56	10	66
Mathematics	8	56	10	66
Algebra 1		58	10	68

Subject/Course	Grade	Number of Operational Items	Number of Field-Test Items	Total Items
Geometry		58	10	68

Reporting categories were utilized to more narrowly define the topics assessed within each content area. Individual scores on reporting categories provide information to help identify areas in which a student may have had difficulty. Table 3, Table 6, and Table 10 provide the percentage of operational items required in the blueprints by content strands, or reporting categories, for each grade level or course. The percentages shown represent an acceptable range of item counts. As many of these items in the ELA Reading component were associated with passages, flexibility was necessary for test construction for practical reasons. The ELA Writing component prompt was not included in these blueprints.

Table 4, Table 7, and Table 11 provide the percentage of test items assessing each reporting category that appeared on the spring 2019 forms. Table 5, Table 8, and Table 12 provide the percentage of test items assessing each reporting category on the spring 2019 paper-based accommodated forms.

Table 3: Blueprint Percentage of Test Items Assessing Each Reporting Category in Reading

Grade	Key Ideas and Details	Craft and Structure	Integration of Knowledge and Ideas	Language and Editing Task
3	15–25%	25–35%	20–30%	15–25%
4	15–25%	25–35%	20–30%	15–25%
5	15–25%	25–35%	20–30%	15–25%
6	15–25%	25–35%	20–30%	15–25%
7	15–25%	25–35%	20–30%	15–25%
8	15–25%	25–35%	20–30%	15–25%
9	15–25%	25–35%	20–30%	15–25%
10	15–25%	25–35%	20–30%	15–25%

Table 4: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in Reading

Grade	Key Ideas and Details	Craft and Structure	Integration of Knowledge and Ideas	Language and Editing Task
3	30%	34%	24%	12%
4	30%	32%	24%	14%
5	36%	30%	20%	14%
6	31%	37%	21%	12%
7	31%	31%	25%	13%
8	25%	33%	27%	15%

Grade	Key Ideas and Details	Craft and Structure	Integration of Knowledge and Ideas	Language and Editing Task
9	30%	30%	28%	13%
10	28%	41%	19%	13%

Table 5: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in Reading—Accommodated Forms

Grade	Key Ideas and Details	Craft and Structure	Integration of Knowledge and Ideas	Language and Editing Task
7	31%	31%	25%	13%
8	25%	33%	27%	15%
9	31%	30%	26%	13%
10	28%	41%	19%	13%

Table 6: Blueprint Percentage of Test Items Assessing Each Reporting Category in Mathematics

Grade	1*	2*	3*	4*	5*
3	48%	17%	35%		
4	21%	21%	25%	33%	
5	39%	28%	33%		
6	15%	30%	15%	19%	21%
7	25%	21%	23%	16%	15%
8	30%	25%	27%	18%	

*See Table 9 for reporting category names.

Table 7: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in Mathematics

Grade	1*	2*	3*	4*	5*
3	48%	17%	35%		
4	20%	20%	26%	33%	
5	39%	28%	33%		
6	14%	30%	14%	20%	21%
7	25%	21%	23%	16%	14%
8	30%	25%	27%	18%	

*See Table 9 for reporting category names.

Table 8: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in Mathematics—Accommodated Forms

Grade	1*	2*	3*	4*	5*
7	25%	21%	23%	16%	14%
8	30%	25%	27%	18%	

Table 9: Reporting Categories Used in Mathematics

Grade	Reporting Category
3	Operations, Algebraic Thinking, and Numbers in Base Ten Numbers and Operations—Fractions Measurement, Data, and Geometry
4	Operations and Algebraic Thinking Numbers and Operations in Base Ten Numbers and Operations—Fractions Measurement, Data, and Geometry
5	Operations, Algebraic Thinking, and Fractions Numbers and Operations in Base Ten Measurement, Data, and Geometry
6	Ratio and Proportional Relationships Expressions and Equations Geometry Statistics and Probability The Number System
7	Ratio and Proportional Relationships Expressions and Equations Geometry Statistics and Probability The Number System
8	Expressions and Equations Functions Geometry Statistics and Probability and The Number System

Table 10: Blueprint Percentage of Test Items Assessing Each Reporting Category in EOC

Course	1*	2*	3*
Algebra 1	41%	40%	19%
Geometry	46%	38%	16%

*See Table 13 for reporting category names.

Table 11: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in EOC

Course	Core Form	1*	2*	3*
Algebra 1	Core 16	41%	40%	19%
	Core 17	41%	40%	19%
	Core 18	41%	40%	19%
	Core 19	41%	40%	19%
Geometry	Core 12	47%	38%	16%
	Core 13	47%	38%	16%
	Core 14	47%	38%	16%
	Core 15	47%	38%	16%

*See Table 13 for reporting category names.

Table 12: Observed Spring 2019 Percentage of Test Items Assessing Each Reporting Category in EOC—Accommodated Forms

Course	1*	2*	3*
Algebra 1	43%	40%	17%
Geometry	47%	38%	16%

*See Table 13 for reporting category names.

Table 13: Reporting Categories Used in EOC

Course	Reporting Category
Algebra 1	Algebra and Modeling Functions and Modeling Statistics and the Number System
Geometry	Congruence, Similarity, Right Triangles, and Trigonometry Circles, Geometric Measurement, and Geometric Properties with Equations Modeling with Geometry

The summary tables show overall that the spring 2019 forms were a match to the blueprint. In almost all cases, the percentages across reporting categories met the blueprint or blueprint range. In the instances where the blueprint was not met, the percentage of items from a reporting category was, at most, 2% away from the blueprint.

In addition to information about reporting categories, the ELA Reading component, Mathematics, and EOC blueprints also contained target information about Depth of Knowledge (DOK). DOK levels are used to measure the cognitive demand of instructional objectives and assessment items. The use of DOK levels to construct the FSA provided a greater depth and breadth of learning and also fulfilled the requirements of academic rigor required by the Every Student Succeeds Act. The DOK level described the cognitive complexity involved when engaging with an item; a higher DOK level required greater conceptual understanding and cognitive processing by the students. It is important to note that the DOK levels were cumulative but not additive. For example, a DOK

level 3 item could potentially contain DOK level 1 and 2 elements; however, DOK level 3 activity cannot be created with DOK level 1 and 2 elements.

Table 14 shows the range of the percentage of items by DOK level by grade and subject or course. Table 15 shows the percentage of items from each DOK on the spring 2019 forms. The table shows that in most cases, the percentage of items from each DOK level met the blueprint. Where the blueprint was not met, there was a maximum of a 6% difference between the blueprint and the forms.

Table 14: Blueprint Percentage of Items by Depth of Knowledge

Grade and Subject	DOK 1	DOK 2	DOK 3
ELA 3–10	10–20%	60–80%	10–20%
Mathematics 3–8	10–20%	60–80%	10–20%
Algebra 1	10–20%	60–80%	10–20%
Geometry	10–20%	60–80%	10–20%

Table 15: Observed Spring 2019 Percentage of Items by Depth of Knowledge

Subject	Grade	DOK 1	DOK 2	DOK 3
Reading	3	26%	64%	10%
Reading	4	14%	68%	18%
Reading	5	16%	66%	18%
Reading	6	13%	63%	23%
Reading	7	19%	56%	25%
Reading	8	19%	67%	13%
Reading	9	13%	61%	26%
Reading	10	13%	63%	24%
Mathematics	3	13%	78%	9%
Mathematics	4	22%	67%	11%
Mathematics	5	13%	74%	13%
Mathematics	6	20%	73%	7%
Mathematics	7	13%	77%	11%
Mathematics	8	21%	66%	13%
Algebra 1	Core 16	21%	69%	10%
	Core 17	21%	71%	9%
	Core 18	17%	74%	9%
	Core 19	19%	72%	9%

Subject	Grade	DOK 1	DOK 2	DOK 3
Geometry	Core 12	17%	72%	10%
	Core 13	17%	71%	12%
	Core 14	16%	72%	12%
	Core 15	16%	74%	10%

The FSA Reading component blueprint also included specifications for the genres of text presented in the passages. Two main types of text were used: literary and informational. Table 16 provides target percentages of test passages assessing each type of text. Summary Table 17 shows that across the grades, the percentage of informational and literary passages was close to the blueprint percentages. There was at most a 11% difference between the blueprint and the forms in grade 4 Reading.

Table 16: Blueprint Percentage of Reading Passage Types by Grade

Grades	Informational	Literary
3–5	50%	50%
6–8	60%	40%
9–10	70%	30%

Table 17: Observed Spring 2019 Percentage of Reading Passage Types by Grade

Grade	Informational	Literary
3	50%	50%
4	61%	39%
5	51%	49%
6	59%	41%
7	64%	36%
8	59%	41%
9	66%	34%
10	72%	28%

2.3 Content-Level and Psychometric Considerations

In addition to test blueprints, several content-level and psychometric considerations were used in the development of the FSA. Content-level considerations included the following:

- Correct responses A–D were evenly represented on the test for multiple-choice (MC) items.
- Selected items addressed a variety of topics (no item clones appeared on the same test).
- Identified correct answer or key was correct.

- Each item had only one correct response (some technology-enhanced items did, in fact, have more than one correct answer, and these items were reviewed to confirm that the number of correct answers matched the number asked for in the item itself).
- Identified item content or reporting category was correct.
- No clueing existed among the items.
- Items were free from typographical, spelling, punctuation, or grammatical errors.
- Items were free of any bias concerns and did not include topics that stakeholders might find offensive.
- Items fulfilled style specifications (e.g., italics, boldface, etc.).
- Items marked do-not-use (DNU) were not selected.

Psychometric considerations included the following:

- A reasonable range of item difficulties was included.
- *p*-values for MC and constructed-response (CR) items were reasonable and within specified bounds.
- Corrected point-biserial correlations were reasonable and within specified bounds.
- No items with negative corrected point-biserial correlations were used.
- Item response theory (IRT) *a*-parameters for all items were reasonable and greater than 0.50.
- IRT *b*-parameters for all items were reasonable and between –2 and 3.
- For MC items, IRT *c*-parameters were less than 0.40.
- Few items with model fit flags were used.
- Few items with differential item functioning (DIF) flags were used.

More information about *p*-values, corrected point-biserial correlations, IRT parameters, and DIF calculations can be found in Volume 1 of this report. The spring 2019 FSA was calibrated and equated to the IRT calibrated item pool. More details about calibration, equating, and scoring can be found in Volume 1 of this report.

3. ITEM DEVELOPMENT PROCEDURES

The item development procedures employed by AIR for the FSA tests were consistent with industry practice. Just as the development of Florida's content and performance standards was an open, consensus-driven process, the development of test items and stimuli to measure those constructs was grounded in a similar philosophy.

Item development began with the following guidelines: the FSA item specifications; the Florida Standards; language accessibility, bias, and sensitivity guidelines; editorial style guidelines; and the principles of universal design. These guidelines ensured that each aspect of a Florida item was relevant to the measured construct and was unlikely to distract or confuse test takers. In addition, these guidelines helped ensure that the wording, required background knowledge, and other aspects of the item were familiar across identifiable groups.

The principles of universal design of assessments mandate that tests are designed to minimize the impact of construct-irrelevant factors in the assessment of student achievement, removing barriers to access for the widest range of students possible. The following seven principles of universal design, as clearly defined by Thompson, Johnstone, and Thurlow (2002), were applied to the FSA development:

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

AIR applied these universal design principles in the development of all test materials, including tasks, items, and manipulatives. Test development specialists receive extensive training in item development. At every step of the review process, adherence to the principles of universal design was confirmed.

In terms of software that supports the item development process, AIR's Item Tracking System (ITS) served as the technology platform to efficiently carry out any item and test development process. ITS facilitated the creation of the item banks, item writing and revision, cataloging of changes and comments, and export of documents (items and passages). ITS enforced a structured review process, ensuring that every item that was written or imported underwent the appropriate sequence of reviews and signoffs; ITS archived every version of each item along with reviewer comments throughout the process. ITS also provided sophisticated pool management features that increased item quality by providing real-time, detailed item inventories and item use histories. Because ITS had the capabilities to be configured to import items in multiple formats (e.g., Microsoft Word, Excel, XML), AIR was able to import items from multiple sources. To support online test delivery, ITS had a unique web preview module that displayed items exactly as they were also presented to students, using the same program code used in the AIR's test delivery

system (TDS). An online test does not have a blueline (print approval) process like a paper-pencil test, and this feature provided an item-by-item blueline capability.

The next section describes the item sources for FSA, and the subsequent sections outline the procedure used for the development and review of new items and the alignment of existing items.

3.1 SUMMARY OF ITEM SOURCES

Items for the spring 2019 FSA came from multiple sources as outlined below.

Student Assessment of Growth and Excellence (SAGE)

AIR, on behalf of the Florida Department of Education (FDOE), negotiated a contract with the State of Utah to use test items from its Student Assessment of Growth and Excellence (SAGE) item bank provisionally until a Florida-specific item bank could be developed. Compared to prior years, the use of SAGE items on spring 2019 forms was significantly less. Only grade 3 Reading included SAGE items as anchor items that were previously tested in Florida.

New Items Written by AIR

New field-test items were also included in the spring 2019 forms, and these items will be used on future FSA test forms. The newly developed field-test items came from two sources: Items were written either for the Florida-specific item bank (denoted as FSA item bank items) or for an AIR item bank to be shared with other states (denoted as AIR Core items). Items were written by AIR content experts or by trained partners. All items undergo a rigorous process of preliminary, editorial, and senior review by AIR and by FDOE's Test Development Center (TDC) content experts, who followed appropriate alignment, content, and style specifications. All of these items were also reviewed by panels of Florida educators and citizens for content accuracy, and to ensure that the test items were fair, unbiased, and included topics acceptable to the Florida public. This review is described in more detail in Section 3.3.1.

Next Generation Sunshine State Standards (NGSSS) Assessment Items

In spring 2019, a total of five NGSSS items were field tested at grades 5 and 7 in Mathematics. They were items that had been field tested in spring 2018 but required re-field testing after edits were implemented based on recommendations from rubric validation committees. NGSSS items that aligned to the Florida Standards were used as core and anchor items on Algebra 1 and Geometry forms. Approximately one-third of the operational items in each course are NGSSS items. These items were previously tested in Florida, and their item statistics were updated based on 2019 performance before being used in scoring.

3.2 ITEM TYPES

One of the important features of the online FSA is the administration of technology-enhanced items. Generally referred to as Machine-Scored Constructed Response (MSCR) items, these include a wide range of item types. MSCR items require students to interact with the test content to select, construct, and/or support their answers.

Table 18 and Table 19 list the Reading, Mathematics, and EOC item types, and provide a brief description of each. For paper-pencil-based accommodations, some of these items must be modified or replaced with other items that assess the same standard and can be scanned and scored electronically. Please see the test design summary/blueprint documents or the test item specifications for specific details. Additional information about the item types can be found in Appendix C for Reading and Appendix D for Mathematics and EOC. Examples of various item types can be found in Appendix E.

Table 18: Reading Item Types and Descriptions

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from a number of options.
multipleselect (MS)	Student selects all correct answers from a number of options.
tablematch (MI)	Student checks a box to indicate if information from a column header matches information from a row. On paper, the student fills in a bubble to indicate if information from a column header matches information from a row.
edittaskwithchoice (ETC)	Student identifies an incorrect word or phrase and chooses the replacement from a number of options. On paper, the student bubbles in the correct word or phrase that should replace the underlined word or phrase from a set of options. One option will always be “correct as is.”
hottext (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. On paper, the student fills in bubbles to indicate which sentences are correct.
multiplechoice, hottextselectable (Two-part HT)	Student selects the correct answers from Part A and Part B. Part A is a multiple-choice or a multiselect item, and Part B is a selectable HT item.
Evidence-Based Selected Response (EBSR)	Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A.

Table 19: Mathematics and EOC Item Types and Descriptions

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from four options.
multipleselect (MS)	Student selects all correct answers from a number of options.
edittaskchoice (ETC)	Student identifies an incorrect word, phrase, or blank and chooses the replacement from a number of options. On paper, the student fills in a bubble to indicate the correct number, word, or phrase that should replace a blank or a highlighted number, word, or phrase.
grid (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
hottext (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. On paper, the student fills in bubbles to indicate which sentences are correct.
equation (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. On paper, the student fills in bubbles indicating numbers and mathematical symbols to create a response. Students respond in response grids

Response Type	Description
	in which they write their answer in the boxes at the top of the grid, then fill in the corresponding bubble underneath each box.
textentrynaturallanguage (NL)	Student uses the keyboard to enter a response into a text field.
tablematch (MI)	Student checks a box to indicate if information from a column header matches information from a row. On paper, the student is directed to fill in a bubble that matches a correct option from a column with a correct option from a row.
tableinput (TI)	Student types numeric values into a given table.
Multi-Interaction (MULTI)	This is an item that contains more than one response type. It could contain more than one of the same response type or a combination of response types.

3.3 DEVELOPMENT AND REVIEW PROCESS FOR NEW ITEMS

3.3.1 Development of New Items

AIR developed field-test items to be embedded in the FSA operational tests. As part of the standard test development process, item writers followed the guidelines in FDOE's approved Test Item Specifications and the Test Design Summary/Blueprint.

AIR staff used the Test Item Specifications to train qualified item writers, each of whom had prior item-writing experience. The item writers were trained at AIR item-writing workshops or had previous training on writing multiple-choice and constructed-response items. An AIR content area assessment specialist worked with Test Development Center content leads to review measurement practices in item writing, and interpret the meaning of the Florida Standards and benchmarks as illustrated by the Test Item Specifications documents. This information, along with the purpose of the assessment, was explained to the item writers. Sample item stems that are included in the specifications documents served as models for the writers to use in creating items to match the Standards. To ensure that the items tapped the range of difficulty and taxonomic levels required, item writers use a method based on Webb's cognitive demands (Webb, 2002) and Depth of Knowledge levels.

Item writing and passage selection were guided by the following principles for each of the item types. When writing items, item writers were trained to develop items that:

- have an appropriate number of correct response options or combinations;
- contain plausible distractors that represent feasible misunderstandings of the content;
- represent the range of cognitive complexities and include challenging items for students performing at all levels;
- are appropriate for students in the assigned grade in terms of reading level, vocabulary, interest, and experience;
- are embedded in a real-world context, where indicated;
- do not provide answers or hints to other items in the set or test;
- are in the form of questions or directions for task completion;

- use clear language and avoid negative constructions unless doing so provides substantial advantages; and
- are free of ethnic, gender, political, socioeconomic, and religious bias.

Similarly, reading passages should:

- represent literary (fiction), informational (nonfiction), multimedia (audio and audio-visual), and practical selections (e.g., nontraditional pieces, including tables, charts, glossaries, indexes);
- provide students with the opportunity to interact with complex, authentic texts that may employ a variety of different structures;
- include multimedia and audio elements when appropriate;
- be of high interest and appropriate readability for the grade level;
- be of appropriate length for the grade level;
- include topics that are in alignment with sensitivity guidelines;
- be free of ethnic, gender, political, and religious bias;
- not provide answers or hints to other items in the test; and
- include real-world texts (consumer or workplace documents, public documents such as letters to the editor, newspaper and magazine articles, thesaurus entries) to the extent possible.

When selecting passages, word count, readability, and text complexity are used in conjunction with other aspects of the passages (level of interest, accessibility of the topic, thematic elements) to determine appropriateness for a particular grade level. Table 20 provides the guidelines used in FSA Reading.

Table 20: Word Counts and Readabilities of Reading Passages in FSA Reading

Grade	Word Count (approximate)	Lexile Range (approximate)
3	100–700	450–900
4	100–900	770–1050
5	200–1000	770–1050
6	200–1100	955–1200
7	300–1100	955–1200
8	350–1200	955–1200
9	350–1300	1080–1400
10	350–1350	1080–1400

In FSA Reading, the texts are categorized into informational and literary texts. *Informational texts* include texts that inform the reader, such as the following:

- Exposition: informational trade books, news articles, historical documents, essays
- Persuasive text: speeches, essays, letters to the editor, informational trade books
- Procedural texts and documents: directions, recipes, manuals, contracts

Literary texts include texts that enable the reader to explore other people's experiences or to simply read for pleasure, such as the following:

- Narrative fiction: historical and contemporary fiction, science fiction, folktales, legends, and myths and fables
- Literary nonfiction: personal essays, biographies/autobiographies, memoirs, and speeches
- Poetry: lyrical, narrative, and epic works; sonnets, odes, and ballads

Department Item Review and Approval

After internal review, the sets of items were reviewed by content specialists at the Test Development Center (TDC). If needed, AIR and TDC content staff discussed requested revisions, ensuring that all items appropriately measured the Florida Standards. The items were then revised by AIR and brought to Florida bias, sensitivity, and content committees for review. After any final adjustments were made to the items, including an editorial review conducted by TDC, the TDC provided a decision for each item: *Accept as Appears*, *Accept as Revised*, or *Reject*. Items that were approved by the TDC were subsequently web-approved and placed on field-test forms.

Committee Review of New Items

All items generated for use on the Florida Statewide Assessments were required to pass a series of rigorous reviews before they could appear as field-test items on operational test forms. The items were reviewed by three committees—the Bias and Sensitivity Committee, the Community Sensitivity Committee, and the Content Item Review Committee.

The Bias and Sensitivity Committees reviewed items for potential bias and controversial content. These committees consisted of Florida reviewers who were selected to ensure geographic and ethnic diversity. These committees ensure that items:

- present racial, ethnic, and cultural groups in a positive light;
- do not contain controversial, offensive, or potentially upsetting content;
- avoid content familiar only to specific groups of students because of race or ethnicity, class, or geographic location;
- aid in the elimination of stereotypes; and
- avoid words or phrases that have multiple meanings.

TDC and AIR reviewed the Bias and Sensitivity Committees feedback and conveyed any issues to the attention of the Content Item Review Committee.

The Content Item Review Committee consisted of Florida classroom teachers or content specialists by grade for each subject area. The primary responsibility of the committee members was to review all new items to ensure that they were free from such flaws as (a) inappropriate readability level, (b) ambiguity, (c) incorrect or multiple answer keys (although some item types may include multiple answer keys by design), (d) unclear instructions, and (e) factual inaccuracy. These items were approved, approved with modifications, or rejected. Only approved items were added to the item pool for the field test stage.

3.3.2 Rubric Validation

After items were field tested, the rubric used for scoring MSCR items was validated by a team of grade-level Florida educators. These individuals reviewed the machine-assigned scores for constructed-response items based on the scoring rubrics and either approved the scoring rubric as it appeared on the field test or suggested revisions to the scoring based on their interpretation of the item task and the rubric.

Similar to the items field tested in 2018, rubrics were reviewed in one of two ways: Items with simpler rubrics were reviewed via frequency tables of all student responses, while items with more complex rubrics were reviewed in 45-response samples.

Items with complex rubrics include grid (GI) items, hottext (HT) draggable items, equation (EQ) items with full keypads, tableinput (TI) items, textentrynaturallanguage (NL) items, and Multi-Interaction (MULTI) items containing at least one of the preceding response types.

Items with simple rubrics include edittaskchoice and edittaskwithchoice (ETC) items, hottext (HT) selectable items, matching (MI) items, equation (EQ) items with simple numeric keypads, multiplechoice and hottextselectable (Two-part HT) items, and any Multi-Interaction (MULTI) items comprised entirely of the preceding response types.

Multiplechoice (MC) items, multipleselect (MS) items, and Evidence-Based Selected Response (EBSR) items do not go through rubric validation.

Prior to the meeting, AIR staff selected a sample of 45 student responses for each item with complex rubrics. The sample consisted of the following:

- 15 responses from students who performed as expected on the item given their overall performance
- 15 responses from students who were predicted to perform well on the item given their overall performance, but instead performed poorly on the item
- 15 responses from students who were predicted to perform poorly on the item given their overall performance, but instead performed well on the item

For items with simple rubrics and all items administered on paper, AIR staff generated frequency tables that contained all student responses for each item.

The Rubric Validation Committee reviewed 45 responses for every item with a complex rubric, having the option to approve the score or suggest a different score based on the committee's understanding of the rubric. For item with simple rubrics, the committee members were shown the

item, along with the correct response and the most frequently selected incorrect responses. TDC and AIR staff ensured that the committee was scoring consistently. The committee meetings used the following procedures:

- ELA committee members were provided with their own binder that contained a PDF version of each item. Materials were collected and shredded at the conclusion of rubric validation for item security. Math committee members were given a laptop allowing them to respond to the items the way a student would be able to respond in a live test.
- Each item was displayed with a projector.
- The committee discussed how to answer the item and how each point was earned.
- For items with complex rubrics, each of the 45 student response papers and machine-assigned scores were displayed with a projector.
- For items with simple rubrics, the item was displayed with a projector, along with the correct response and the most frequently selected incorrect responses.
- If the committee members reached a consensus that a score was incorrect, the committee proposed modifications to the rubric.
- AIR rescored the responses using the revised rubric.
- AIR reviewed the responses that received changed scores to determine if they were correctly scored.
- TDC reviewed the rescored responses and approved the rubric.

If any scores changed based on the Rubric Validation Committee review, AIR staff revised the machine rubric and rescored the item. After the item was rescored, AIR staff reviewed at least 10% of responses for which the score changed. Please note that the immediate rescoring of an item is not feasible in paper-based assessments. This review ensured that committee suggestions were honored, that the item was scored consistently, and that no unintended changes in scoring occurred as a result of the revision to the machine rubric. AIR staff reviewed changes with TDC staff, and TDC staff had one final opportunity to revise the rubric or approve or reject the item.

The approved items were embedded into the spring 2019 operational test forms. At the end of the testing window, AIR conducted classical item analysis on these field-test items to ensure that the items functioned as intended with respect to the underlying scales. AIR's analysis program computed the required item and test statistics for each multiple-choice and constructed-response item to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistical analyses included item discrimination, distractor analysis, item difficulty analysis, and fit analysis. Details of these analyses are presented in Section 5 of Volume 1.

3.4 Development and Maintenance of the Item Pool

As described earlier, 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 reporting category.

In spring 2019, field-test items were embedded in online forms but were field tested on paper in grades 3 through 6 for both Reading and Mathematics. All assessments were fixed-form tests with a predetermined number and location of field-test items. Table 21 and Table 22 provide the number of field-test items by type for Reading, Mathematics, and EOC.

Table 21: Number of Reading Field-Test Items by Type

Item Type	3	4	5	6	7	8	9	10
EBSR	36	22	31	37	9	6	12	36
ETC	0	0	0	30	0	0	0	0
HT	10	7	2	2	1	3	2	6
MC	140	98	99	99	31	36	33	100
MI	23	19	15	6	2	0	1	6
MS	10	12	19	17	8	6	4	15
Two-Part HT	1	1	1	0	0	1	0	0

Table 22: Number of Mathematics and EOC Field-Test Items by Type

Item Type	3	4	5	6	7	8	Algebra 1	Geometry
EQ	55	54	57	36	32	18	2	4
ETC	3	4	5	31	12	13	41	19
GI	0	0	0	0	1	2	9	2
HT	0	0	0	2	0	0	1	0
MC	66	60	65	46	16	11	40	20
MI	11	15	5	16	0	4	2	2
MS	15	19	13	18	1	4	7	4
MULTI	9	6	8	5	8	12	48	33
TI	0	0	0	0	0	0	0	0

3.5 Alignment Process for Existing Items and Results from Alignment Studies

A third-party, independent alignment study was conducted in February 2016. This report can be found in Volume 4 Appendix D of the 2015–2016 FSA Annual Technical Report.

4. TEST CONSTRUCTION

4.1 OVERVIEW

During summer 2018, psychometricians and content experts from FDOE, TDC, and AIR convened for two weeks to build forms for the spring 2019 administration. FSA test construction utilized a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

Beginning in spring 2016, anchor items were included for all grades. Anchor items may be either internal or external. Internal anchor items are operational and count toward a student's score. In grades and subjects that use internal anchor items, internal anchor items appear on all forms. External anchor items are located in embedded slots and do not count toward a student's score. Anchor items, whether internal or external, will be used to link the current year's calibrations to the IRT calibrated item pool.

Anchor items were selected first, and the set of anchor items in any given grade represented the blueprint for that grade to the greatest extent possible. Since anchor items can be considered a mini-test form, the targets for the set of anchor items were the same as the set of operational items.

The form-construction process is highly iterative. Appendix H, the test construction specifications, provides the details of this process. While the subsequent sections also elaborate the process, including the roles and responsibilities of participants, the key steps involved in test construction are summarized here.

1. AIR content staff select the items for the form that follow the test specifications. The anchor items are selected first, and then the "core" items are selected. The anchor item sets and core item sets are designed to match the statistical qualities and content coverage.

2. AIR content staff consult AIR psychometricians to ensure that the form meets the psychometric considerations. The forms are then submitted to TDC content specialists for review. Both TDC and AIR content specialists collaborate to revise the forms and select replacement items as needed. Once a form is approved by TDC content leads, it is sent for review to the AIR psychometric team and then to the FDOE psychometric team.

3. Both the AIR and FDOE psychometric teams evaluate the statistical properties of the constructed forms against the statistical targets outlined in the test construction specifications. This step is also intended to minimize the conditional standard error of measurement around the achievement-level cut scores. The proposed forms are either returned to the content teams for suggested improvements or are approved and forwarded to FDOE leadership for final review.

4. The FDOE leadership team identifies the suitability of the selected items and test forms as a whole and considers the factors such as diversity of topics, the projected level of difficulty, statistical summaries, and match to the test specifications. The FDOE leadership team can either approve the proposed forms or return them with comments to the AIR and FDOE content teams for further revision.

4.1.1 Roles and Responsibilities of Participants

AIR Content Team

AIR ELA and Mathematics content teams were responsible for the initial form construction and subsequent revisions. These initial forms were pivotal to the test construction activities during the preparation period and during onsite test construction. AIR content teams performed the following tasks:

- Selection of the initial set of anchor items
- Selection of the initial set of operational items
- Revision of the anchor and operational item sets according to feedback from senior AIR content staff
- Revision of the anchor and operational item sets according to feedback from AIR psychometricians
- Assistance in the generation of materials for TDC and FDOE review
- Revision of the forms to incorporate feedback from TDC and FDOE

AIR Technical Team

The AIR technical team, which included psychometricians and statistical support associates, prepared the item bank by updating the Item Tracking System (ITS) with current item statistics and provided test construction training to the internal content team. During onsite test construction, at least one psychometrician was facilitating the process with each content area. The technical team performed the following tasks:

- Preparing item bank statistics and updating of AIR's ITS
- Creating the master data sheets (MDS) for each grade and subject
- Providing feedback on the statistical properties of initial item pulls
- Providing explanations surrounding the item bank
- Providing feedback on the statistical properties of each subsequent item selection
- Creating materials for FDOE psychometrician and leadership review

TDC Content Specialists and Leads

TDC content specialists collaborated with AIR content specialists to revise forms and select replacement items. Both parties selected items with respect to the statistical guidelines and the FSA content and blueprint guidelines. Content specialists communicated with content leads and psychometricians if they had concerns about either blueprints or statistical summaries.

TDC content leads reviewed the test forms and provided either approval or feedback to AIR content specialists. Once a form was approved, content leads completed verification logs for FDOE psychometricians to review.

FDOE Psychometrics

The FDOE psychometrics team evaluated the statistical properties of the constructed forms against statistical targets. These targets are outlined in the sample verification log in Appendix F. The proposed forms were either returned to TDC and AIR content teams for additional edits or approved and forwarded to FDOE and TDC leadership for final review.

FDOE and TDC Leadership

All proposed forms were reviewed by the FDOE leadership team to determine the overall suitability of the proposed forms. When evaluating any given form, leadership considered the diversity of topics, projected level of difficulty, statistical summaries, adherence to blueprint, overall challenge to the test takers, and acceptability of test content to the Florida public. The leadership team was given the opportunity to approve proposed forms or return them with comments to AIR's content team for further revision.

4.2 TEST CONSTRUCTION PROCESS

The FSA test construction process began in early summer with the following tasks:

1. Confirmation of test construction checklists and blueprints
2. Identification of key dates for each activity
3. Preparation for onsite meetings, including room reservations and agendas
4. Update of verification logs

After the test construction checklists and blueprints were approved, offsite test construction began.

4.2.1 Offsite Test Construction

Once item calibrations were complete, AIR's technical team updated the item bank with all possible items for test construction. AIR's Item Tracking System (ITS) was updated with the most current item statistics for any given item. Master data sheets (MDS) were also created to assist the content teams at AIR and TDC to select the items and to assist FDOE psychometricians in their form review. For each grade and subject, the MDS lists all items from each administration and provides item characteristics, classical statistics, and item response theory statistics. Items that have been administered multiple times have multiple listings in the MDS.

AIR's content team created initial anchor item lists according to test construction checklists and blueprints. These preliminary versions of the anchor sets were given to AIR's technical team for review. AIR psychometricians compiled statistical summaries and provided feedback. The selection of anchor items was updated to incorporate this feedback. There were often several iterations of the proposed preliminary anchor sets between AIR's teams before final approval of initial anchor item lists. This communication and interaction ensured that the initial anchor item sets delivered to FDOE and TDC were of high quality and representative in terms of both content and item statistics.

At least one week before the onsite meetings, initial anchor item lists and summaries were provided to FDOE and TDC. This allowed for review before onsite face-to-face meetings.

4.2.2 Onsite Meetings

Onsite meetings occurred at the AIR offices. All parties, including program management, were actively involved in onsite test construction. On the morning of the first day, a commencement meeting was held to introduce all team members, explain any changes to test specifications or blueprints, discuss proposed forms, and prioritize upcoming activities. ELA, Mathematics, and EOC content specialists proceeded to their respective rooms to discuss proposed forms. For each grade and subject, there was at least one AIR content specialist and one TDC content specialist present for deliberations; at least one AIR psychometrician was available in each room.

Content specialists discussed proposed anchor item sets considering each item individually, ensuring that the composition of the items satisfied the blueprint and content-level considerations. For spring 2019 test forms, anchor items were selected from previous anchor, core, or Florida field-tested items. In only grade 3 Reading, SAGE owned items with FSA specific parameters were used. Each item was carefully reviewed to confirm that it aligned with Florida Standards and fulfilled statistical criteria. If content experts had questions about item statistics, psychometricians were available to provide clarification.

AIR conducted additional activities during spring 2019 due to transition from Computer-based Testing (CBT) to Paper-based Testing (PBT) for grades 4–6 ELA and grades 3–6 Mathematics. During the summer 2018 test construction, content experts from AIR and TDC created a list of "watch items" among the anchor and Core items selected for the form where the mode effect can possibly exist. The selected watch items were categorized into minimum, moderate, and high based on their potential for mode effect. During operational calibration, the item statistics for these items was monitored and discussed in the calibration call.

Once anchor item sets were judged to be satisfactory from a content perspective, item sets were again reviewed by AIR psychometricians to ensure that they met the psychometric considerations. The psychometric considerations for each form included the test difficulty, target test information, standard error of measurement, and test characteristic curves. The information reviewed at the item level included classical item statistics, differential item functioning (DIF) statistics, item response theory (IRT) parameters, and fit statistics. If any particular item did not meet the statistical criteria, content specialists were asked to submit a replacement item. Once all items satisfied both content and statistical considerations, the verification log was completed, and summary materials were prepared. An example of the verification log can be found in Appendix F. Summary materials are discussed in Section 4.3.

FDOE psychometricians were given the verification log and summary materials to perform their own item-by-item review. If questions about content level or statistical criteria arose, discussions were held with all parties. Anchor item sets were either returned to content specialists with feedback to replace problematic items or approved and passed on to FDOE leadership.

FDOE leadership reviewed the verification log, summary materials, and comments from the FDOE psychometricians. Anchor item sets were once again either approved or returned to content specialists with feedback to replace problematic items, as necessary.

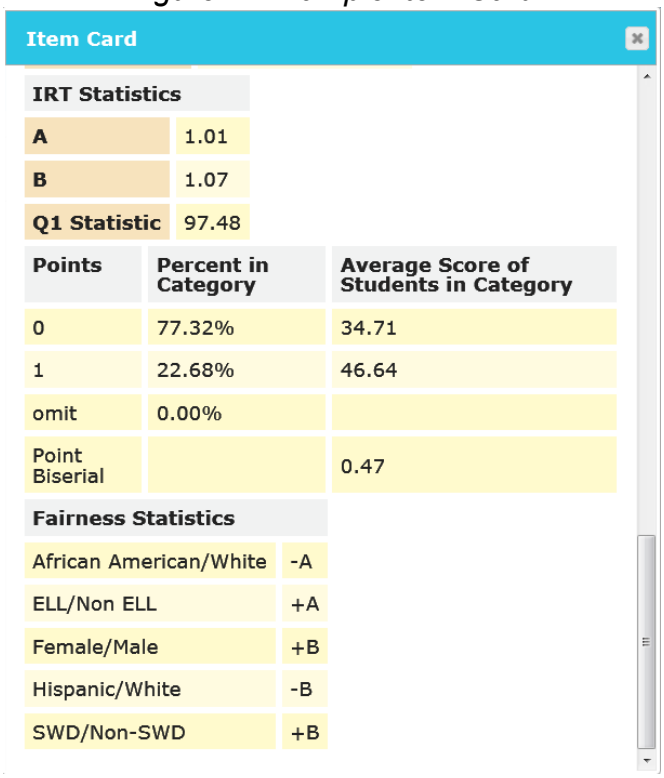
Once an anchor item set was approved, the same process was used to select operational items. Once both anchor item sets and operational items were approved, forms were entered into ITS, where they were evaluated for a final time to confirm that the intended items were placed on the individual forms. Final verification of approval from FDOE was obtained, and the necessary steps were taken to prepare the form for use in AIR's test delivery system (TDS).

4.3 Test Construction Summary Materials

4.3.1 Item Cards

Item cards, generated within ITS, contained statistical information about an individual item. Item cards contained classical item statistics, IRT statistics, and DIF statistics. When possible, item cards also contained a screenshot of the item. This was not possible in the case of some technology-enhanced items. In these instances, the items were viewed directly in ITS. Item cards were typically used to determine the viability of an individual field-test item for operational use in the next administration. Figure 1 provides an example item card.

Figure 1: Example Item Card



4.3.2 Bookmaps

A bookmap is a spreadsheet that lists characteristics of all items on a form. Bookmaps contain information such as:

- Item ID
- Item position

- Form
- Grade
- Role (e.g., operational or field test)
- Item format (e.g., multiple choice)
- Point value
- Answer key
- Reporting category
- Depth of Knowledge (DOK)

Bookmaps were used as an accessible resource to both content specialists and psychometricians to find information about a test form. Bookmaps differed from item cards in that there were no statistical summaries in a bookmap.

4.3.3 Graphical Summaries

In addition to numerical summaries and spreadsheets, it was often useful to create graphical summaries for visualization.

Test Characteristic Curve

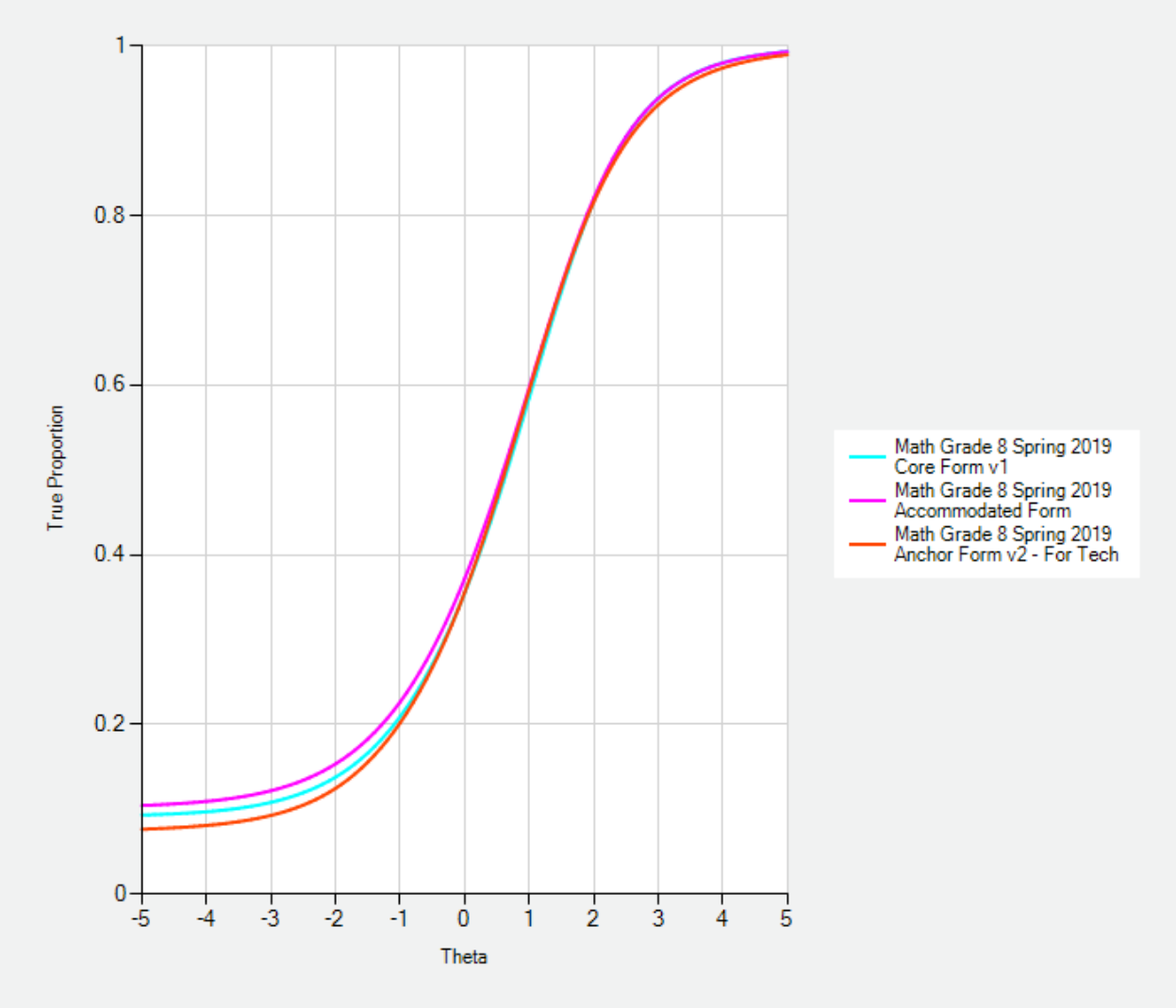
An item characteristic curve (ICC) shows the probability of a correct response as a function of ability, given an item's parameters. Test characteristic curves (TCCs) can be constructed as the sum of ICCs for the items included on any given test. The TCC can be used to determine test taker raw scores or percentage-correct scores that are expected at a given ability level. When two tests are developed to measure the same ability, their scores can be equated through the use of TCCs.

The spring 2018 core form TCCs were the target for the spring 2019 forms. The spring 2019 online TCC was used as a target while building the spring 2019 paper-pencil accommodated forms. Items were selected for the paper-pencil form so that the form TCC matched the online form TCC as closely as possible.

Figure 2 compares the TCCs for both online and paper-pencil forms of grade 8 Mathematics.

Efforts were made to maximize information at the performance cut scores. These general targets were used for guidance, but not as a definitive rule.

Figure 2: TCC Comparisons of Grade 8 Mathematics Online and Paper-Pencil Forms

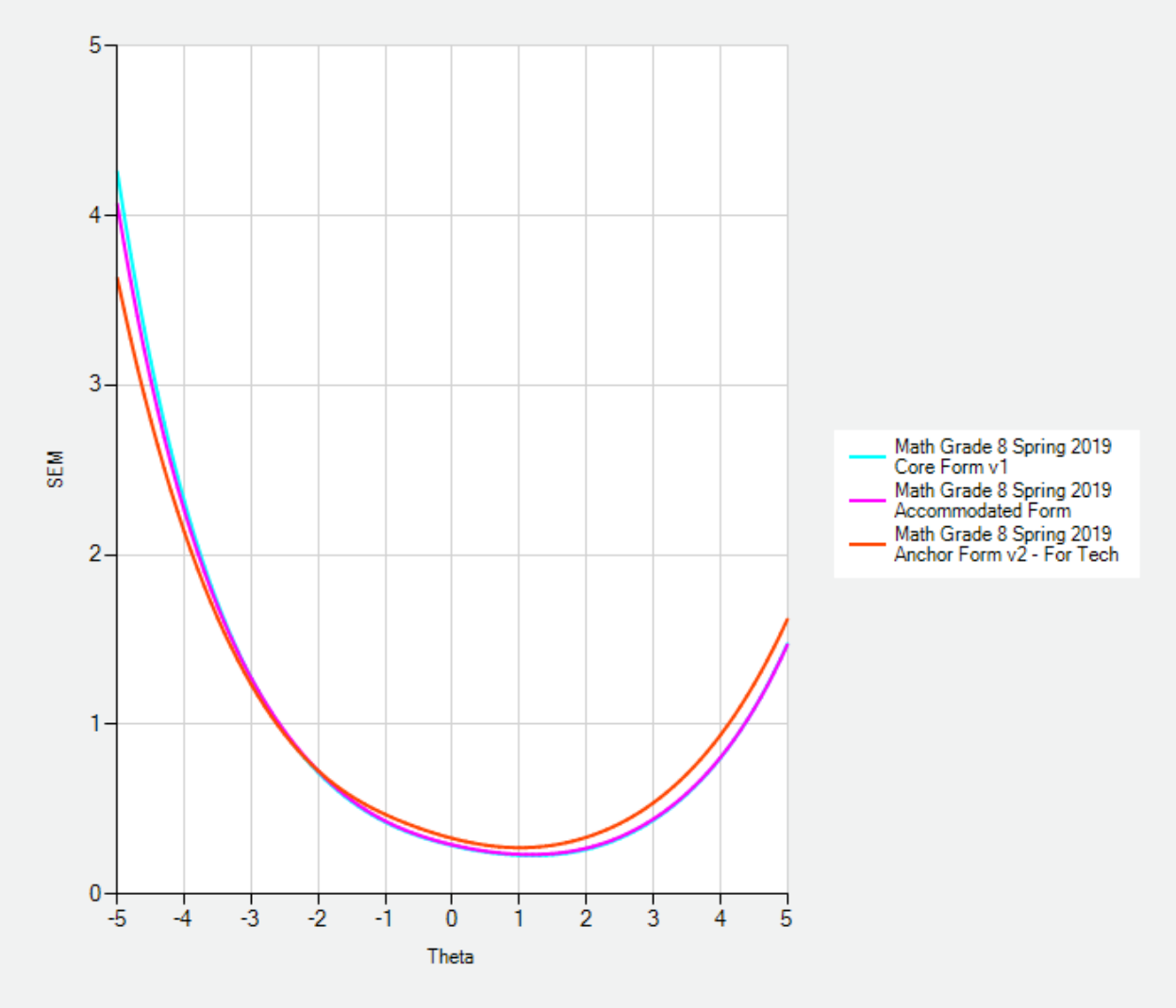


Conditional Standard Error of Measurement Curve

The conditional standard error of measurement (CSEM) curve shows the level of error of measurement expected at each ability level. The CSEM is calculated as the reciprocal of the square root of the test information function, and thus the CSEM is lowest when information is highest. Ability estimates in the middle of the distribution often appear more reliable than the ability estimates at the high and low ends of the scale. Figure 3 compares the CSEM of the grade 8 Mathematics online and paper-pencil forms.

The spring 2018 core forms were the target CSEMs for the spring 2019 forms. However, efforts were made to minimize the standard error at the performance cuts and improve the precision of the test over time rather than adhering to matching the targets. Appendix H, the test construction specifications, provides additional details.

Figure 3: CSEM Comparison of Grade 8 Mathematics Online and Paper-Pencil Forms



4.4 Paper-Pencil Accommodation Form Construction

Student scores should not depend upon the mode of administration or type of test form. Because the FSA grades 7–10 ELA and grades 7–8 Mathematics tests were administered in an online test system, scores obtained via alternate modes of administration must be established as comparable to scores obtained through online testing. This section outlines the overall test development plans that ensured the comparability of online and paper-pencil tests.

During test development, forms across all modes were required to adhere to the same test blueprints, content-level, and psychometric considerations. To build paper-pencil forms, content specialists began with the online form and removed any technology-enhanced items that could not be rendered on paper or machine scored. These items were then replaced with either multiple-choice items or other technology-enhanced items that could be rendered on paper from the same reporting category. In some instances, it was necessary to select replacement items from a different reporting category in order to satisfy statistical expectations; however, all parties ensured that each

reporting category was still appropriately represented in the final test forms. Table 23 provides the number of items replaced between the online and paper-pencil accommodated forms.

Table 23: Number of Item Replacements for Paper-Pencil Accommodated Forms

Test	Number of Items Replaced
Grade 9 ELA	1
Grade 7 Mathematics	15
Grade 8 Mathematics	7
Algebra 1	7
Geometry	9

The online and paper-pencil accommodated forms were then reviewed for their comparability of item counts and point values, both at the overall test level and at the reporting category levels. ELA Reading tests in both administration modes were additionally compared for the distribution of passages by length. The forms were then submitted for psychometric reviews, during which the following statistics were computed and compared between the online and paper-pencil accommodated forms:

- Maximum possible score
- IRT *b*-parameter mean and standard deviation
- IRT *b*-parameter minimum and maximum
- IRT *a*-parameter mean and standard deviation
- IRT *a*-parameter minimum and maximum
- IRT *c*-parameter mean and standard deviation
- IRT *c*-parameter minimum and maximum
- Item *p*-value mean and standard deviation
- Item *p*-value minimum and maximum
- Lowest biserial/polyserial
- Mean biserial/polyserial
- Expected raw score at cut points

A sample output with summary statistics for grade 8 Mathematics is presented in Table 24. As the table shows, the IRT *b*-parameter mean and the item *p*-value mean are similar between the forms.

Parallelism among test forms was further evaluated by comparing TCCs, test information curves, and CSEMs between the online and paper-pencil forms.

Table 24: Test Summary Comparison for Grade 8 Mathematics Online and Paper-Pencil Forms

Type	Statistics	Spring 2018 Core Form	Spring 2019 Core Form	Spring 2019 Accommodated
Overall	Number of Items	56.00	56.00	56.00
	Possible Score	57.00	56.00	56.00
	Difficulty Mean	0.70	0.74	0.69
	Difficulty Standard Deviation	1.11	0.81	0.86
	Difficulty Minimum	-2.71	-1.91	-1.91
	Difficulty Maximum	2.37	2.03	2.03
	Parameter-A Mean	0.81	0.80	0.79
	Parameter-A Standard Deviation	0.25	0.18	0.18
	Parameter-A Minimum	0.35	0.50	0.39
	Parameter-A Maximum	1.58	1.31	1.31
	Parameter-C Mean	0.08	0.14	0.14
	Parameter-C Standard Deviation	0.08	0.08	0.08
	Parameter-C Minimum	0.00	0.01	0.01
	Parameter-C Maximum	0.26	0.33	0.33
	Raw Score Sum	20.72	21.21	22.02
	p-Value Mean	0.36	0.38	0.39
	p-Value Standard Deviation	0.21	0.18	0.19
	p-Value Minimum	0.10	0.11	0.11
	p-Value Maximum	0.88	0.85	0.85
	Lowest Bi/Poly-Serial	0.26	0.30	0.30

5. REFERENCES

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved 02-15-2012, from http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis44.html.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.