State of Florida

Florida Assessment of Student Thinking (FAST), Benchmarks for Excellent Student Thinking (B.E.S.T.), and Science & Social Studies Statewide Assessments Technical Report

2024–2025

Volume 2
Test Development

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 FDOE (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Dr. Myvan Bui, Dr. Sherry Li, Matt Gordon, Zoe Dai, Oliver Brown, Daniel Lam, and Dr. Yuan Hong. The major contributors from FDOE are as follows: Dr. Salih Binici, Susie Lee, Catherine Altmaier, Racquel Harrell, Sally Donnelly, Dr. Shakia Johnson, Dr. Stacy Skinner, Leah Glass, Kristina Lamb, Dr. Wenyi Li, Jieling Ming, Saeyan Yun, and Yiting Yao.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

By statute, all Florida public school students are required to participate in statewide assessments. Beginning with the 2022–2023 school year, Florida's statewide, standardized assessments in English language arts (ELA) reading, ELA writing, and mathematics were aligned with the Benchmarks for Excellent Student Thinking (B.E.S.T.). Social studies assessments are aligned to Florida's State Academic Standards. In December 2008, the State Board of Education adopted the Next Generation Sunshine State Standards (NGSSS) for social studies. These standards were used to develop the U.S. History end-of-course (EOC) assessment, which continues to assess the NGSSS. The grade 7 Civics and U.S. Government strands of the NGSSS informed the development of the initial Civics EOC assessment. In July 2021, the State Board of Education adopted new Civics and U.S. Government standards. Beginning with the spring 2024 test administration, the Civics EOC assessment reflects these updated standards.

The State of Florida implemented new online computer-adaptive tests (CATs) for operational use beginning with the 2022–2023 school year for ELA writing and mathematics and in 2023–2024 for science and social studies. Before this, they were fixed-form online and paper-based tests.

The assessment program for ELA and mathematics, referred to as the Florida Assessment of Student Thinking (FAST), replaced the Florida Standards Assessments (FSA) in ELA reading and mathematics. FAST consists of progress monitoring (PM) assessments administered three times a year. Grades 4–10 writing and mathematics EOC Algebra 1 and geometry are considered B.E.S.T. assessments and are not part of the PM FAST assessments. FAST and B.E.S.T. were first administered to students in fall 2022. Writing was administered in spring 2023 as a standalone field test administered to a representative sample of Florida students. Beginning with the 2023–2024 school year, writing will be administered during each spring test administration. During spring testing windows, all CATs are given to students as summative assessments.

Beginning spring 2024, the Summative Statewide Science Assessment in grades 5 and 8, as well as the Biology 1, Civics, and U.S. History EOC assessments, were delivered in a computer-adaptive format, which allows for immediate reporting. While the core content for these tests did not change, some administration details (e.g., reduced test length) and blueprint specifications (e.g., number of items each student will receive) have been updated.

The science and social studies assessments include multiple-choice (MC) items, and the ELA and mathematics assessments include (but are not limited to) the use of several technology-enhanced item (TEI) types. For all online assessments, accommodated versions are available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicate such a need. IEPs or Section 504 Plans indicating the need for versions other than online are addressed accordingly.

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

- Test item development process employed was consistent with industry standards.
- The development and maintenance of the item pool plan established an item bank in which test items cover the range of measured standards, grade-level difficulties, and cognitive

complexity using both selected-response keyed items and constructed-response machine-scored or handscored item types.

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

2. TEST SPECIFICATIONS

The test design summary and blueprint stipulated the range of operational items from each reporting category required on each form. These documents guided item selection and test construction for the assessments. The item specifications provided detailed guidance for item writers and reviewers to ensure that items were aligned to the standards they were intended to measure. The Florida Department of Education (FDOE) and committees of experienced Florida educators developed and approved the specifications that define the content and format of the tests and test items.

Following the adoption and integration of Florida's State Academic Standards and Benchmarks for Excellent Student Thinking (B.E.S.T.) Standards into the school curriculum, items and test item specifications were developed to ensure that the tests and their items were aligned to the benchmarks and grade-level expectations that they were intended to measure. FDOE and content specialists developed test item specifications. Educator committees also reviewed and provided feedback for the initial development of B.E.S.T. test item specifications.

The assessments are based on their relevant standards, course descriptions, and test item specifications. Florida Assessment of Student Thinking (FAST) and B.E.S.T. test item specifications are based on the B.E.S.T. Standards, the state's science assessments are based on Florida's science standards, and U.S. History and Civics are based on their respective standards. In July 2021, the new State Academic Standards for Civics were adopted. Beginning with the spring 2024 test administration, these standards were assessed by the Civics end-of-course (EOC) assessment. The specifications define 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 Florida's state academic standards and informs all stakeholders about the scope and function of the assessments. In addition to these general guidelines, specifications for FAST English language arts (ELA) reading and ELA writing also include guidelines for developing reading and writing passages and prompts, such as length, type, and complexity.

2.1 DEVELOPMENTS

A blueprint for each assessment specifies the reporting categories with respective benchmarks and percentages. A test design summary specifies the blueprint, duration of test (time), and length of test (item range). The blueprint construction for the assessments are evidenced by test design summary documents found at https://www.fldoe.org/accountability/assessments/k-12-student-assessment. These documents were created using Florida's course descriptions as the basis for the design. Course descriptions can be found on the CPALMS website at http://www.cpalms.org/Public/search/Course.

After switching ELA reading and mathematics to computer-adaptive tests (CATs), Content Advisory Committee (CAC) meetings were held with educators to propose and approve the revised blueprints. An in-person CAC meeting was held to discuss blueprints for ELA reading and mathematics for grades 3–8 in April 2022. There was a virtual CAC meeting for mathematics and EOC assessments in September 2022. In December 2022, a virtual meeting was held for both subjects to discuss shortening the blueprints further, after they had been shortened during the move from Florida Standards Assessments (FSA) to B.E.S.T. (although the decision was to not do so).

The blueprints were also discussed at Technical Advisory Committee (TAC) meetings in the summer and November 2022. Reporting categories for ELA reading were derived from the applicable "cluster" naming convention in the Florida B.E.S.T. Standards, and percentages of the reporting categories within the tests were derived from the number, complexity, and breadth of standards to be assessed. Vocabulary standards were folded in with the Reading Across Genres Standards to create the Reading Across Genres & Vocabulary reporting category. Guidelines for the weight of each reporting category for FAST ELA reading were determined by Florida's TAC. To avoid "statistical noise" generated from items scored in a small reporting category, the TAC advised FDOE that 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 B.E.S.T. Standards. As with ELA reading, if a mathematics domain has 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 benchmark information provides benchmark clarification statements, assessment limits, stimulus attributes, response attributes, prior knowledge, and a sample item for each benchmark that could be assessed. Test item specifications for ELA and mathematics were revised in 2021 and 2022 after adoption of the B.E.S.T. Standards and the decision to move to CATs. The test design summaries for both mathematics and ELA reading were updated during the 2022–2023 school year to represent the reduced test length and new reporting categories. The design summary now states specifically that the ELA reading and ELA writing components are tested and reported separately.

The social studies test design summaries were updated during the 2023–2024 school year. Both Civics and U.S. History summaries reflect revised item ranges in the length of test section to accommodate CATs. Additionally, the Civics summary was updated to integrate new benchmarks into existing reporting categories.

The Civics EOC Assessment Test Item Specifications were revised in 2023 to align with the Civics and U.S. Government standards approved by the State Board of Education in 2021. The revised summary retained the existing reporting categories, incorporated new benchmarks within those categories, and updated the item range in the length of test section to reflect CATs.

The U.S. History EOC Assessment Test Item Specifications (2010) have not been revised; and they include the assessment's initial test design summary. The updated test design summary is available only at https://www.fldoe.org/accountability/assessments/k-12-student-assessment/end-of-course-eoc-assessments/index.stml. At this time, test item specification updates may occur only after the State Board of Education adopts new or revised standards and attaches them to course descriptions. Then, the Office of Assessment initiates a comprehensive process to develop interpretive materials, including test item specifications, which contain test design summary, reporting category statements, and Achievement-Level Descriptors.

All test item specifications for all grades and subjects are always marked as drafts to emphasize that they are living documents. Most often, changes are minor and do not result in significant changes to test design, development, or content.

Detailed descriptions for the constructs of the reporting categories are presented in Appendices A1–A4 (Science Reporting Categories Descriptors, ELA Reporting Categories Descriptors, Science and Social Studies EOC Reporting Categories Descriptors, and Mathematics and EOC Reporting Categories Descriptors) for all assessments.

2.1.1 Target Blueprints

Test blueprints provided the following guidelines:

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

This section provides a summary of the blueprints. Detailed blueprints for each content level are presented in Appendices B1–B5 (ELA Blueprints, Mathematics and EOC Blueprints, U.S. History Blueprints, Civics Blueprints, and Biology and Science Blueprints) for all subjects.

The following assessments are administered as CATs: grades 3–10 ELA reading, grades 3–8 mathematics, mathematics EOC assessments (Algebra 1 and geometry), grades 5 and 8 science, and EOC assessments in science and social studies (Biology 1, U.S. History, and Civics). Additionally, ELA writing is administered online for grades 4–10. In spring 2023, typed written response accommodations were provided for students taking ELA writing assessments; therefore, responses from these students were collected online. For grades and subjects testing online, accommodations are provided if indicated by a student's Individualized Education Plan (IEP) or Section 504 Plan.

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. The exception is for writing, where field testing for the current items in rotation was conducted in 2023. Table 2 displays the number of operational and field-test items available in the spring 2025 item pool. ELA writing items are not included in the item counts listed for ELA reading tests.

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

Subject/Course	Grade	Total Number of Items
ELA Reading	3	36–40
	4	36–40
	5	36–40
	6	36–40
	7	36–40
	8	36–40
	9	36–40
	10	36–40
Mathematics	3	35
	4	35

Subject/Course	Grade	Total Number of Items
	5	35
	6	36
	7	36
	8	36
Algebra 1	9	45
Geometry	9	45
Biology 1	10	45
Grade 5 Science	5	45
Grade 8 Science	8	45
Civics	7	44
U.S. History	9	46

Table 2: Number of Items Available in Spring 2025 Item Pool, by Grade and Subject or Course

Subject/Course	Grade	Number of Operational Items	Total Item Counts in the Field-Test Pool	Total Item Counts in the Field-Test and Operational Pool
ELA Reading	3	650	180	830
	4	661	150	811
	5	668	101	769
	6	511	215	726
	7	611	181	792
	8	606	223	829
	9	641	192	833
	10	614	158	772
Mathematics	3	637	147	784
	4	671	149	820
	5	673	152	825
	6	633	150	783
	7	611	180	791
	8	624	198	822
Algebra 1		528	172	700
Geometry		510	169	679
Biology 1		996	236	1,232
Science	5	1,107	190	1,297
	8	872	177	1,049
Civics		809	311	1,120

Subject/Course	Grade	Number of Operational Items	Total Item Counts in the Field-Test Pool	Total Item Counts in the Field-Test and Operational Pool
U.S. History		817	300	1,117

Reporting categories were used to define the topics assessed in each content area more narrowly. Individual scores on reporting categories provide information to help identify areas in which a student may have had difficulty. Table 3 and Table 4 provide the percentages 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 practical construction. The ELA writing component prompt was not included in these blueprints. Table 5 provides the reporting categories for mathematics grades 3–8, Table 6 provides the percentage of operational items required in the blueprints by content strands, or reporting categories, for mathematics EOC, and Table 7 provides the reporting categories for mathematics EOC. Table 8 provides the percentage of operational items required in the blueprints by content strands, or reporting categories, for science and social studies. Table 9 provides the reporting categories for science and social studies.

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

Grade	Reading Prose and Poetry	Reading Informational Text	Reading Across Genres and Vocabulary
3	25%–35%	25%–35%	35%–50%
4	25%–35%	25%–35%	35%–50%
5	25%–35%	25%–35%	35%–50%
6	25%–35%	25%–35%	35%–50%
7	25%–35%	25%–35%	35%–50%
8	25%–35%	25%–35%	35%–50%
9	25%–35%	25%–35%	35%–50%
10	25%–35%	25%–35%	35%–50%

Table 4: Blueprint Percentage of Test Items Assessing Each Reporting Category, Mathematics

Grade	1*	2*	3*	4*
3	23%–29%	23%–29%	23%–29%	23%–29%
4	31%–37%	31%–37%	31%–37%	
5	23%–29%	23%–29%	23%–29%	23%–29%
6	33%–42%	25%–36%	25%–36%	
7	25%–31%	22%–31%	22%–28%	22%–28%
8	22%–28%	22%–28%	25%–31%	22%–28%

*Refer to Table 5 for reporting category names.

Table 5: Reporting Categories Used in Mathematics

Grade	Reporting Category
3	Number Sense and Additive Reasoning Number Sense and Multiplicative Reasoning Fractional Reasoning Geometric Reasoning, Measurement, and Data Analysis and Probability
4	Number Sense and Operations with Whole Numbers Number Sense and Operations with Fractions and Decimals Geometric Reasoning, Measurement, and Data Analysis and Probability
5	Number Sense and Operations with Whole Numbers Number Sense and Operations with Fractions and Decimals Algebraic Reasoning Geometric Reasoning, Measurement, and Data Analysis and Probability
6	Number Sense and Operations Algebraic Reasoning Geometric Reasoning, Data Analysis, and Probability
7	Number Sense and Operations and Algebraic Reasoning Proportional Reasoning and Relationships Geometric Reasoning Data Analysis and Probability
8	Number Sense and Operations and Probability Algebraic Reasoning Linear Relationships, Data Analysis, and Functions Geometric Reasoning

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

Course	1*	2*	3*
Algebra 1	31%–38%	31%–38%	31%–38%
Geometry	33%–40%	27%–33%	33%–40%

*Refer to Table 7 for reporting category names.

Table 7: EOC Reporting Categories

Course	Reporting Category
Algebra 1	Expressions, Functions, and Data Analysis Linear Relationships Non-Linear Relationships
Geometry	Logic, Relationships, and Theorems Congruence, Similarity, and Constructions Measurement and Coordinate Geometry

Table 8: Blueprint Percentage of Items Assessing Each Reporting Category, Science and Social Studies

Grade/Course	1*	2*	3*	4*
Biology 1	35%	25%	40%	
Grade 5 Science	17%	29%	29%	25%
Grade 8 Science	19%	27%	27%	27%
Civics	25%–30%	15%–20%	20%–25%	20%–25%
U.S. History	33%	34%	33%	

*Refer to Table 9 for reporting category names.

Table 9: Reporting Categories Used in Science and Social Studies

Grade/Course	Reporting Category
Biology 1	Molecular and Cellular Biology Classification, Heredity, and Evolution Organisms, Populations, and Ecosystems
Grade 5 Science	Nature of Science Earth and Space Sciences Physical Sciences Life Sciences
Grade 8 Science	Nature of Science Earth and Space Sciences Physical Sciences Life Sciences
Civics	Origins and Purposes of Law and Government Roles, Rights, and Responsibilities of Citizens Government Policies and Political Processes Organization and Function of Government
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

The FAST ELA reading blueprint also included specifications for genres of text presented in the passages. Two main types of text were used: literary and informational. Table 10 provides target percentages of the test passages assessing each type of text.

Table 10: Blueprint Percentage of Reading Passage Types by Grade

Grades	Informational	Literary
3–5	50%	50%
6–8	50%	50%
9–10	50%	50%

2.1.2 Cognitive Complexity

Cognitive complexity refers to the cognitive demand associated with a test item. The cognitive classification system implemented by FDOE is based on Dr. Norman L. Webb's Depth of Knowledge (DOK) levels (Webb, 2002). The rationale for classifying a test item by its cognitive complexity focuses on expectations made of the test item, not on the ability of the student. When classifying a test item's demands on thinking (i.e., what the test item requires the student to recall, understand, analyze, and do), it is assumed that the student is familiar with the basic concepts of the task. Test items are chosen for the Statewide Science Assessment based on the standards and grade-level appropriateness, but the complexity of test items remains independent of the curriculum a student has experienced. On any given assessment, the cognitive complexity of a multiple-choice (MC) item may be affected by the distractors (answer options). The cognitive complexity of a test item depends on the grade level of the assessment; a test item that has a high level of cognitive complexity at one grade may not be as complex at a higher grade. The complexity categories—low, moderate, and high—form an ordered description of the demands a test item may make on a student. For example, low-complexity test items may require a student to solve a one-step problem. Moderate-complexity test items may require multiple steps; however, the number of steps is not always indicative of cognitive level. High-complexity test items may require a student to analyze and synthesize information. The distinctions made in terms of complexity ensure that test items will assess the depth of student knowledge at each benchmark. The intent of the item writer weighs heavily in determining the complexity of a test item.

Cognitive complexity is addressed at the item bank level through deliberate item development, including item writer training on the DOK levels, ensuring sufficient representation across complexity categories. Item development plans target a variety of intended cognitive complexities to explore the different components of the benchmarks and the rigor variety they offer within a grade level. It is also enhanced during item review committees, where items are updated as needed based on the feedback. Additionally, achievement levels are added to operational items after they are calibrated and assigned to those levels the first time they are operational.

Table 11 presents the target range of the percentage of items by cognitive complexity on Statewide Science and Social Studies Assessments at the CAT item bank level. Table 12 presents the actual percentages in the item banks.

Table 11: Blueprint Percentage of Items by Cognitive Complexity

Grade/Course	Low	Moderate	High
Grade 5 Science	10%–20%	60%–80%	10%–20%
Grade 8 Science	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 12: Item Bank Observed Percentage of Items by Cognitive Complexity

Grade/Course	Low	Moderate	High
Grade 5 Science	14.9%	73.5%	11.6%
Grade 8 Science	16.7%	70.6%	12.6%

Grade/Course	Low	Moderate	High
Biology 1	14.5%	71.8%	13.8%
US History	25.0%	53.5%	21.5%
Civics	25.3%	50.2%	24.5%

Table 13 presents the observed percentages in item banks for ELA and mathematics.

FAST and B.E.S.T. assessments attend to cognitive complexity in the item development process. Items were written with the intention of building an item bank that includes items at DOK levels 1, 2, and 3, with an emphasis on DOK 2. An independent alignment study conducted in winter 2024–2025 affirmed that the item bank includes items across the range of DOK levels, as intended. Nearly all test forms (94%) were found to meet alignment expectations overall either fully or acceptably. Only four out of the 64 test forms (6%) were found to need slight or major adjustments due to unmet DOK consistency. More details about the alignment study can be found in Section 3.7, Alignment Process for Existing Items and Results from Alignment Studies, of this volume.

Table 13: Item Bank Observed Percentage of Items by Depth of Knowledge

Grade/Course	Depth of Knowledge		
Mathematics	**1**	**2**	**3**
3	21.7%	65.8%	12.6%
4	18.2%	70.8%	11.0%
5	16.3%	71.9%	11.7%
6	19.6%	69.8%	10.6%
7	20.0%	66.6%	13.4%
8	20.7%	70.3%	9.0%
Algebra 1	16.5%	69.5%	14.0%
Geometry	18.0%	64.9%	17.1%
ELA	**1**	**2**	**3**
3	5.1%	75.8%	19.1%
4	3.5%	74.9%	21.6%
5	2.0%	74.1%	24.0%
6	1.6%	74.2%	24.3%
7	0.5%	71.4%	28.1%
8	0.5%	74.6%	24.9%
9	0.3%	71.3%	28.4%
10	1.3%	73.8%	24.9%

2.2 CONTENT-LEVEL AND PSYCHOMETRIC CONSIDERATIONS

In addition to test blueprints, several content-level and psychometric considerations were used in the development of Florida's K–12 statewide student assessments. Content-level considerations included the following:

- Selected items addressed a variety of topics.

- Item content was accurate. This is important for social studies items that may include content affected by executive, legislative, and/or judicial actions at local, state, and national levels or because of changes emanating from international organizations.
- Correct answers were identified, or key was correct.
- Each item had only one correct response. Some technology-enhanced items (TEIs) had more than one correct answer. 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.
- Items were reviewed for potential bias and community sensitivity to ensure that they did not advantage or disadvantage any student based on personal characteristics and that the content was appropriate for Florida students and communities.

Items marked as do-not-use were not selected. Items adhered to required style specifications at time of development (e.g., italics, boldface) and were free from typographical, spelling, punctuation, and grammatical errors. Psychometric considerations included the following:

- A reasonable range of item difficulties was included.
- *p*-values for MC and constructed-response 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 –3 and 5.
- For MC items, IRT *c*-parameters were less than 0.40.
- Fit plots for items with model fit issues were evaluated.
- Items with differential item functioning (DIF) flags were evaluated.

Items that violate these specifications severely are removed automatically. Borderline items are taken to official data review meetings for content evaluation. More information on data review can be found in this volume under Section 3.5.2, Rubric Validation and Data Review.

More information about *p*-values, corrected point-biserial correlations, IRT parameters, and DIF calculations can be found in Volume 1, Annual Technical Report, of this report. Spring 2025 assessments were calibrated and equated to the IRT-calibrated item pool. More details about calibration, equating, and scoring can be found in Volume 1, Annual Technical Report, of this report.

3. ITEM DEVELOPMENT PROCEDURES

Item development procedures employed by Cambium Assessment, Inc. (CAI), for the Florida assessments were consistent with industry practice. The development of Florida's content and performance standards, test items, and stimuli was an open, consensus-driven process.

Item development began with test item specifications; Florida's State Academic Standards; language accessibility, bias, and sensitivity guidelines; editorial style guidelines; and principles of universal design (UD). 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 assessment principles of UD 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 possible range of students. The following seven principles of UD, as clearly defined by Thompson, Johnstone, & Thurlow (2002), were applied to assessment 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 applied these UD 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 UD was confirmed.

The application of 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, item development procedures implemented for Florida tests are consistent with industry practice. Specifically, Florida implements 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 item development processes, minimizing access barriers associated with tests to the greatest extent possible for all students, including SWDs and ELLs.

Test development specialists receive extensive training in item development, including instruction on UD principles and guidance on designing accessible content. Adherence to 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 Assessments Guide (FDOE, 2024), the processes of item development and test construction are guided carefully and include quality-control measures as follows:

- Item content on accommodated forms matches item content as administered online to the extent possible (e.g., wording, graphics, paragraph breaks, option order) via multiple rounds of content reviews. Note that some interactions will have accommodated form-specific language, such as equation (EQ) and table match (MI) items. This additional language is needed to guide students on how to appropriately answer some items on accommodated forms.
- Students see two-page items on an even page and an odd page simultaneously, just as they would see the entire item on one screen. Appropriate language is used for directives on the accommodated forms.

Regarding software that supports the item development process, CAI's Item Tracking System (ITS) served as the technology platform to carry out any item and test development process efficiently. ITS facilitated the creation of item banks, item writing and revision, cataloging of changes and comments, and exporting 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 added 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 could be configured to import items in multiple formats (e.g., Microsoft Word, Microsoft Excel, XML), CAI was able to import items from multiple sources. To support online test delivery, ITS has a unique Web Preview feature that displays items exactly as 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 paper-based test, but this feature provides an item-by-item blueline capability.

Before test administration, user acceptance testing (UAT) is performed on all approved platforms to ensure that items are rendered as expected and have similar appearances across platforms to minimize potential device effects.

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

The next section describes item sources, 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 2025 assessments came from multiple sources. New field-test items were included in the spring 2025 item pool. Only field-test items that meet psychometric criteria will be used on future tests as operational items. The newly developed field-test items were written for the Florida-specific item bank. Mathematics and English language arts (ELA) items were written by CAI content experts or by trained partners. Pearson contracted item writers to create new items for science and social studies. All items underwent a rigorous process of preliminary, editorial, and senior reviews by CAI and Florida Department of Education (FDOE) Test Development Center

(TDC) content and editorial teams, who followed appropriate test item and style specifications. All items were also reviewed by Florida educators and citizens for content accuracy and to ensure they were fair, unbiased, and included topics acceptable to the Florida public. Science and social studies item development processes also included subject-matter expert reviews.

3.2 ITEM TYPES

One important feature of online assessments is the administration of technology-enhanced items (TEIs). 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 support their answers.

Table 14–Table 16 list the item types and provide a brief description of each. For accommodated forms, 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 refer to the test design summary/blueprint documents or the test item specifications for specific details. Examples of various item types can be found in Appendix C, Example Item Types.

Table 14: ELA Reading Item Types and Descriptions

Response Type	Description
Multiple-Choice (MC)	Student selects one correct answer from a number of options.
Multiple-Select (MS)	Student selects all correct answers from a number of options.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row. On accommodated forms, the student fills in a bubble to indicate if information from a column header matches information from a row.
Hot-Text (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 accommodated forms, the student fills in bubbles to indicate which sentences are correct.
MC HT Selectable (Two-part HT)	Student selects the correct answers from Part A and Part B. Part A is an MC or an MS item, and Part B is a selectable HT item.
External Copy Interaction (ECI)	Student selects information directly from the passage to support an analysis or make an inference.
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 15: Mathematics and Mathematics EOC Item Types and Descriptions

Response Type	Description
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple-Select (MS)	Student selects all correct answers from a number of options.
Edit Task Inline Choice (ETIC)	Student specifies the number, word, or phrase. This includes items with one or more ETIC interactions. On accommodated forms, the student fills in a bubble to indicate the correct number, word, or phrase that should fill in the blank.

Response Type	Description
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. This item type is not used on accommodated forms.
Graphing	The student creates a graph or number line. The student can create a bar graph, line plot, or histogram by clicking on parts of a display. The student can plot a point, create a graph of an equation, draw a shape, or construct other responses by clicking on a grid or a coordinate grid. The student can plot a point on a number line by clicking on the number line or graph an inequality by clicking on the number line and choosing an arrow. This item type is not used on accommodated forms.
Hot-Text (HT)	Student is directed to click on one or more correct answers from among a number of options. When the student hovers over the options (e.g., phrases, sentences, numbers, or expressions), the text will highlight. This indicates that the text is selectable ("hot"). The options may be presented in various ways (e.g., as a list, embedded within text, or in a table). The student can then click on an option to select it. Student is directed to select text to support an analysis or make an inference. On accommodated forms, the student fills in bubbles to indicate which choices are correct.
Equation (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. On accommodated forms, the student uses an empty response box to write in their answer.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row. On accommodated forms, the student is directed to fill in a bubble that matches a correct option from a column with a correct option from a row.
Multi-Interaction (MULTI)	This is an item that contains more than one response type. It could contain more than one of the same interaction type (except for multiple combinations of ETIC) or a combination of interaction types.

Table 16: Science and Social Studies Item Types and Descriptions

Response Type	Description
Multiple-Choice (MC)	Student selects one correct answer from four options.

3.3 COGNITIVE LABORATORIES

In a United States Department of Education-funded grant report investigating the accessibility of computerized assessments, Shaftel et al. (2015) point out that TEIs present greater accessibility barriers than traditional item types on accommodated tests and that they should be examined to ensure that no construct-irrelevant variance is introduced. If some aspects of the technology impede, or advantage, 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 the Smarter Balanced Assessment Consortium (SBAC); therefore, research evidence on the SBAC platform can also be generalized to Florida assessments. Two types of research were conducted for SBAC: usability studies on system tools and features; and cognitive lab studies evaluating the validity of various item types. Findings show that various aspects of the test delivery system (e.g., tools, navigation, directions) provide students with equitable access to the assessed content; and TEI types do not introduce construct-irrelevant variance into scores. The full research report is provided in Volume

7, Special Studies, of the *Florida Standards Assessments 2014–2015 Technical Report*, which was included in an earlier submission for peer review.

Florida Assessment for Student Thinking (FAST), Benchmarks for Excellent Student Thinking (B.E.S.T.), and science cognitive labs were conducted to examine the response processes of test takers for grades 3, 7, and 10 ELA; Grades 3 and 7 mathematics and Algebra 1; and Grades 5 and 8 science and Biology 1. These grades/courses were selected because they represent the item types, share similar blueprints (including the same content categories), and have the same test development procedures as the non-selected grades/courses. The assessments are all based on the same content standards, benchmarks, and extensive content limits that define what is to be assessed. For all grades/courses, committees of educators collaborate with item development experts, assessment experts, and FDOE staff annually to review new and field-test items so that each test adequately samples the relevant domain of material the test is intended to cover. These committees review and verify the alignment of the test items with content standards and measurement specifications so that items measure the appropriate content. Given these commonalities between the selected and non-selected grades/courses, results from cognitive lab studies from the selected grades/courses are generalizable to non-selected grades/courses and non-selected item types.

In the studies, students worked through sample items. Eight students responded to each item, and their thinking processes were elicited through a combination of concurrent think-aloud (thinking aloud while reading and responding to an item) and focused probes that were tailored based on the anticipated solution path for a given item.

The cognitive lab interviews used recorded audio, and student responses to test items were captured by TDS. Following the cognitive lab, the interviewer reviewed all relevant information and files in a report that included, for each item attempted by the student, a detailed record of the student's think-aloud and responses to probes, as well as a record of the student's test item response.

Content experts evaluated these reports to determine whether the evidence for any given item met the following criteria:

- Students who receive full credit on an item display—through think-aloud and responses to probes—defensible evidence that they based their response on a combination of skills and knowledge that make up the "intended construct."
- Students who do not receive full credit on an item display—through think-aloud and responses to probes—defensible evidence that they understood (at a general level) what the item was asking them to do, and they were unable to provide a full-credit response as a result of deficiencies in one or more aspects of the skills or knowledge that make up the "intended construct." For example, they lacked the necessary procedural knowledge for manipulating fractions, or they were unable to apply reasoning skills required by the item.

The studies were delayed due to the COVID-19 pandemic and school closings in 2020–2021. They were completed in 2024. In comparison to the intended cognitive complexity, it was found that the enacted cognitive complexity either met or exceeded the intended cognitive complexity in 58%–88% of the items. Evidence of linguistic complexity that was construct-irrelevant was not found; however, students had significant difficulty reading algebra equations accurately,

suggesting a focal point teachers should consider targeting during instruction. Study findings were generalized across sampled grades. This study provides response process validity evidence that assessment items measure the intended cognitive processes represented in Florida's State Academic Standards. Full findings can be found in the cognitive laboratory report in Appendix G, Cognitive Lab Final Report.

3.4 ITEM TRANSLATIONS TO BRAILLE FORMAT

As noted in Allman (2009), it is common that portions of a test may need to be modified to be translatable to braille format. Modifications may include substituting words, reformatting item layout, 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 SWDs to properly demonstrate the knowledge, skills, and abilities the construct represents.

Florida uses a rigorous process, outlined in the *Florida Statewide Assessments Production Specifications*, when creating braille translations of its summative assessments and works with the Florida Instructional Materials Center for the Visually Impaired (FIMC-VI) and the American Printing House for the Blind (APH), both of which are leaders in the industry. 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 when modifications for the braille reader are necessary. For instance, sometimes, 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 Nemeth Unified English Braille (UEB) and Technical UEB versions of mathematics and science tests, and for all tests, we provide contracted and uncontracted versions to ensure that students with visual impairments 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*, and *UEB-Technical Contracted*.

We ensure that 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 students with visual impairments, who have vast experience and knowledge for 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 an advantage nor a 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 currently created for text-to-speech, to which all students have access. If tactile graphics are created, they are kept as true to the original version 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 tactfully 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

Field-test items were developed to be embedded in Florida's K–12 statewide student assessment program operational tests. As part of the standard test development process, item writers followed guidelines in FDOE-approved test item specifications and test design summary/blueprint.

CAI and Pearson staff used test item specifications to train qualified item writers, each of whom had prior item-writing experience. The item writers were trained at CAI item-writing workshops or had previous training on writing MC and constructed-response items. CAI content-area assessment specialists worked with TDC content leads and specialists to review measurement practices in item writing and interpret the meaning of Florida's State Academic Standards and benchmarks as illustrated by test item specification documents. This information, along with the purpose of the assessment, was explained to item writers. Sample item stems that are included in the specification documents serve as models for writers to use in creating items to match the standards. To ensure that item development addressed required ranges of cognitive complexity and difficulty, item writers used a cognitive classification system implemented by FDOE, based on Dr. Norman L. Webb's Depth of Knowledge (DOK) framework (Webb, 2002). 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 biases.

Similarly, reading passages should

- represent literary (fiction), informational (nonfiction), and practical selections (e.g., nontraditional pieces, including tables, charts, glossaries, indices);
- provide students with the opportunity to interact with complex, authentic texts that may employ a variety of different structures;
- include multimedia 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 biases;
- 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, etc.) to determine appropriateness for a particular grade level. Table 17 provides the guidelines used in FAST ELA reading assessments.

Table 17: Word Counts and Readabilities of Reading Passages in FAST ELA Reading

Grade	Word Count (approximate)	Lexile Range (approximate)
3	100–700	450–900
4	100–900	770–1,050
5	200–1,000	770–1,050
6	200–1,100	955–1,200
7	300–1,100	955–1,200
8	350–1,200	955–1,200
9	350–1,300	1,080–1,400
10	350–1,350	1,080–1,400

In FAST ELA reading assessments, texts are categorized as either informational or literary. *Informational texts* inform the reader and include the following types of publications:

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

Literary texts enable the reader to explore other people's experiences or to read for pleasure and include the following genres:

- 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 an internal review, 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 measured Florida's State Academic Standards appropriately. Items were then revised by CAI (for ELA and mathematics) or Pearson (for science and social studies) and brought to Florida Bias, Community Sensitivity, and Content Committees for review. After any final adjustments were made to the items, including a TDC editorial review, TDC provided a decision for each item: *Accept as Appears*, *Accept as Revised*, or *Reject*. Items approved by the TDC were web-approved and placed on field-test forms.

Committee Review of New Items

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

The Bias and Community Sensitivity Committees reviewed items for potential bias and controversial content. These committees consisted of Florida reviewers 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 content team members review the Bias and Community Sensitivity Committees' feedback with respective contractors and convey any issues to the attention of Content Item Review Committee participants.

The Content Item Review Committee consists of Florida educators and TDC content specialists by grade for each subject area. The social studies item review committee includes subject-matter experts. The primary responsibility of the committee members was to review all new items to ensure that they were free from flaws such as inappropriate readability level, ambiguity, incorrect or multiple answer keys (although some item types may include multiple answer keys by design),

unclear instructions, and 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.

In addition, science convenes an Expert Review Panel to confirm item accuracy and longevity and that best practices are represented. This review meeting takes place after Bias, Community Sensitivity, and Content committees.

3.5.2 Rubric Validation and Data Review

After items were field-tested, the rubric used for scoring MSCR items was validated by a team of grade-level Florida educators for ELA and mathematics. 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 rubric. The rubric validation meeting occurred in May 2025 in person in Tallahassee, Florida.

Like 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 GI and ECI items, HT draggable items, EQ items with full keypads, text entry natural language (NL) items, and MULTI items containing at least one of the preceding response types.

Items with simple rubrics include edit task choice and ETIC items, HT selectable items, MI items, EQ items with simple numeric keypads, MC items, HT selectable (two-part HT) items, and any MULTI items comprised entirely of the preceding response types.

MC items, MS items, and EBSR items do not go through rubric validation.

Before 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 data:

- Fifteen responses from students who performed as expected on the item given their overall performance
- Fifteen responses from students who were predicted to perform well on the item given their overall performance but instead performed poorly on the item
- Fifteen 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, CAI staff generated frequency tables that contained all student responses for each item. Frequency tables were generated out of CAI's Database of Record (DOR).

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, 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 responses were being scored consistently. 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, the 45 student responses 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 the score was incorrect or needed additional keys, the committee proposed modifications to the rubric.
- CAI rescored responses using the revised rubric.
- CAI reviewed responses that received changed scores to determine if they were scored correctly.
- TDC reviewed 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 was 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.

At the end of the testing window, CAI conducted classical item analysis on these field-test items to ensure that items functioned as intended with respect to underlying scales. CAI's analysis program computed the required item and test statistics for each MC and constructed-response item to check the integrity of the item and to verify the appropriateness of the item's difficulty level. Key statistical analyses included item discrimination, distractor analysis, item difficulty analysis, and fit analysis. Details of these analyses are presented in Section 5, Item Analyses Overview, of Volume 1.

Field-test items that survived statistical analysis were reviewed at the data review meeting. At the start of the data review meeting, CAI (ELA and mathematics) and Pearson (science and social studies) staff lead panels of Florida educators in a formal training to familiarize them with the item development process, the purpose of data review, the meanings of statistical flags, and how a data review committee operates. The training included a review of item cards, which detail specific item attributes (including grade level and alignment to the standards), content and rubric of the item, and various item statistics. Participants reviewed items on laptops to interact with item types as students do and to view all statistics associated with each item. Data review was conducted using CAI's Content Rater system in ITS.

The Content Rater also enables educators to see the online "item card" stored for each item in ITS. Item cards allow a viewer to see information regarding IRT statistics, fairness statistics, percentage

and average score of students in each point category, biserial correlations, and more. CAI can customize the Content Rater questions to the data review process and then use the system to capture individual educator decisions about the items. CAI facilitated group discussions of item data once individual reviews were completed. CAI recorded notes from group discussions of the item data directly in ITS. Final decisions made by the committee were noted in ITS, and CAI and Pearson worked with FDOE to discuss any items on which the committee could not reach a consensus.

Additionally, FDOE reviewed the items based on the IRT statistics flagging criteria. The lists of items dropped from both the data review with educators and through discussions with FDOE were combined, and ITS was updated to disqualify these items from being added to the operational pool. Items that pass both reviews are added to the operational pool for the next winter test administration for ELA and mathematics and during spring for science and social studies.

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 difficulties, and numbers in each reporting category. The psychometric team provides further guidance through a pool analysis of the overlap between item difficulty and student ability, items at each performance level for each reporting benchmark, and item difficulty spread with regards to the on-grade performance level cut score. This can be seen in Appendix H, Item Pool Analysis 2025. This is to ensure that continuous item development will result in a bank that can produce precise estimates of student performance. Adaptive tests select items tailored to the test taker's estimated ability. If the item bank lacks items near a student's true ability, the test will struggle to assess them accurately. The bank needs to span the full range of abilities in the population. Furthermore, the item difficulty to student ability overlap needs to exist across all required blueprint reporting categories and reported benchmarks, since the item administration algorithm must satisfy blueprint requirements before it can prioritize the match to student ability. Lastly, the distribution of item difficulty should also cover the area near the On Grade Cut Score sufficiently. Information about the adaptive test item selection algorithm is found in Section 4, Test Construction, of this volume.

In spring 2025, field-test items were embedded in online forms in grades 3–10 ELA reading and grades 3–8 mathematics, mathematics EOC tests, science, and social studies. An independent field test for ELA writing in grades 4–10 was also conducted in spring 2023. All assessments were computer-adaptive tests (CATs) with a predetermined number and location of field-test items. Table 18–Table 20 provide the number of field-test items by type for ELA reading, mathematics, science, and social studies. Table 21 provides the number of writing prompts for each grade.

Table 18: Number of ELA Reading Field-Test Items by Type

Item Type	3	4	5	6	7	8	9	10
EBSR	23	16	9	21	13	20	22	20
HT	6	8	6	4	10	23	14	14
MC	124	100	65	157	128	159	130	99
MI	13	10	9	17	19	13	10	13
MS	14	16	11	16	11	8	16	12
Two-Part HT	0	0	1	0	0	0	0	0
ECI	0	0	0	1	1	0	0	1

Table 19: Number of Mathematics and EOC Field-Test Items by Type

Item Type	3	4	5	6	7	8	Algebra 1	Geometry
EQ	51	69	79	71	79	65	64	74
ETIC	15	12	6	9	16	20	13	25
GI	0	1	0	0	0	1	0	0
HT	0	0	0	0	0	1	1	0
MC	33	34	44	52	64	80	66	51
MI	6	8	5	0	2	4	2	2
MS	29	18	7	9	6	10	11	9
MULTI	12	7	9	9	10	14	15	8
Graphing	1	0	2	0	3	3	0	0

Table 20: Number of Science and Social Studies Field-Test Items by Type

Item Type	Biology 1	Grade 5 Science	Grade 8 Science	Civics	U.S. History
MC	236	190	177	311	300

Table 21: Number of ELA Writing Field-Test Prompts by Grade in 2023

Grade	Number of Prompts
4	10
5	10
6	11
7	10
8	10
9	12

Grade	Number of Prompts
10	16

3.7 Alignment Process for Existing Items and Results from Alignment Studies

In a standards-based education system, evidence of alignment between standards and assessments is a central component of a comprehensive validity argument for the proposed interpretation and use of assessment results (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). A third-party, independent alignment study was conducted in February 2016. This report can be found in Volume 4, Evidence of Reliability and Validity, Appendix D, FSA Alignment Report, of the *Florida Standards Assessments 2015–2016 Technical Report*.

A new third-party, independent content alignment analysis was conducted by the Wisconsin Center for Education Products and Services (WCEPS) for FDOE over winter 2024–2025 to provide information about the degree of alignment of FAST for ELA reading grades 3–10, mathematics grades 3–8, and B.E.S.T. end-of-course (EOC) assessments of Algebra 1 and geometry with corresponding Florida B.E.S.T Standards for ELA and mathematics. The study was designed to yield evidence as pertains to fulfilling requirements within the Elementary and Secondary Education Act (ESEA) as amended by the Every Student Succeeds Act (ESSA, 2015), related to the alignment of a state's academic content assessments with the state's corresponding academic content standards. These federal expectations for technical quality are consistent with Florida Statute Title XLVIII K-20 Education Code §1008.22 *Student assessment program for public schools,* which expects the content of statewide assessments to be aligned with the content of Florida's State Academic Standards. Adequate alignment helps support a validity argument that student scores can be used to make appropriate inferences about student proficiency as relates to Florida's State Academic Standards. Details about any alignment weaknesses identified can help inform ongoing improvements to an assessment system.

Four main research questions guided this work:

1. Framework Analysis: To what extent do the test blueprints, item selection algorithm, and other relevant test specifications and documentation reflect a structure and design that support the capacity for alignment of test events with corresponding grade-level academic benchmarks?
2. Aggregate Data Review: To what extent do the available aggregate data for test events administered in spring 2024 provide evidence that the algorithm and blueprints are yielding test forms as expected?
3. Test Form and Level Alignment: What is the degree of alignment of actual test events sampled from achievement Level 2 (*below grade level*), achievement Level 3 (*on grade level*), and between achievement Level 4 (*proficient*) and Level 5 (*exemplary*) with corresponding Florida B.E.S.T. Standards, based on agreed-on criteria and minimum cutoffs?

4. Validation of Internal Metadata: To what extent is independent majority coding of assessment items by benchmark and DOK consistent with the assessment targets identified within internal item metadata?

For each subject area, three grade-band panels of six to 10 educators conducted an item-level analysis of four test events per grade: one test form sampled from at or near achievement Level 2 (*below grade level*), two test forms sampled from at or near achievement Level 3 (*on grade level*), and one test form sampled from between achievement Level 4 (*proficient*) and Level 5 (*exemplary*). For each subject area, a total of 32 test forms were analyzed. A total of 105–133 unique items were included in the analysis for each grade and subject area. Several sources of evidence were used to gauge generalizability of results.

Based on the alignment criteria and cutoffs agreed upon and used in this study, all FAST mathematics, B.E.S.T. EOC, and FAST ELA reading test forms analyzed for grades 3–7 and for grades 9 and 10 were found to fully or acceptably meet expectations for overall alignment with corresponding benchmarks. FAST ELA reading test forms for grade 8 were found to need slight or major adjustments to meet the expectations for alignment used in this analysis.

For FAST ELA reading assessments, validation of the internal metadata, based on the independent item-level content analysis, supports confidence in the generalizability of alignment findings across all test forms generated by the assessment program. For FAST mathematics and B.E.S.T. EOC assessments, independent verification of internally assigned benchmarks supports confidence in the generalizability of alignment findings for the criteria of Categorical Concurrence, Range of Knowledge (breadth), and Balance of Representation across all test forms generated by the assessment program. Independent assignment of DOK had limited-to-weak agreement with internal assignments for FAST mathematics. However, consistency of findings across the set of four test forms by grade, based on an independent analysis, suggests that it is still reasonable to expect that DOK consistency (depth) would be met or weakly met across all test forms generated by the assessment program.

Panelists provided a wide range of positive feedback about items. They also provided constructive and actionable feedback about a variety of issues and concerns. In the process of item-level content analyses, several errors were identified, some of which could impact student scores. In circumstances when errors are identified incidentally, it is recommended to closely and systematically check through all items as other instances may be found. All item-level coding and qualitative feedback were provided to FDOE and can serve as rich sources of information for ongoing item bank maintenance and improvement.

FDOE acknowledged the findings presented in the FAST ELA and Mathematics Alignment Report by WCEPS. FDOE has committed to reviewing and addressing the areas identified for adjustment, including targeted revisions to grade 8 ELA reading forms, reconciliation of metadata discrepancies, and continued quality assurance on item alignment and scoring accuracy. The full alignment report can be found in Appendix E, ELA and Mathematics Alignment Study Report, of this volume.

4. TEST CONSTRUCTION

4.1 OVERVIEW

During the 2022–2023 school year, the Florida Department of Education (FDOE) began transitioning from the fixed-form Florida Standards Assessment (FSA) to the computer-adaptive Florida Assessment of Student Thinking (FAST). In spring 2022, the first set of FAST items developed to align with the Benchmarks for Excellent Student Thinking (B.E.S.T.) Standards were field-tested. In summer 2022, field-test items were calibrated and placed on the FSA scale. Consistent with the progress monitoring (PM)1 and PM2 administrations, the spring 2023 FAST summative PM3 administration (as well as end-of-course [EOC] Algebra 1 and geometry) utilized Cambium Assessment, Inc. [CAI]'s adaptive algorithm to administer tests using these pre-equated items on the FSA scale. During this transition year, scores were reported to students on the FSA scale.

Subsequently, calibrations in summer 2023 placed items in English language arts (ELA) reading at grades 3–10 and mathematics at grades 3–8 on a common vertical FAST scale via a linking design. EOC Algebra 1 and geometry assessments were placed on the B.E.S.T. scale. Standard settings were conducted for all grades in ELA reading, mathematics, ELA writing, Algebra 1, and geometry. In the 2023–2024 school year and beyond, FDOE reported scores on the new FAST scale.

In spring 2023–2024, science and social studies also transitioned to computer-adaptive test (CAT) delivery. Calibrations in summer 2024 updated all item parameters, which were then linked back and placed on the original scale, utilizing existing cut scores and performance levels.

In addition to the online CAT, Florida also has accommodated forms. Accommodated forms were administered to students instead of the online forms if such a need was indicated on their Individualized Education Plan (IEP) or Section 504 Plan. Accommodated forms used online parameters for scoring purposes, and no calibrations were done on the accommodated forms.

4.2 ITEM SELECTION ALGORITHM

CAI's adaptive algorithm takes two sources of information as input: an item pool and a test blueprint. The adaptive algorithm is then configured to execute maximally adaptive test administrations under the constraint of blueprint match. Configuration of the adaptive algorithm is critical because composition of the item pool, which changes from administration to administration, interacts with the blueprint to influence the performance of the adaptive algorithm.

Item Pool

CAI's ability to administer various state item pools has been proven. For example, CAI administered items from the Smarter Balanced item bank during the 2013 pilot test and the 2014 field test. CAI designed and built the item renderers shared by the open-source version of the test delivery engine and CAI's version of the item-rendering software. These renderers ensure that items appear to students exactly as they did in the field test.

Test Blueprint

Test blueprints may contain specifications from content hierarchy (strand, benchmark, standard, etc.) and other constraints, such as item type or any other test item attribute that may be stored. CAI's adaptive engine supports blueprints that meet the following conditions (which have been advocated by the Consortium for Citizens with Disabilities, an umbrella group encompassing most national advocacy groups for students with disabilities [SWDs] and other exceptional students):

1. Every student is tested on the full range of grade-level content, with no discernible differences in the content assessed.
2. Every student is tested on items measuring the same mix of cognitively complex skills, with no discernible difference—regardless of student proficiency.
3. Every student is tested on items reflecting the full range of other aspects of the grade-level curriculum as may be appropriate for the grade and subject.
4. Students are tested on items that provide the best measurement possible within these constraints.

These four principles ensure that every student can demonstrate their academic skills and knowledge accurately across the entire grade-level curriculum. CAI's adaptive algorithm supports blueprints that align with these principles.

Item Selection

The adaptive algorithm, built on our partnerships with client states over the years, ensures that each student will receive a test that matches the blueprint and contains the items that best match students' performance levels, as defined by the blueprint. To accomplish this goal, the algorithm implements a highly parameterized multiple-objective utility function that includes

- a measure of the content match to the blueprint;
- a measure of overall test information; and
- measures of test information for each reporting category on the test.

We define an objective function that measures an item's contribution to each of these objectives, weighting them to achieve the desired balance among them. The following equation sketches this objective function for a single item.

$$f_{ijt} = w_2 \left(\frac{\sum_{r=1}^{R} s_{rit} p_r d_{rj}}{\sum_{r=1}^{R} d_{rj}} \right) + w_1 \sum_{k=1}^{K} q_k h_{1k}(v_{kijt}, V_{kit}, t_k) + w_0 h_0(u_{ijt}, U_{it}, t_0)$$

Where the w terms represent user-supplied weights that assign relative importance to meeting each of the objectives, d_{rj} indicates whether item j has the blueprint-specified feature r, and p_r is the user-supplied priority weight for feature r. The term s_{rit} is an adaptive control parameter that is described in this section. In general, s_{rit} increases for features that have not met their designated minimum as the end of the test approaches.

The remainder of terms represent an item's contribution to measurement precision:

- v_{kjit} is the value of item j toward reducing the measurement error for reporting category k for test taker i at time of selection t.
- u_{ijt} is the value of item j in terms of reducing the overall measurement error for test taker i at time of selection t.

U_{it} and V_{kit} represent the total information overall and on reporting category k, respectively.

q_k is a user-supplied priority weight associated with the precision of the score estimate for reporting category k. The t terms represent precision targets for the overall score (t_0) and each score reporting category score.

The functions $h(.)$ are given by

$$h_0(u_{ijt}, U_{it}, t_0) = \begin{cases} au_{ijt} \text{ if } U_{it} < t_0 \\ bu_{ijt} \text{ otherwise} \end{cases}$$

$$h_{1k}(v_{kijt}, V_{kit}, t_k) = \begin{cases} c_k v_{kijt} \text{ if } V_{kit} < t_k \\ d_k v_{kijt} \text{ otherwise} \end{cases}$$

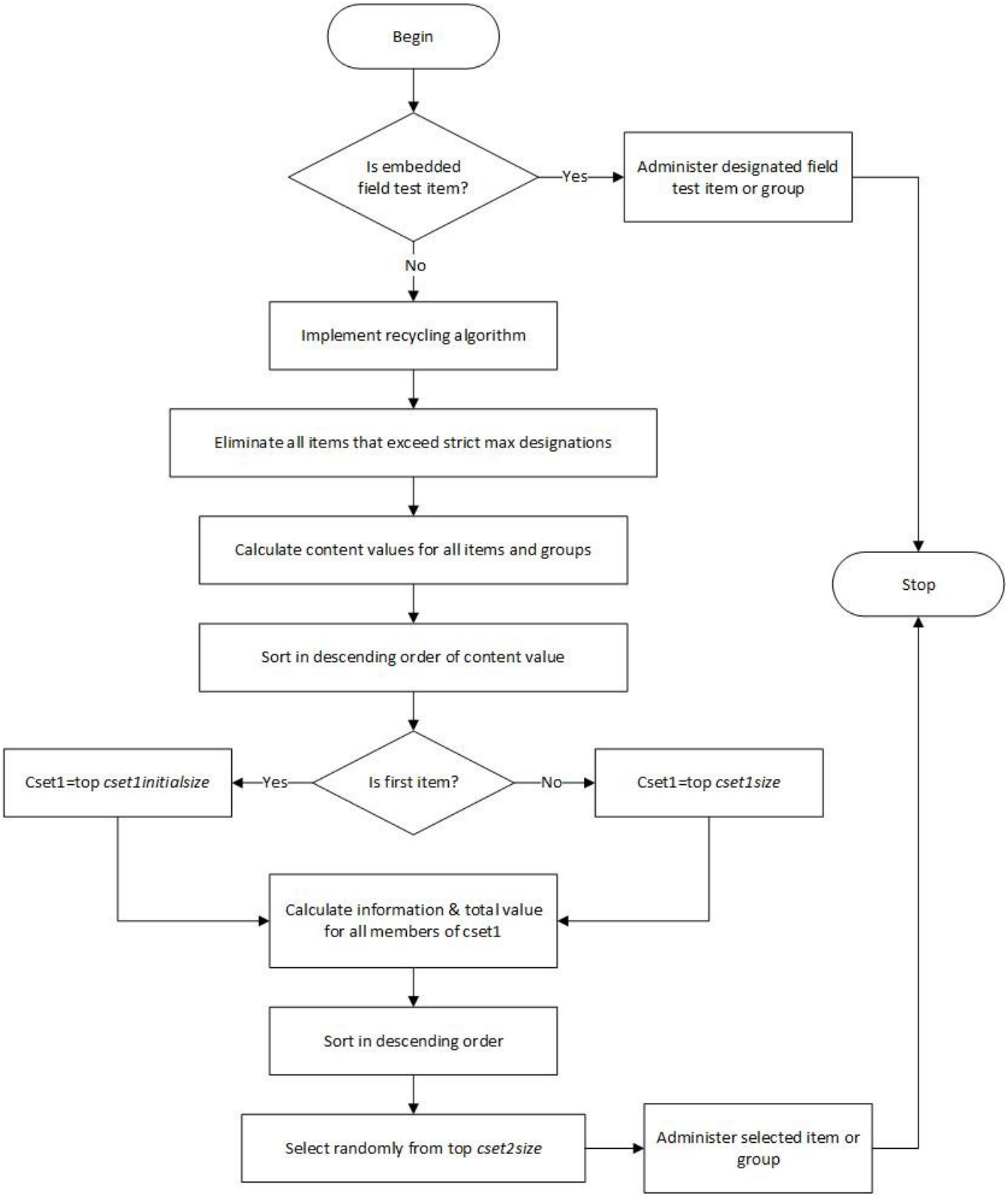
Items can be selected to maximize the value of this function. This objective function can be manipulated to produce a pure, standards-free adaptive algorithm by setting w_2 to zero or to produce a completely blueprint-driven test by setting $w_1 = w_0 = 0$. Adjusting the weights to optimize performance for a given item pool will enable users to maximize information subject to the constraint that the blueprint is virtually always met. We note that the computations of the content values and information values generate values on very different scales and that the scale of the content value varies as the test progresses. Therefore, we normalize both the information and content values before computing the value of the equation.

This normalization is given by $x = \begin{cases} 1 \; if \; min = max \\ \frac{v - min}{max - min} \; otherwise \end{cases}$, where min and max represent the minimum and maximum, respectively, of the metric computed over the current set of items or item groups.

Items (or groups of items for ELA tests) are sorted by their "content value," their value toward meeting the content constraints in the blueprint. Information measures are added to the content measures, and the items are sorted based on their overall value for the objective function. Final item selection is made based on a random selection from the small subset of items that have the highest combined content and information value.

Figure 1 summarizes the item selection process. If the item position has been designated for a field-test item, then a field-test item is administered. Otherwise, the adaptive algorithm is triggered.

Figure 1: Item Selection Process



Blueprint Match

Configuration of the adaptive algorithm for spring 2025 test administration was designed to administer tests meeting blueprint specifications while maximizing test information to student ability for all subjects. In the adaptive item selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability.

While the simulation results described in the Spring 2025 Simulation Summary Report (refer to Appendix F, Spring 2025 Simulation Results) indicated that the configuration resulted in the test administrations meeting all blueprint match requirements, it is also important to evaluate the blueprint match rate for the actual test administrations.

Appendix D, Spring 2025 Operational Item Blueprint Match, contains the operational item blueprint match results for spring 2025 grades 3–8 mathematics, grades 3–10 ELA reading, EOC Algebra 1 and geometry, grades 5 and 8 science, EOC Biology 1, U.S. History, and Civics assessments. For the blueprint match analysis, only students who completed all parts of the test were included. As can be seen in Appendix D, in all assessments, all reporting categories met the blueprint or blueprint range. The observed percentage of reading passage types by grade is also documented.

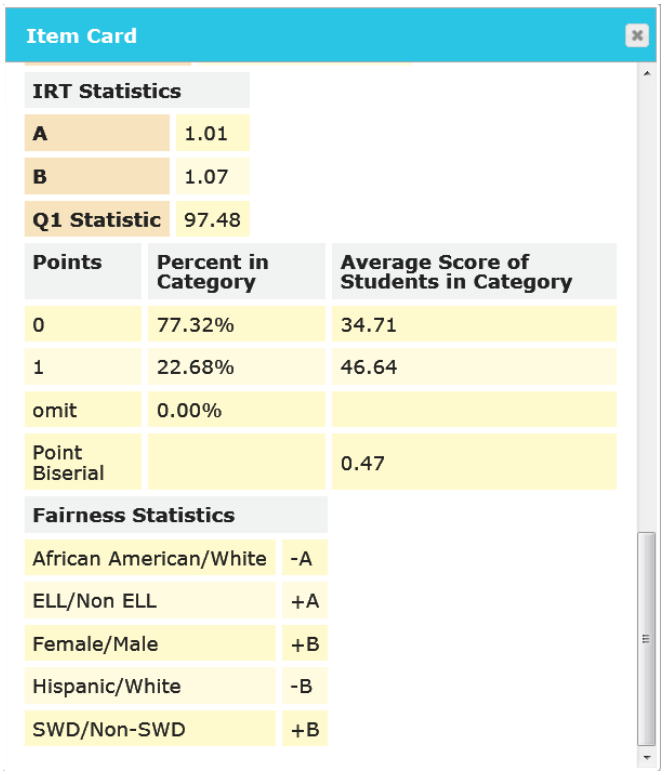
In 2025, 100% blueprint match was observed for all reporting categories, including passage maximums for all grades. For some tests, item recycling in the adaptive algorithm was necessary. Eventually, with a large enough item pool, item recycling will not be necessary.

4.3 TEST CONSTRUCTION SUMMARY MATERIALS

4.3.1 Item Cards

Item cards, generated within CAI's Item Tracking System (ITS), contained statistical information about an individual item. Item cards contained classical item statistics, item response theory (IRT) statistics, and differential item functioning (DIF) statistics. Items cards are linked to the items and can be 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 2 shows one example of an item card.

Figure 2: 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 Workspace Files for Accommodated Forms

Workspace files were provided for accommodated forms only. They are generated by CAI's automated form-building tool inside ITS. A Workspace export file is a spreadsheet that lists the characteristics of all items on a form and form-level statistical information. They can also be viewed directly in ITS. Workspace files contain information such as the following:

- Item ID
- Item position
- Form
- Grade
- Role (e.g., operational, field-test)
- Item format (e.g., multiple choice [MC])
- Point value
- Answer key
- Reporting category
- Cognitive complexity
- Test characteristic curves

Workspace files are used as an accessible resource by content specialists and psychometricians to find information about a test form. Workspace files contain useful information regarding the forms

that are built in ITS. During the form construction phase, content and psychometric teams can use the Workspace files directly in ITS to construct forms collaboratively.

4.4 ACCOMMODATED FORM CONSTRUCTION

Student scores should not depend on the mode of administration or type of test form. Because Florida's 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. During test development, forms across all modes were required to adhere to the same test blueprints and content-level considerations. This section outlines the overall test development plans that ensured the comparability of online tests and accommodated tests.

To create spring 2025 accommodated forms, CAI's automated form-building tool inside CAI's ITS was used. ITS is a web-based software application that, in conjunction with the Item Authoring Tool (IAT), which can be accessed through ITS, enables users to create items and stimuli for testing purposes. Once the items and stimuli have been created, users can review, approve, and publish these items in ITS so that they can be administered to students through the Test Delivery System (TDS). ITS serves as an item repository and comprises several databases, referred to as item banks, which contain items specific to a client or project. It is an integrated system that supports all phases of test development. It facilitates direct item entry by item writers, online review and editing by reviewers, automated reporting for clients, and management and production of test forms.

Psychometric targets were set for each item and test form by CAI psychometricians in conjunction with FDOE, using the automated form-building tool inside ITS. Florida-accommodated fixed-form assessments have two distinct psychometric targets: item-level targets and test-level targets. Item-level targets are related to item statistics such as desired item difficulty, discrimination, parameter ranges (such as possibility of guessing), fit, and content bias. Test-level targets are related to test-level summaries and aggregated information, including desired average test difficulty, average item difficulty, target test information, standard error of measurement (SEM), and test characteristic curves. The nature of the item-level targets is predefined and does not change from one test administration to the next; however, the overall test targets might be updated with respect to policy needs and scale drift. All test forms must always meet blueprint targets. Items that failed the psychometric targets are flagged for review and removed if it is still possible to meet the blueprint without them.

For the test-level targets, the most important consideration is the SEM curve that shows the level of error of measurement expected at each ability level. SEM is calculated as the reciprocal of the square root of the test information curve; therefore, the SEM 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. Test construction always aims to minimize error at the score points at which relevant decisions (e.g., pass, fail) may be made.

Content specialists reviewed the forms and made any necessary item replacements, considering suitability for inclusion in an accommodated form and psychometric feedback. To build accommodated forms, content specialists began with the selected forms and removed any technology-enhanced items (TEIs) that could not be rendered on accommodated or machine-scored forms. These items were then replaced with either MC items or other TEIs that could be

rendered on accommodated forms from the same reporting category. In some instances, it was necessary to select replacement items from a different reporting category to satisfy statistical expectations; however, all parties ensured that each reporting category was still represented appropriately in the final test forms. Two of the forms with the best statistics were selected to be sent to FDOE for evaluation and selection of a final form.

5. REFERENCES

Allman, C. (2009). *Test access*: *Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel* (4th ed.). American Printing House for the Blind. https://sites.aph.org/wp-content/uploads/2017/09/Test-Access-Making-Tests-Accessible-2009.pdf

Florida Department of Education. (2024, August). *Florida Statewide Assessments Guide.* https://www.fldoe.org/core/fileparse.php/20102/urlt/K12SAG.pdf

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*. University of Kansas. https://ateassessments.atlas4learning.org/sites/default/files/ATEA%20Project%20Report%20111615.pdf

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. 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). Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED543067.pdf