Florida
Statewide Assessments

Florida Statewide Assessments

2020–2021

Volume 2
Test Development

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Dr. Christina Schneider, Dr. Dipendra Subedi, Dr. Haiyan Lin, Dr. Tzu-Chun (June) Kuo, Patrick Kozak, Joshua Wallace, Cameron Clark. Contributing staff from Pearson are Dr. Jie (Serena) Lin, Dr. Bradley Ungurait, Ebony Gaines, and Michael Watson. Major contributors from the FDOE are as follows: Vince Verges, Susie Lee, Jenny Black, Dr. Qian Liu, Racquel Harrell, Sally Donnelly, Travis Barton, Leah Glass, Dr. Stacy Skinner, Dr. Salih Binici, Jiajing Huang, Yachen Luo, Gertrudes Velasquez, and Wenyi Li.

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. The Next Generation Sunshine State Standards (NGSSS) were adopted in 2008 to replace the 1996 Sunshine State Standards. The first operational administration of the Science assessments (in grades 5 and 8) and Biology 1 end-of-course (EOC) was during the spring 2012 administration window. During the spring 2013 administration window the first operational administration of the U.S. History EOC assessment occurred. The following administrative year (2014), the first operational Civics EOC assessment was administered. Since fall 2020, all FSA and NGSSS assessments have been collectively referred to as the Florida Statewide Assessments. Additional details on the implementation of the assessments can be found in Volume 1 of this technical report.

In spring testing windows, students in Grades 3–6 Reading and Mathematics are administered fixed operational forms on paper. Students in Grades 7–8 Mathematics and Grades 7–10 Reading are administered fixed operational forms online. The NGSSS Science assessments for students in grades 5 and 8 occurs during the spring testing window. Science is a fixed operational form delivered on paper. Online operational EOC assessments are given to students taking Algebra 1, Geometry, Biology 1, U.S. History, and Civics. The online versions of the Reading, Mathematics, Algebra 1, and Geometry assessments include the use of several technology-enhanced item types. For all online assessments, paper accommodated versions are available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicate such a need. For the ELA Writing component, the forms are 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 stipulate the need.

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 Florida Statewide Assessments 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, ELA, Science, and Social Studies.
    - 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 design summaries for NGSSS Science and EOC have not changed since adoption of each operational test.
- The test item specifications provided detailed guidance for item writers and reviewers to ensure that Florida Statewide Assessments 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 FDOE and committees of experienced Florida educators developed and approved the NGSSS Specifications. The Specifications serve as a resource that defines the content and format of the NGSSS tests and test items.
- The item development procedures employed for Florida Statewide Assessments were consistent with industry standards.
- The development and maintenance of the Florida Statewide Assessments 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]) using 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-based tests (PBTs).

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. NGSSS Science and EOC tests are developed according to the content outlined in the NGSSS at each grade level for each tested subject area. The FDOE and content specialists developed test item specifications.

The Florida Statewide Assessments test item specifications are based on the Florida Standards and the Florida course descriptions. The specifications 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 Florida Statewide Assessments. 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 DOK, and reporting categories.

The blueprint construction for the FSA in ELA and Mathematics is evidenced by the ELA and Mathematics Test Design Summary documents found at https://fsassessments.org/. The NGSSS test design summary is posted on the FDOE website (NGSSS Science and EOC). 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 FDOE's Bureau of Standards and Instructional Support and Just Read, Florida! office to solidify the blueprint content. 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. As with 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.

The NGSSS Science and EOC reporting categories assessed are defined based upon the 2007 and 2008 NGSSS adoption. The NGSSS are divided into benchmarks that identify what a student should be able to do following the completion of each course. The Test Item Specifications document for NGSSS Science and EOC contain benchmark-specific information. The benchmark information provides benchmark clarification statements, content limits, stimulus attributes, and a sample item for each benchmark that could be assessed.

Detailed descriptions for the construct of reporting categories are presented in Appendix A for ELA and Appendix B for Mathematics and EOCs. Similarly, Appendix C provides the detailed descriptions for Science and Appendix D provides the information for NGSSS EOC assessments.

2.1.1 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 E for ELA, Appendix F for Mathematics and EOCs, and Appendix G for NGSSS Science and EOCs.

In all grades and subjects, the assessments are administered as fixed-form tests. The grades 3–6 Reading and Mathematics tests are administered on paper, while the grades 7–10 ELA Reading, grades 7–8 Mathematics, and EOC assessments (Algebra 1 and Geometry) are administered online. Additionally, ELA Writing is administered on paper for grades 4–6 and online for grades 7–10. In spring 2021, typed writing response accommodations were provided for students taking Writing assessments in grades 4-6, therefore, their responses were collected online instead of on paper. NGSSS Science grades 5 and 8 were administered on paper, while the NGSSS EOC assessments were administered online. For grades and subjects testing online, paper-pencil-based accommodations are 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 are combined such that the data are calibrated concurrently 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 6-10 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 2021 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
Science	5	60-66
Science	8	60-66
Biology 1		60-66
U.S. History		50-60
Civics		52-56

Table 2: Observed Spring 2021 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

Subject/Course	Grade	Number of Operational Items	Number of Field-Test Items	Total Items
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
Geometry		58	10	68
Science	5	56	10	66
Science	8	56	10	66
Biology 1		56	10	66
U.S. History		52	8	60
Civics		48	8	56

Reporting categories were used 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, Table 10, and Table 14 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, Table 11, and Table 15 provide the percentage of test items assessing each reporting category that appeared on the spring 2021 forms. Table 5, Table 8, Table 12, and Table 16 provide the percentage of test items assessing each reporting category on the spring 2021 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%

Grade	Key Ideas and Details	Craft and Structure	Integration of Knowledge and Ideas	Language and Editing Task
9	15–25%	25–35%	20–30%	15–25%
10	15–25%	25–35%	20–30%	15–25%

Table 4: Observed Spring 2021 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	28%	40%	16%	16%
4	32%	32%	20%	16%
5	31%	31%	24%	14%
6	25%	40%	19%	15%
7	25%	38%	21%	15%
8	31%	38%	17%	13%
9	28%	37%	20%	15%
10	22%	37%	28%	13%

Table 5: Observed Spring 2021 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	25%	38%	21%	15%
8	31%	38%	17%	13%
9	28%	37%	20%	15%
10	20%	38%	25%	17%

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 2021 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 2021 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 Science

Grade	1*	2*	3*	4*
5	17%	29%	29%	25%
8	19%	27%	27%	27%

*See Table 13 for reporting category names.

Table 11: Observed Spring 2021 Percentage of Test Items Assessing Each Reporting Category in Science

Grade	1*	2*	3*	4*
5	18%	29%	29%	25%
8	20%	27%	27%	27%

*See Table 13 for reporting category names.

Table 12: Observed Spring 2021 Percentage of Test Items Assessing Each Reporting Category in Science—Accommodated Forms

Grade	1*	2*	3*	4*
5	18%	29%	29%	25%
8	20%	27%	27%	27%

Table 13: Reporting Categories Used in Science

Course	Reporting Category
5	Nature of Science Earth and Space Science Physical Science Life Science
8	Nature of Science Earth and Space Science Physical Science Life Science

Table 14: Blueprint Percentage of Test Items Assessing Each Reporting Category in EOC

Course	1*	2*	3*	4*
Algebra 1	41%	40%	19%	
Geometry	46%	38%	16%	
Biology 1	35%	25%	40%	
U.S. History	33%	34%	33%	
Civics	25%	25%	25%	25%

*See Table 17 for reporting category names.

Table 15: Observed Spring 2021 Percentage of Test Items Assessing Each Reporting Category in EOC

Course	Core Form	1*	2*	3*	4*
Algebra 1	Core 20	41%	40%	19%	
	Core 21	41%	40%	19%	
	Core 22	41%	40%	19%	
	Core 23	41%	40%	19%	
Geometry	Core 16	47%	38%	16%	
	Core 17	47%	38%	16%	
	Core 18	47%	38%	16%	
Biology 1	Core 100	36%	25%	39%	
	Core 200	36%	25%	39%	
	Core 300	36%	25%	39%	
	Core 400	36%	25%	39%	
U.S. History	Core 100	33%	35%	33%	
	Core 200	33%	35%	33%	
	Core 300	33%	35%	33%	
	Core 400	33%	35%	33%	
Civics	Core 100	25%	25%	25%	25%
	Core 200	25%	25%	25%	25%
	Core 300	25%	25%	25%	25%
	Core 400	23%	26%	26%	26%

*See Table 17 for reporting category names.

Table 16: Observed Spring 2021 Percentage of Test Items Assessing Each Reporting Category in EOC—Accommodated Forms

Course	1*	2*	3*	4*
Algebra 1	41%	40%	19%	
Geometry	47%	38%	16%	
Biology 1	36%	25%	39%	
U.S. History	33%	35%	33%	
Civics	25%	25%	25%	25%

*See Table 17 for reporting category names.

Table 17: 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
Biology 1	Molecular and Cellular Biology Classification, Heredity, and Evolution Organisms, Populations, and Ecosystems
U.S. History	Late Nineteenth and Early Twentieth Century, 1860–1910 Global Military, Political, and Economic Challenges, 1890–1940 The United States and the Defense of the International Peace, 1940–Present
Civics	Origins and Purposes of Law and Government Roles, Rights, and Responsibilities of Citizens Government Policies and Political Processes Organization and Function of Government

The summary tables show overall that the spring 2021 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, deviated from the blueprint by 7%.

In addition to information about reporting categories, the ELA Reading component, Mathematics, Science, and EOC blueprints also contained target information about DOK. DOK levels are used to measure the cognitive demand of instructional objectives and assessment items. The use of DOK levels to construct the Florida Statewide Assessments 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 Levels 1 and 2 elements; however, DOK Level 3 activity cannot be created with DOK Levels1 and 2 elements.

Table 18 shows the range of the percentage of items by DOK level by grade and subject or course. Table 19 shows the percentage of items from each DOK on the spring 2021 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 11% difference between the blueprint and the forms.

Table 18: 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%
Science 5 and 8	10-20%	60-80%	10-20%
Biology 1	10-20%	60-80%	10-20%
U.S. History	20-30%	45-65%	15-25%
Civics	15-25%	45-65%	15-25%

Table 19: Observed Spring 2021 Percentage of Items by Depth of Knowledge

Subject	Grade	DOK 1	DOK 2	DOK 3
Reading	3	18%	72%	10%
Reading	4	18%	58%	24%
Reading	5	14%	71%	14%
Reading	6	15%	56%	29%
Reading	7	15%	65%	19%
Reading	8	15%	62%	23%
Reading	9	19%	59%	22%
Reading	10	13%	56%	31%
Mathematics	3	20%	72%	7%
Mathematics	4	19%	67%	15%
Mathematics	5	13%	76%	11%
Mathematics	6	20%	68%	13%
Mathematics	7	14%	73%	13%
Mathematics	8	23%	66%	11%
Science	5	16%	71%	13%
Science	8	13%	75%	13%
Algebra 1	Core 20	19%	71%	10%

Subject	Grade	DOK 1	DOK 2	DOK 3
	Core 21	24%	66%	10%
	Core 22	19%	71%	10%
	Core 23	21%	71%	9%
Geometry	Core 16	19%	71%	10%
	Core 17	16%	74%	10%
	Core 18	19%	69%	12%
Biology 1	Core 100	14%	71%	14%
	Core 200	16%	70%	14%
	Core 300	14%	71%	14%
	Core 400	20%	64%	16%
U.S. History	Core 100	27%	52%	21%
	Core 200	29%	48%	23%
	Core 300	21%	54%	25%
	Core 400	23%	58%	19%
Civics	Core 100	27%	52%	21%
	Core 200	25%	50%	25%
	Core 300	25%	58%	17%
	Core 400	23%	55%	21%

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 20 provides target percentages of test passages assessing each type of text. Table 21 shows that across the grades, the percentage of informational and literary passages was close to the blueprint percentages. There was at most a 12% difference between the blueprint and the forms in grade 7 Reading.

Table 20: Blueprint Percentage of Reading Passage Types by Grade

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

Table 21: Observed Spring 2021 Percentage of Reading Passage Types by Grade

Grade	Informational	Literary
3	52%	48%
4	50%	50%
5	48%	52%

6	64%	36%
7	48%	52%
8	62%	38%
9	67%	33%
10	66%	34%

2.2 CONTENT-LEVEL AND PSYCHOMETRIC CONSIDERATIONS

In addition to test blueprints, several content-level and psychometric considerations were used in the development of the Florida Statewide Assessments. 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 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 2021 Florida Statewide Assessments tests were calibrated and equated to the IRT-calibrated item pool. More details about calibration, equating, and scoring can be found in Volume 1 of this technical report.

3. ITEM DEVELOPMENT PROCEDURES

The item development procedures employed by CAI for the FSA tests and the item development procedures employed by Pearson for Science and NGSSS EOC 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 Florida Statewide Assessments 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, & Thurlow (2002), were applied to the Florida Statewide Assessments development:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

CAI and Pearson 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.

The application of universal design (UD) principles as defined by Thompson, Johnstone, & Thurlow (2002) helps develop assessments that are usable to the greatest number of test takers, including Students with Disabilities (SWDs) and English language learners (ELLs).

As documented in this technical report, the item development procedures implemented for the Florida tests are consistent with industry practice. Specifically, Florida implements the UD principles throughout every stage of the assessment development process (i.e., initial design, item development, field testing, and implementation) to minimize the need for individual accommodations. As noted by Shaftel et al. (2015), under UD principles, accessibility is integral to the item development processes, thus minimizing access barriers associated with the tests themselves to the greatest extent possible for all students, including SWDs and ELLs.

Test development specialists receive extensive training in item development, including instruction on the UD principles and guidance on designing accessible content. Adherence to the UD principles is confirmed at every step of the review process so that the test maximizes readability, legibility, and compatibility with accommodations. Checklists that align to the Council of Chief State School Officers (CCSSO) Principles for High Quality Summative Assessment are used at each phase of the development cycle. As described in the Statewide Assessment Program Information Guide (FDOE, 2019), the processes of item development and test construction are carefully guided and include many quality control (QC) measures.

Examples of QC measures implemented to ensure that the constructs do not change from online to paper are presented here. Appendix K of Volume 2 provides details of test construction that includes roles and responsibilities of personnel involved in test item selection for test form assembly and QC measures to ensure the following:

- Item content on paper matches item content as administered online (e.g., wording, graphics, paragraph breaks, option order) via multiple rounds of content reviews.
- Items on both the accommodated forms and online forms are in the same order/locations.
- The student sees two-page items on an even then odd-numbered page simultaneously, just as they would see the entire item on one screen. Paper-appropriate language is used for directives on the paper accommodated forms.

In terms of software that supports the item development process, CAI'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), CAI was able to import items from multiple sources. To support online test delivery, ITS had a unique Web Preview feature that displayed items exactly as they were also presented to students, using the same program code used in CAI's Test Delivery System (TDS). An online test does not have a blueline (print approval) process like a PBT, and this feature provided an item-by-item blueline capability.

Prior to test administration, a series of user acceptance testing is performed on all approved platforms to ensure that items are rendered as expected and have similar appearance across platforms to minimize potential device effects.

Rigorous review is in place to ensure that the content of the item on paper matches the content of the item as administered online (e.g., wording, graphics, paragraph breaks, option order).

The next section describes the item sources for Florida Statewide Assessments, 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 2021 Florida Statewide Assessments came from multiple sources as outlined here.

New Items Written by CAI/Pearson

New field-test items were included in the spring 2021 forms, and these items will be used on future Florida Statewide Assessments test forms. The newly developed field-test items were written for the Florida-specific item bank (denoted as Florida Statewide Assessments item bank items). Mathematics and ELA items were written by CAI content experts or by trained partners. Pearson works with Florida educators to write new items in Science and Social Studies. All items undergo a rigorous process of preliminary, editorial, and senior review by CAI, Pearson, and FDOE's 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

Items field-tested in spring 2021 were developed in 2019 for an intended administration in spring 2020. However, because of the cancelation of the spring 2020 examinations due to the COVID-19 pandemic, the items were not field-tested until 2021.

Next Generation Sunshine State Standards (NGSSS) Assessment Items for FSA

In spring 2021, no NGSSS items were selected to be field-tested in FSA Mathematics. Some NGSSS items that aligned to the Florida Standards were used as core and anchor items in all grades. These items were previously field-tested on FSA tests.

3.2 ITEM TYPES

One of the important features of the online Florida Statewide Assessments 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 22, Table 23, and Table 24 list the Reading, Mathematics, Science 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 E for Reading, Appendix F for Mathematics and EOC, and Appendix G for Science and NGSSS EOC. Examples of various item types can be found in Appendix I.

Table 22: Reading Item Types and Descriptions

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from a number of options.

Response Type	Description
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 23: 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 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.

Table 24: NGSSS Science and EOC Item Type and Description

Response Type	Description
multiplechoice (MC)	Student selects one correct answer from four options.

3.3 COGNITIVE LABORATORIES

In a recent United States Department of Education (ED)-funded grant report investigating the accessibility of computerized assessments, Shaftel et al. (2015) point out that technology-enhanced (TE) items present greater accessibility barriers than traditional item types on PBTs, and that they should be examined to ensure that no construct-irrelevant variance is introduced. If some aspect of the technology impedes, or advantages, students in their responses to items, this could affect item responses and inferences regarding abilities on the measured construct.

Florida assessments are delivered by the same test delivery system as Smarter Balanced Assessment Consortium (SBAC), therefore, research evidence on the SBAC platform can also be generalizable to Florida assessments. Two types of research were conducted for SBAC: (1) usability studies on system tools and features; and (2) cognitive lab studies evaluating validity of various item types. Findings show that (1) various aspects of the test delivery system (e.g., tools, navigation, directions) provide students equitable access to the assessed content; and (2) TE item types do not introduce construct-irrelevant variance into scores. The full research report is provided in Volume 7 of the *2014–2015 Florida Standards Assessments Technical Report*, which was included in an earlier submission for peer review. In addition, cognitive labs on Florida items were planned to be carried out in spring 2020, but they were canceled due to the COVID-19 pandemic. The goal of the study is to evaluate whether Florida items provide valid measures of students' mastery of the intended constructs. This study will be conducted in spring 2021.

3.4 ITEM TRANSLATIONS TO BRAILLE FORMAT

As is noted in Allman (2006), it is common that portions of a test may need to be modified in order to be translatable to braille format. Modifications may include substituting word, reformatting the layout of the item, and replacing untranslatable items with others of equal weight, content, and difficulty. As Winter (2010) acknowledges, this can pose a challenge to comparability, but this accommodation is needed for students with disabilities to properly demonstrate the knowledge, skills, and abilities the construct represents.

Florida uses a rigorous process outlined in the *Florida Statewide Assessments Production Specifications* in creating the braille translations of the summative tests and works with the Florida Instructional Materials Center for the Visually Impaired (FIMC-VI) and American Printing House (APH) who are both leaders in the industry to create those translations. Both FIMC-VI and APH follow practices determined by the Braille Authority of North America (BANA).

When forms are translated into braille, our contractors ensure that the braille forms match the regular print forms and make exceptions only when modifications for the braille reader are necessary. For instance, sometimes the item directions need to be modified for the braille reader instructing them to write in the letter instead of filling in the bubble. We also provide both UEB-Nemeth and Unified English Braille (UEB)-Technical versions of Mathematics and Science tests, and for all tests we provide both contracted and uncontracted versions to ensure that visually

impaired students have the type of braille they read available. This means that in some cases, four braille transcriptions are made for each grade and subject: *UEB-Nemeth Uncontracted, UEB-Nemeth Contracted, UEB-Technical Uncontracted, UEB-Technical Contracted*. We ensure that the students who read braille are tested and challenged at the same level as their sighted peers. By working with FIMC-VI and APH, we ensure that all tests are reviewed and proofread by certified braille transcribers/proofreaders and teachers of the visually impaired that have vast experience and knowledge regarding students in this demographic. If modifications are made, a subject content specialist must approve any suggestions made by FIMC-VI and APH. Our content team ensures that the information vital to the item is retained in the braille format and that the student who reads braille is not given either an advantage or disadvantage.

When transcribing pictures, cartoons, and graphics, images are either described or made in a tactual format for the braille reader, or with permission from content specialists, are sometimes omitted from the test if they do not provide any additional information. If graphics are described, we often use the descriptions already created for text-to-speech, which all students have access to. If tactile graphics are created, they are kept as true to the original as possible. When deviation is needed, we comply with best practices in the field. Examples are as follows:

- Extraneous details such as decorative pictures, icons, or sections of a map that are not needed for the item are sometimes omitted—as the amount of information that can be interpreted through fingers is less than the amount of information the eye can process.
- Occasionally, especially with three-dimensional figures represented as two-dimensional drawings, graphics are too complex to be created tactually and description alone either would not provide enough information or would give away the answer. In situations such as this, we develop manipulatives of the three-dimensional figures with specific directions to the Test Administrator on how to present them.

3.5 Development and Review Process for New Items

3.5.1 Development of New Items

CAI and Pearson developed field-test items to be embedded in the Florida Statewide Assessments 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.

CAI and Pearson 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 either CAI or Pearson item-writing workshops or had previous training on writing MC and CR items. CAI and Pearson content area assessment specialists worked with TDC 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 DOK 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 (e.g., 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 25 provides the guidelines used in FSA Reading.

Table 25: 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 TDC. If needed, CAI, Pearson, and TDC content staff discussed requested revisions, ensuring that all items appropriately measured the Florida Standards. The items were then revised by CAI and Pearson 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 the 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 Committee, the Community Sensitivity Committee,

and the Content Item Review Committee. These committee reviews occurred in 2019 for items field tested in spring 2021.

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.

The TDC, CAI, and Pearson 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.

3.5.2 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 CR 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. The meetings occurred in June 2021 in a hybrid on-site (TDC and educators) and virtual (CAI staff) manner.

Similar to the items field-tested in previous years, 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 rubric validation meeting, CAI 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, CAI staff generated frequency tables that contained all student responses for each item. Frequency tables were either generated out of CAI's Database of Record (DOR) for computer-based grades, or CAI's Key Verification System (KVS) for paper-based grades. Although sourced from different databases, the frequency tables included the same information regardless of origin.

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 items with simple rubrics, the committee members were shown each item, along with the correct response and the most frequently selected incorrect responses. TDC and CAI staff ensured that the committee was scoring consistently. The committee meetings used the following procedures:

- All 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.
- CAI rescored the responses using the revised rubric.
- CAI reviewed the responses that received changed scores to determine if they were correctly scored.
- The TDC reviewed the rescored responses and approved the rubric.

If any scores changed based on the Rubric Validation Committee review, CAI staff revised the machine rubric and rescored the item. After the item was rescored, CAI staff reviewed at least 10% of responses for which the score changed. This review ensured that committee suggestions were

honored, that the item was scored consistently, and that no unintended changes in scoring occurred because of the revision to the machine rubric. CAI 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 operational test forms. At the end of the testing window, CAI conducted classical item analysis on these field-test items to ensure that the items functioned as intended with respect to the underlying scales. CAI's analysis program computed the required item and test statistics for each MC and CR 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.

3.6 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, DOK levels, and numbers in each reporting category.

In spring 2021, field-test items were embedded in online forms in grades 7 through 10 Reading, grades 7 through 8 Mathematics, and EOC. They were administered on paper in grades 3 through 6 for Reading, Mathematics, and Science. All assessments were fixed-form tests with a predetermined number and location of field-test items. Table 26, Table 27, and Table 28 provide the number of field-test items by type for Reading, Mathematics, Science, and EOC.

Table 26: Number of Reading Field-Test Items by Type

Item Type	3	4	5	6	7	8	9	10
EBSR	13	9	11	20	19	16	20	1
HT	5	2	5	3	3	3	7	3
MC	89	87	81	75	76	85	79	99
MI	8	6	9	6	6	4	4	4
MS	10	13	8	15	14	10	7	10
Two-Part HT	0	0	2	0	0	1	1	0

Table 27: Number of Mathematics and EOC Field-Test Items by Type

Item Type	3	4	5	6	7	8	Algebra 1	Geometry
EQ	45	32	36	26	25	32	5	8
ETC	5	5	6	9	10	13	43	45
GI	0	0	0	0	1	2	6	6
HT	0	0	0	0	0	0	2	0
MC	66	77	64	52	36	64	64	59
MI	9	10	7	7	4	8	0	3

MS	17	29	18	21	13	13	7	11
Multi	10	6	15	5	7	10	10	19

Table 28: Number of NGSSS Science and EOC Field-Test Items by Type

Item Type	5	8	Biology 1	U.S. History	Civics
MC	331	306	333	214	152

3.7 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 Florida Standards Assessments Annual Technical Report*.

4. TEST CONSTRUCTION

4.1 OVERVIEW

The test forms administered in spring 2021 are generated from the test construction activities that happened in summer 2019 and summer 2020. During summer 2019, psychometricians and content experts from FDOE, the TDC, and CAI convened in person for two weeks, and FDOE, the TDC, and Pearson convened in person for one week, to build forms for spring 2020. When the spring 2020 administration was cancelled due to the COVID-19 pandemic, it was decided to reuse the forms built for 2020 in 2021, with some exceptions. Tests that were scheduled for release in grade 3 Mathematics and ELA, grade 10 ELA, and Algebra 1, as well as tests in Civics which needed to be rebuilt due to COVID-19-related sensitivity concerns, were built during one-week virtual test construction meetings held in summer 2020. In both instances, FDOE Florida Statewide Assessments test construction used 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 K, 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. CAI/Pearson 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. CAI/Pearson content staff consult CAI/Pearson psychometricians to ensure that the form meets the psychometric considerations. The forms are then submitted to TDC content specialists for review. Both TDC and CAI/Pearson 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 CAI/Pearson psychometric team and then to the FDOE psychometric team.

3. Both the CAI/Pearson 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 (CSEM) 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 CAI/Pearson and FDOE content teams for further revision.

4.1.1 Roles and Responsibilities of Participants

CAI/Pearson Content Team

CAI/Pearson ELA, Mathematics, Science, and Social Studies 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. CAI/Pearson 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 CAI/Pearson content staff
- Revision of the anchor and operational item sets according to feedback from CAI/Pearson psychometricians
- Assistance in the generation of materials for TDC and FDOE review
- Revision of the forms to incorporate feedback from the TDC and FDOE

CAI/Pearson Technical Team

The CAI/Pearson technical team, which included psychometricians and statistical support associates, prepared the item bank by updating ITS with current item statistics and provided test construction training to the internal content team. During on-site 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 CAI'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 CAI/Pearson content specialists to revise forms and select replacement items. Both parties selected items with respect to the statistical guidelines and the Florida Statewide Assessments 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 CAI/Pearson 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 H. The proposed forms were either returned to TDC and CAI/Pearson 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 CAI/Pearson's content team for further revision.

4.2 TEST CONSTRUCTION PROCESS

The test construction process for spring 2021 occurred in both summer 2019 (on-site) and summer 2020 (virtual). During summer 2019 psychometricians and content experts from FDOE, the TDC, CAI, and Pearson convened to build forms for spring 2020. When the spring 2020 administration was cancelled due to the COVID-19 pandemic, it was decided to reuse the forms built for 2020 in 2021, with some exceptions. Tests that were scheduled for release in grade 3 Mathematics and ELA, grade 10 ELA, and Algebra 1, as well as tests in Civics which needed to be rebuilt due to COVID-19-related sensitivity concerns, were built during one-week virtual test construction meetings held in summer 2020. In both instances, Florida test construction used a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

The Florida Statewide Assessments 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 Off-Site Test Construction

Once item calibrations were complete, CAI/Pearson's technical team updated the item bank with all possible items for test construction. CAI's ITS was updated with the most current item statistics for any given item. MDSs were also created to assist the content teams at CAI/Pearson and the 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 IRT statistics. Items that have been administered multiple times have multiple listings in the MDS.

CAI/Pearson'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 CAI/Pearson's technical team for review. CAI/Pearson 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 CAI's or Pearson'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 the 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 the TDC. This allowed for review before onsite face-to-face meetings.

4.2.2 On-Site/Virtual Meetings

Due to the ongoing COVID-19 pandemic restrictions, the Test Construction meetings for the spring 2021 administration of Florida's K-12 Statewide Student Assessment Program was held virtually rather than face-to-face, as had been done in previous years. This necessity provided challenges as well as opportunities to use new processes. Fortunately, because all grades and courses except those originally scheduled for release in 2021 or where sensitivity issues relating to the COVID-19 pandemic required adjustments, used the forms developed for the 2020 spring administration, the 2021 Test Construction was smaller in scope compared with previous years. That made Virtual Test Construction more manageable than it might have been in a normal year. Although the processes for Test Construction were slightly different in terms of logistics, the end products were the same as previous administrations. All Test Construction resources (e.g., Verification Checklists/Logs, form-building spreadsheets, statistical worksheets, item cards, item PDF files, form-tracking logs and posters) were delivered digitally.

All parties, including program management, were actively involved in Virtual Test Construction. On the morning of the first day, a commencement meeting was held online to introduce all team members, explain any changes to test specifications or blueprints, discuss proposed forms, and prioritize upcoming activities. ELA, Mathematics, Science, and Social Studies content specialists proceeded to their respective rooms to discuss proposed forms. For each grade and subject, there was at least one CAI or Pearson content specialist and one TDC content specialist present for deliberations; at least one CAI or Pearson 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 2021 test forms, anchor items were selected from previous anchor, core, or Florida field-tested items. If content experts had questions about item statistics, psychometricians were available to provide clarification.

Although the transition from CBTs to PBTs for grades 4–6 ELA and grades 3–6 Mathematics occurred in spring 2019, content experts from CAI and the TDC continued the process of creating a list of possible mode effect "watch items" for anchor items with most recent statistics from online administrations if needed. 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 CAI/Pearson 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, DIF statistics, 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 H. 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.

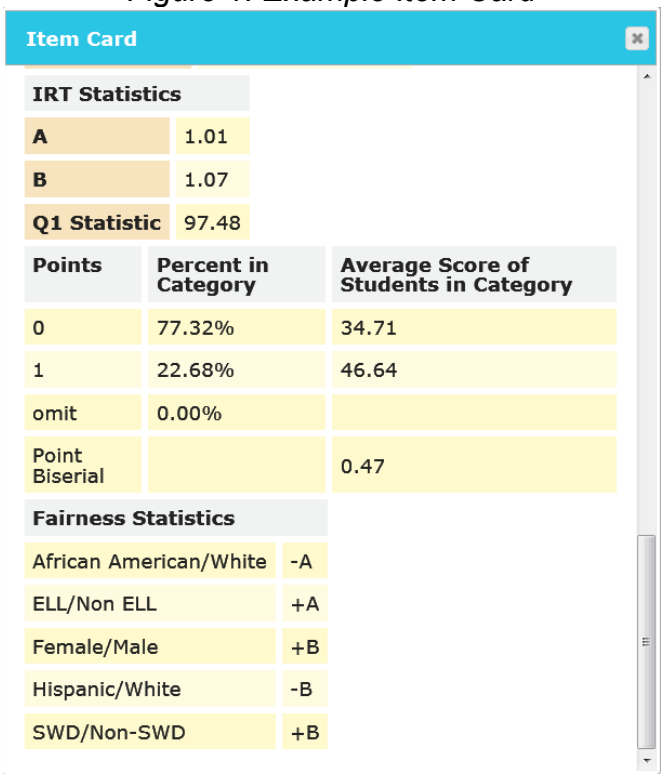
Once an anchor item set was approved, the same process was used to select operational items. Once both anchor item sets and operational items were approved, forms were entered into ITS, where they were evaluated for a final time to confirm that the intended items were placed on the individual forms.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



Item Card

IRT Statistics	
A	1.01
B	1.07
Q1 Statistic	97.48

Points	Percent in Category	Average Score of Students in Category
0	77.32%	34.71
1	22.68%	46.64
omit	0.00%	
Point Biserial		0.47

Fairness Statistics	
African American/White	-A
ELL/Non ELL	+A
Female/Male	+B
Hispanic/White	-B
SWD/Non-SWD	+B

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., MC)
- Point value
- Answer key
- Reporting category
- DOK

Please note that bookmaps cannot be generated until after the anchor and core forms are finalized and put into ITS. Bookmaps are used as an accessible resource to both content specialists and psychometricians to find information about a test form. Bookmaps differ from item cards in that there are no statistical summaries in a bookmap. Bookmaps provide useful information regarding the forms that are built in ITS.

4.3.3 Graphical Summaries

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

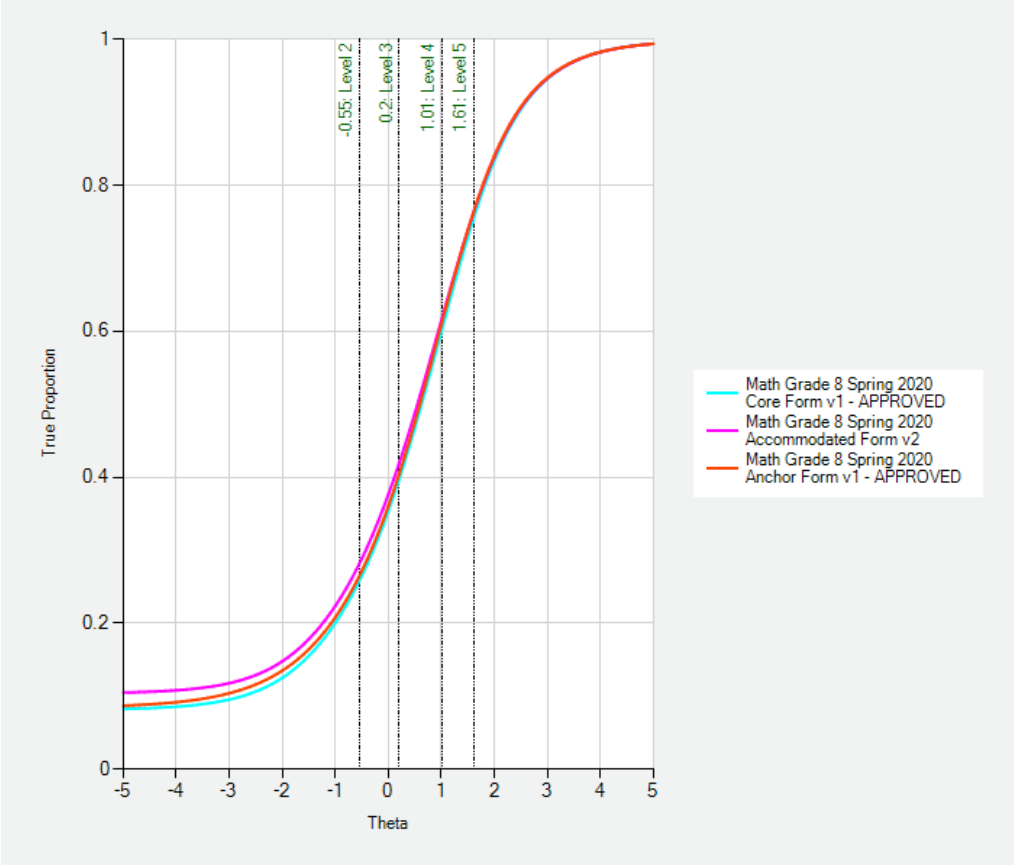
Test Characteristic Curve

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 using TCCs.

The spring 2019 core form TCCs were the target for the spring 2021 forms. The spring 2021 online TCC was used as a target while building the spring 2021 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

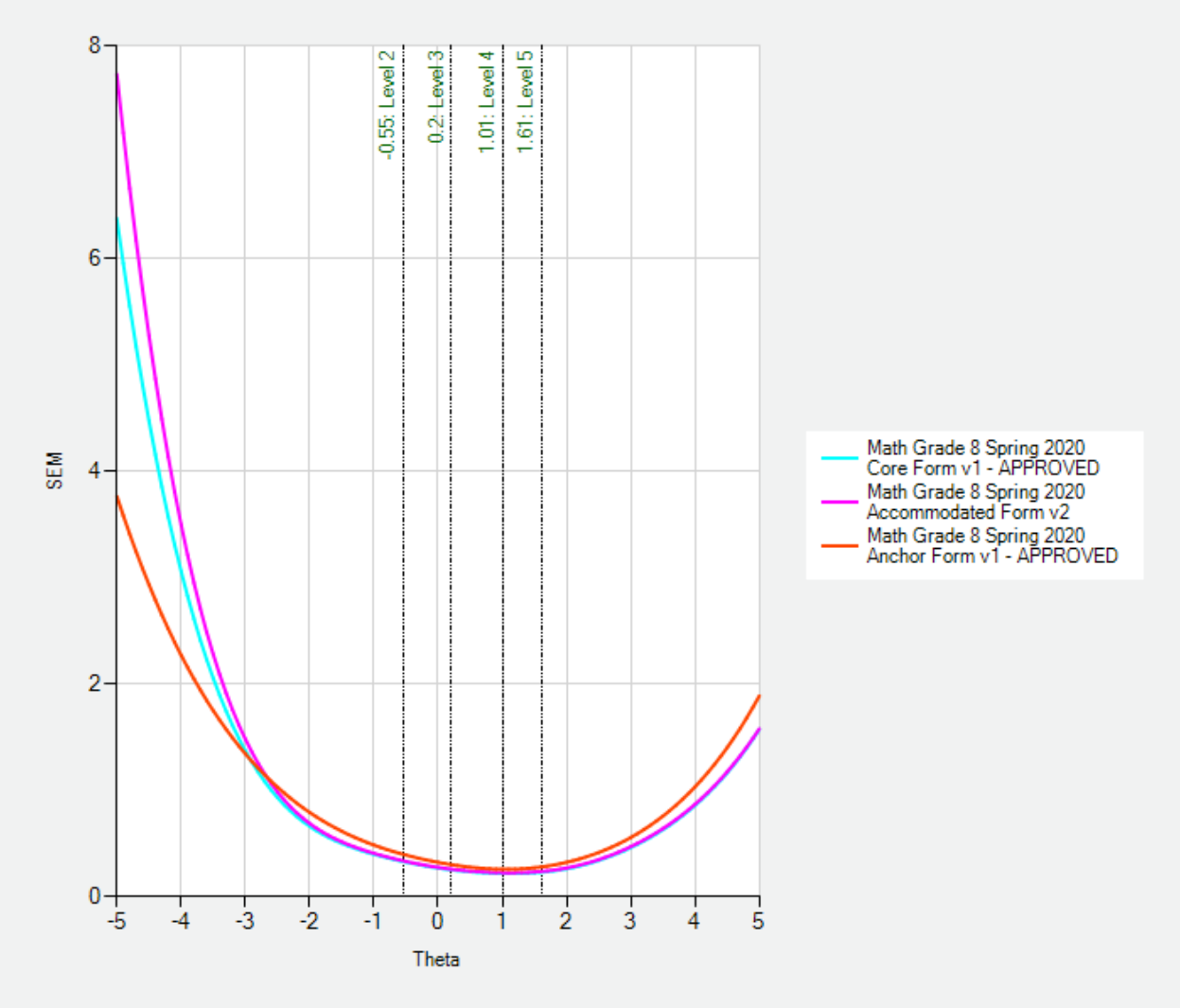


Conditional Standard Error of Measurement Curve

The 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 2019 core form CSEMs were the target for the spring 2021 forms. The spring 2021 online CSEM was used as a target while building the spring 2021 paper-pencil accommodated 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 tests and PBTs.

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 MC 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 29 provides the number of items replaced between the online and paper-pencil accommodated forms.

Table 29: Number of Item Replacements for Paper-Pencil Accommodated Forms

Test	Number of Items Replaced
Grade 7 Mathematics	8
Grade 8 Mathematics	9
Algebra 1	11
Geometry	5

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 30. 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 30: Test Summary Comparison for Grade 8 Mathematics Online and Paper-Pencil Forms

Type	Statistics	Spring 2021 Core Form	Spring 2021 Accommodated
Overall	Number of Items	56	56
	Possible Score	56	56
	Difficulty Mean	0.68	0.66
	Difficulty Standard Deviation	0.77	0.79
	Difficulty Minimum	-1.41	-1.41
	Difficulty Maximum	1.86	1.86
	Parameter-A Mean	0.85	0.84
	Parameter-A Standard Deviation	0.22	0.22
	Parameter-A Minimum	0.49	0.49
	Parameter-A Maximum	1.58	1.58
	Parameter-C Mean	0.12	0.13
	Parameter-C Standard Deviation	0.10	0.10
	Parameter-C Minimum	0.00	0.01
	Parameter-C Maximum	0.34	0.34
	Raw Score Sum	20.83	21.99
	p-Value Mean	0.37	0.39
	p-Value Standard Deviation	0.18	0.18
	p-Value Minimum	0.14	0.14
	p-Value Maximum	0.85	0.85
	Lowest Biserial/Polyserial	0.31	0.29

Using Grade 8 Mathematics as an example to provide a statistical comparison for online and accommodated forms on some essential form-level statistical properties, Section 4.3 further evaluates the parallelism between the online and accommodated forms by comparing TCCs and CSEMs. In both TCC and CSEM plots, the curves for online and accommodated forms are superimposed with each other demonstrating the degree to which two forms are statistically parallel.

REFERENCES

Allman, C. (2006). Position paper: *Accommodations for testing students with visual impairments.* Louisville, KY: American Printing House for the Blind. Retrieved from

Shaftel, J., Benz, S., Boeth, E., Gahm, J., He, D, Loughran, J., Mellen, M. Meyer, E., Minor, E., & Overland, E. (2015). *Accessibility for Technology-Enhanced Assessments (ATEA) Report of Project Activities*. Research Report. University of Kansas.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (NCEO Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved 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.

Winter, P. (2010). Comparability and test variations. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations (pp. 1-11).* Washington, DC: Council of Chief State School Officers.