

Connecticut Smarter Balanced Assessments 2021–2022 Technical Report



**Submitted to
Connecticut State Department of Education
by Cambium Assessment, Inc.**

TABLE OF CONTENTS

1. OVERVIEW	1
2. TEST ADMINISTRATION.....	4
2.1 Testing Windows	4
2.2 Test Options and Administrative Roles	4
2.2.1 Administrative Roles	5
2.2.2 Online Test Administration.....	7
2.2.3 Paper-Pencil Test Administration	8
2.2.4 Braille Test Administration	8
2.3 Training and Information for Test Coordinators and Administrators	9
2.3.1 Online Training.....	9
2.3.2 District Test Coordinator Training Workshops	12
2.4 Test Security	12
2.4.1 Student-Level Testing Confidentiality	12
2.4.2 System Security	13
2.4.3 Security of the Testing Environment.....	13
2.4.4 Test Security Violations	14
2.5 Student Participation	15
2.5.1 Home-Schooled Students.....	15
2.5.2 Exempt Students	15
2.6 Online Testing Features and Testing Accommodations.....	15
2.6.1 Online Universal Tools for All Students.....	16
2.6.2 Designated Supports and Accommodations.....	17
2.7 Testing Time	26
2.8 Data Forensics Program	28
2.8.1 Changes in Student Performance.....	29
2.8.2 Test-Taking Time.....	29
2.8.3 Inconsistent Item Response Pattern (Person Fit).....	30
2.8.4 Item Response Change	30

2.9	Prevention and Recovery of Disruptions in Test Delivery System.....	31
2.9.1	<i>High-Level System Architecture</i>	31
2.9.2	<i>Automated Backup and Recovery</i>	33
2.9.3	<i>Other Disruption Prevention and Recovery Systems</i>	33
3.	SUMMARY OF 2021–2022 OPERATIONAL TEST ADMINISTRATION	34
3.1	Student Population.....	34
3.2	Summary of Student Performance.....	35
3.3	Distribution of Student Ability and Item Difficulty	45
4.	VALIDITY	52
4.1	Evidence on Test Content	52
4.2	Evidence on Internal Structure	57
5.	RELIABILITY	60
5.1	Marginal Reliability.....	60
5.2	Standard Error Curves	61
5.3	Reliability of Achievement Classification.....	64
5.4	Reliability for Subgroups	69
5.5	Reliability for Claim Scores	72
6.	SCORING.....	74
6.1	Estimating Student Ability Using Maximum Likelihood Estimation.....	74
6.2	Rules for Transforming Theta to Vertical Scale Scores	75
6.3	Lowest/Highest Obtainable Scores (LOSS/HOSS).....	76
6.4	Scoring All Correct and All Incorrect Cases	77
6.5	Rules for Calculating Strengths and Weaknesses for Claim Scores	77
6.6	Target Scores	77
6.6.1	<i>Target Scores Relative to Student’s Overall Estimated Ability</i>	77
6.6.2	<i>Target Scores Relative to Proficiency Standard (Level 3 Cut)</i>	78
6.7	Hand-scoring	79
6.7.1	<i>Rater Selection</i>	80
6.7.2	<i>Rater Training and Scoring</i>	81
6.7.3	<i>Rater Statistics and Monitoring</i>	83

6.7.4 Rater Retraining and Dismissal.....	84
6.7.5 Rater Agreement.....	84
7. REPORTING AND INTERPRETING SCORES.....	86
7.1 Centralized Reporting System.....	86
7.1.1 Dashboard.....	88
7.1.2 Aggregate Score Reports: Overall Performance.....	89
7.1.3 Aggregate Score Reports: Claim and Target Performance.....	90
7.1.4 Roster Performance Report.....	91
7.1.5 Trend Report.....	92
7.1.6 Individual Student Report.....	93
7.1.7 Paper Family Score Reports.....	95
7.2 Interpretation of Reported Scores.....	97
7.2.1 Scale Score.....	97
7.2.2 Conditional Standard Error of Measurement.....	97
7.2.3 Achievement Level.....	97
7.2.4 Performance Category for Claims.....	98
7.2.5 Performance Category for Targets.....	98
7.2.6 Aggregated Scale Score.....	98
7.3 Appropriate Uses of Test Results.....	99
8. QUALITY CONTROL PROCEDURE.....	100
8.1 Adaptive Test Configuration.....	100
8.1.1 Platform Review.....	100
8.1.2 User Acceptance Testing and Final Review.....	101
8.2 Quality Assurance in Document Processing.....	101
8.3 Quality Assurance in Data Preparation.....	101
8.4 Quality Assurance in Online Test Delivery System.....	101
8.4.1 Score Report Quality Check.....	102
REFERENCES.....	105

LIST OF TABLES

Table 1. 2021–2022 Testing Windows	4
Table 2. 2021–2022 Testing Options.....	4
Table 3. Number of Students Who Took Paper-Pencil Tests in the 2021–2022 Summative Test Administration	8
Table 4. 2021–2022 Universal Tools, Designated Supports, and Accommodations	22
Table 5. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations	23
Table 6. ELA/L Total Students with Allowed Embedded Designated Supports	23
Table 7. ELA/L Total Students with Allowed Non-Embedded Designated Supports.....	24
Table 8. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations....	24
Table 9. Mathematics Total Students with Allowed Embedded Designated Supports.....	25
Table 10. Mathematics Total Students with Allowed Non-Embedded Designated Supports	26
Table 11. ELA/L Testing Times	27
Table 12. Mathematics Testing Times.....	28
Table 13. Participation Rates in ELA/L Summative Assessment	34
Table 14. Participation Rates in Mathematics Summative Assessment	34
Table 15. Number of Students in ELA/L Summative Assessment.....	35
Table 16. Number of Students in Mathematics Summative Assessment	35
Table 17. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 3–5).....	36
Table 18. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: ELA/L (Grades 6–8).....	37
Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 3–5)	38
Table 20. Descriptive Statistics and Percentage of Students in Achievement Levels for Overall and by Subgroup: Mathematics (Grades 6–8)	39
Table 21. ELA/L Percentage of Students in Performance Categories for Claims.....	44
Table 22. Mathematics Percentage of Students in Performance Categories for Claims	45
Table 23. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: ELA/L Grades 3–5	53
Table 24. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: ELA/L Grades 6–8	54

Table 25. Percentage of Delivered Tests Meeting Blueprint Requirement for Each Claim and Target: Mathematics Grades 3–5.....	55
Table 26. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: Mathematics Grades 6–8.....	56
Table 27. Average and the Range of the Number of Unique Targets Assessed within Each Claim Across All Delivered CAT Components	57
Table 28. Correlations among Claims for ELA/L.....	58
Table 29. Correlations among Claims for Mathematics.....	59
Table 30. Marginal Reliability for ELA/L and Mathematics	61
Table 31. Average Conditional Standard Errors of Measurement by Achievement Levels	64
Table 32. Average Conditional Standard Errors of Measurement at Each Achievement Level Cut and Difference of the Standard Errors of Measurement between Two Cuts.....	64
Table 33. Classification Accuracy and Consistency	68
Table 34. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 3–4)	69
Table 35. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 5–6)	69
Table 36. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 7–8)	70
Table 37. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 3–4)....	70
Table 38. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 5–6)....	71
Table 39. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 7–8)....	71
Table 40. Marginal Reliability Coefficients for Claim Scores in ELA/L.....	72
Table 41. Marginal Reliability Coefficients for Claim Scores in Mathematics	73
Table 42. Vertical Scaling Constants on the Reporting Metric	75
Table 43. Cut Scores in Scale Scores	76
Table 44. Lowest and Highest Obtainable Scores.....	76
Table 45. Number of Hand-Scored Items in 2021–2022 Connecticut Summative Item Pool, by Grade and Subject	80
Table 46. Inter-Rater Agreement for ELA/L Short-Answer Items.....	84
Table 47. Inter-Rater Agreement for Mathematics Items.....	85
Table 48. Types of Online Score Reports by Level of Aggregation.....	87
Table 49. Types of Subgroups.....	87
Table 50. Overview of Quality Assurance Reports.....	102

LIST OF FIGURES

Figure 1. ELA/L Percent Proficient Across Years.....	40
Figure 2. Mathematics Percent Proficient Across Years	41
Figure 3. ELA/L Average Scale Score Across Years.....	42
Figure 4. Mathematics Average Scale Score Across Years	43
Figure 5. Student Ability—Item Difficulty Distribution for ELA/L.....	46
Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5).....	47
Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8).....	48
Figure 8. Student Ability—Item Difficulty Distribution for Mathematics	49
Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)	50
Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8)	51
Figure 11. Conditional Standard Errors of Measurement for ELA/L	62
Figure 12. Conditional Standard Errors of Measurement for Mathematics.....	63

LIST OF EXHIBITS

Exhibit 1. Dashboard: District Level	88
Exhibit 2. Detailed Dashboard: District Level.....	89
Exhibit 3. Overall Performance Summary Results for Grade 3 ELA/L: District Level	90
Exhibit 4. Overall Performance Summary Results for Grade 3 ELA/L by Gender: District Level.....	90
Exhibit 5. Claim and Target Level Results for Grade 5 Mathematics: District Level	91
Exhibit 6. Roster Performance Report for Grade 3 ELA/L	92
Exhibit 7. Trend Report for ELA/L: Student Level	93
Exhibit 8. Individual Student Report for Grade 5 ELA/L.....	94
Exhibit 9. Sample Paper Family Score Report	96

LIST OF APPENDICES

Appendix A: Summary of the 2021–2022 Interim Assessments 107
Appendix B: Student Performance Across Four Years for All Students and by Subgroups 124
Appendix C: Classification Accuracy and Consistency Index by Subgroups..... 130

1. OVERVIEW

The Smarter Balanced Assessment Consortium (SBAC) developed a next-generation assessment system. The assessments are designed to measure the Common Core State Standards (CCSS) in English language arts/literacy (ELA/L) and mathematics for grades 3–8 and 11, and to provide valid, reliable, and fair test scores about student academic achievement. Connecticut was among 18 member states (plus the U.S. Virgin Islands) leading the development of assessments in ELA/L and mathematics. The system includes both summative assessments, for accountability purposes, as well as optional interim assessments that provide meaningful feedback and actionable data that teachers and educators can use to help students succeed. SBAC, a state-led enterprise, is intended to provide leadership and resources to improve teaching and learning by creating and maintaining a suite of summative and interim assessments and tools aligned to the CCSS in ELA/L and mathematics.

The Connecticut State Board of Education formally adopted the CCSS in ELA/L and mathematics on July 7, 2010. All students in Connecticut, including students with significant cognitive disabilities who are eligible to take the Connecticut Alternate Assessment (CTAA), an alternate assessment based on alternate academic achievement standards (AA-AAAS), are taught content that aligns to the same academic standards. Connecticut CCSS define the knowledge and skills students need to succeed in college and careers after graduating from high school. These standards include rigorous content and application of knowledge through higher-order skills and align with college and workforce expectations.

The Connecticut statewide assessments in ELA/L and mathematics aligned with the CCSS were administered for the first time in spring 2015 to students in grades 3–8 and 11 in all public elementary and secondary schools. In 2015–2016, Connecticut adopted the Scholastic Aptitude Test (SAT) to replace the Smarter Balanced grade 11 assessments for high school students.

The Smarter Balanced assessments are composed of the end-of-year summative assessment designed for accountability purposes and the optional interim assessments designed to support teaching and learning throughout the year. The summative assessments are used to determine student achievement based on the CCSS and track student progress toward college and career readiness in ELA/L and mathematics. The summative assessments consist of two parts: a computer-adaptive test (CAT) and a performance task (PT).

- **Computer-Adaptive Test (CAT).** The CAT is an online adaptive test that provides an individualized online assessment for each student.
- **Performance Task (PT).** A PT is a task that challenges students to apply their knowledge and skills to respond to real-world problems. PTs can best be described as collections of items and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis, none of which can be adequately assessed with selected-response or constructed-response items. Some PT items can be scored by the computer, but most are handscored.

Optional interim assessments allow teachers to monitor student progress throughout the year and provide information that teachers can use to improve their instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to gauge students' progress in mastering specific concepts at strategic points during the school year. Two types of interim assessments are available as fixed-form tests:

- **Interim Comprehensive Assessments (ICAs).** ICAs test the same content and report scores on the same scale as the summative assessments.
- **Interim Assessment Blocks (IABs).** IABs focus on smaller sets of related concepts and provide more detailed information about student learning.

Starting in the 2015–2016 summative test administration, Connecticut made four changes in the summative tests:

- Replaced the summative ELA/L and mathematics assessments in grade 11 with the SAT reading and writing, language, and mathematics tests.
- Removed the summative field-test items and off-grade items from the ELA/L and mathematics CAT item pool.
- Removed PTs in ELA/L while keeping PTs in mathematics assessment. For the paper-pencil tests, the test booklet will include both non-PT and PT components, but only the non-PT component will be scored for ELA/L.
- Reported scores for combining claim 2 (writing) and claim 4 (research/inquiry) in ELA/L.

Due to the COVID-19 pandemic, the U.S. Department of Education waived testing requirements in the 2019–2020 school year (<https://www2.ed.gov/policy/gen/guid/secletter/200320.html>). For the 2022–2021 school year, the U.S. Department of Education did not grant waivers for standardized testing but did waive certain accountability requirements (e.g., mandatory high participation rates) due to the impacts of the pandemic in many states, resulting in lower participation rates than in previous years. In the 2020–2021 school year, all public schools in Connecticut participated in the Smarter Balanced summative assessments in grades 3–8, with the participation rates of 90.7–95.2% in ELA/L and 88.1–94.4% in mathematics. In the 2020–2021 test administration, the Connecticut State Department of Education (CSDE) allowed remote testing in addition to in-person testing. The percentage of the students who took the summative tests remotely ranged from 9.8% to 14.4% in ELA/L and from 9.7% to 13.8% in mathematics in grades 3–8.

In the 2021–2022 school year, the participation rates increased, ranging from 96.7% to 97.9% in ELA/L and from 96.0% to 97.7% in mathematics, and remote testing was not allowed for the summative tests.

This report provides a technical summary of the 2021–2022 summative assessments in ELA/L and mathematics administered in grades 3–8 under the Connecticut Smarter Balanced assessments. The report is divided into eight chapters: Overview; Test Administration; Summary of the 2021–2022 Operational Test Administration; Validity; Reliability; Scoring; Reporting and Interpreting Scores; and Quality Control Procedures. The data included in this report are based on Connecticut data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs and a summary of their performance are provided in Appendix A.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration for Connecticut, it is an addendum to the 2021–2022 Smarter Balanced technical report. The Smarter Balanced technical report contains information on item and test development, item content review, field-test administration, item-data review, item calibrations, content alignment study, standard setting, and other validity information.

Smarter Balanced produces a technical report on the Smarter Balanced assessments that covers all aspects of their compliance with the technical qualities described in the *Standards for Educational and*

Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) and the requirements of the U.S. Department of Education, outlined in *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States* (U.S. Department of Education, 2015). The Smarter Balanced technical report includes analysis of the data at the consortium level, combining data from the consortium members.

2. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2021–2022 Smarter Balanced assessments testing window spanned approximately two and a half months for the summative assessments and eight months for the interim assessments. The paper-pencil fixed-form tests for summative assessments were administered concurrently during the online summative window. Table 1 shows the testing windows for the online, remote, and paper-pencil assessments.

Table 1. 2021–2022 Testing Windows

Tests	Grades	Start Date	End Date	Mode
Summative Assessments	3–8	March 28, 2022	June 3, 2022	Online Adaptive
	3–8	March 28, 2022	June 3, 2022	Paper-Pencil Fixed Forms
Interim Comprehensive Assessments	3–8, 11	September 2, 2021	June 10, 2022	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	September 2, 2021	June 10, 2022	Online Fixed Forms

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The Smarter Balanced assessments are administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the Smarter Balanced assessments, several assessment options were available for the 2021–2022 administration to accommodate students’ needs. Table 2 lists the testing options that were offered in 2021–2022. A testing option is selected by content area. Once a testing option is selected, it applies to all tests in the content area.

Table 2. 2021–2022 Testing Options

Assessments	Test Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Braille HAT (Hybrid Adaptive Test) (mathematics only)	Online
	Spanish (mathematics only)	Online
	Paper-Pencil, Large-Print, Fixed-Form Test*	Paper-Pencil
	Paper-Pencil, Braille, Fixed-Form Test*	Paper-Pencil
Interim Assessments	English	Online
	Braille	Online
	Spanish (mathematics only)	Online

* For the paper-pencil fixed-form tests, all student responses on the paper-pencil tests were entered in the Data Entry Interface (DEI) by test administrators.

To ensure standardized administration conditions, teachers (TEs) and test administrators (TAs) follow procedures outlined in the *Smarter Balanced ELA/L and Mathematics Online, Summative Test Administration Manual* (TAM). TEs and TAs must review the TAM prior to the beginning of testing to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures should be established for any students who are absent on testing days. TEs and TAs follow required administration procedures and directions and read the boxed directions verbatim to students, ensuring standardized administration conditions.

2.2.1 Administrative Roles

The key personnel involved with the test administration for the Connecticut State Department of Education (CSDE) are District Administrators (DAs), District Test Coordinators (DTCs), School Test Coordinators (STCs), teachers (TEs), and test administrators (TAs). Their main responsibilities are described in the following subsections. More detailed descriptions can be found in the TAM provided online at this URL: <https://ct.portal.cambiumast.com/resources/>.

District Administrator

The DA may add users with DTC roles in the Test Information Distribution Engine (TIDE). For example, a director of special education may need DTC privileges in TIDE to access district-level data for the purposes of verifying test settings for designated supports and accommodations. DAs have the same test administration responsibilities as DTCs. Their primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

District Test Coordinator

The DTC is primarily responsible for coordinating the administration of the Smarter Balanced assessments at the district level.

DTCs are responsible for the following:

- Reviewing all Smarter Balanced policies and test administration documents
- Reviewing scheduling and testing requirements with STCs, TEs, and TAs
- Working with STCs and technology coordinators (TCs) to ensure that all systems, including the CAI Secure Browser, are properly installed and functional
- Importing users (including STCs, TEs, and TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all STCs, TEs, TAs, and TCs
- Ensuring that all personnel are trained on how to administer the Smarter Balanced assessments properly
- Monitoring the secure administration of the tests
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE and Smarter Balanced policies

School Test Coordinator

The STC is primarily responsible for coordinating the administration of the Smarter Balanced assessments at the school level and ensuring that testing within his or her school is conducted in accordance with the test procedures and security policies established by the CSDE.

STC responsibilities include the following:

- Based on testing windows, establishing a testing schedule with DTCs, TEs, and TAs

- Working with technology staff to ensure timely computer setup and installation
- Working with TEs and TAs to review student information in TIDE to ensure that student information and test settings for designated supports and accommodations are correctly applied
- Identifying students who may require designated supports and test accommodations, and ensuring that procedures for testing these students follow CSDE and Smarter Balanced policies
- Attending all district trainings and reviewing all Smarter Balanced policies and test administration documents
- Ensuring that all TEs and TAs attend school or district trainings and review online training modules posted on the Connecticut state portal
- Establishing secure and separate testing rooms if needed
- Downloading and planning the administration of the classroom activity with TEs and TAs
- Monitoring secure administration of the tests
- Monitoring testing progress during the testing window, and ensuring that all students participate, as appropriate
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE and Smarter Balanced policies

Teacher

A TE who is responsible for administering the Smarter Balanced assessments must have the same qualifications as a TA. TEs also have the same test administration responsibilities as TAs. TEs are able to view their own students' results when they are made available. This role may also be assigned to teachers who do not administer the test but will need access to student results.

Test Administrator

A TA is primarily responsible for administering the Smarter Balanced assessments. The TA's role does not allow access to student results and is designed for TAs, such as technology staff, who administer tests but do not have access to student results.

TAs are responsible for the following:

- Completing Smarter Balanced test administration training
- Reviewing all Smarter Balanced policy and test administration documents before administering any Smarter Balanced assessments
- Viewing student information before testing to ensure that a student receives the proper test with the appropriate supports, and reporting any potential data errors to STCs and DTCs, as appropriate
- Administering the Smarter Balanced assessments
- Reporting all potential test security incidents to the STCs and DTCs in a manner consistent with Smarter Balanced, CSDE, and district policies

2.2.2 Online Test Administration

Within Connecticut’s testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long testing period, minimizing the interruption of classroom instruction and efficiently utilizing its facility. With online testing, schools do not need to handle test booklets and address the storage and security problems inherent in large shipments of materials to a school site.

STCs oversee all aspects of testing at their schools and serve as the main point of contact, while TEs and TAs administer the online assessments only. TEs and TAs are trained in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the test administration are provided online. All school personnel who serve as TEs and TAs are encouraged to complete CAI’s online TA Certification Course. Staff who complete this course receive a certificate of completion.

To start a test session, the TE or TA must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA must enter their State Student Identification Number (SSID), their first name, and the session ID into the Student Interface using computers provided by the school. The TE or TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility features (see Section 2.6 for a list of accommodations). Students can begin testing only when the TA or TE confirms the settings. The TA or TE then reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the students and guides them through the login process.

Once an assessment has started, the student must answer all the test items presented on a page before proceeding to the next page. Skipping items is not permitted. For the online computer-adaptive test (CAT), students are allowed to scroll back to review and edit previously answered items, as long as these items are in the same test session and this session has not been paused for more than 20 minutes. Students may review and edit responses they have previously provided before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, then all items that follow to which the student already responded remain the same. If a student changes the answers, no new items are assigned. For example, a student pauses for 10 minutes after completing Item 10. After the pause, the student goes back to Item 5 and changes the answer. If the response change in Item 5 changes the item score from wrong to right, the student’s overall score will improve; however, there will be no change in Items 6–10.

There is no pause rule implemented for the performance tasks (PTs). The same rules that apply to the CAT for reviews and changes to responses also apply to PTs.

For the summative test, an assessment can be started in one component and completed in another. For the CAT, the assessment must be completed within 45 calendar days of the start date or the assessment opportunity will expire. For the PTs, the assessment must be completed within 20 calendar days of the start date.

During a test session, TEs or TAs may pause the test for a student or group of students to take a break. It is up to the TEs or TAs to determine an appropriate stopping point; however, to ensure the integrity of test scores or testing, the CAT cannot be paused for more than 30 minutes for English language arts/literacy (ELA/L) and mathematics. If that happens, the student must restart a new test session, which starts from where the student left off. The viewing and editing options of previous responses are no longer available.

The TAs or TEs must always remain in the room during a test session to monitor student testing. Once the test session ends, the TAs or TEs must ensure that each student has successfully logged out of the system. Then the TAs or TEs must collect and send for secure shredding any handouts or scratch paper that students used during the assessment.

2.2.3 Paper-Pencil Test Administration

The paper-pencil versions of the Smarter Balanced ELA/L and mathematics assessments are provided as an accommodation for students who do not have access to a computer and students who are visually impaired. For Connecticut, paper-pencil tests were offered only in braille and large print.

The DA must order the accommodated test materials on behalf of the students who need to take the paper-pencil test via TIDE. Based on the paper-pencil orders submitted in TIDE, the testing contractor ships the appropriate test booklets and the *Paper-Pencil Test Administration Manual* to the district.

Separate test booklets are used for ELA/L and mathematics assessments. The items from the CAT and the PT components are combined into one test booklet, including two sessions for CAT and one session for PTs in both content areas. The TEs and TAs are asked not to administer the ELA PT on the paper-pencil test.

After the student has completed the assessments, the TEs and TAs enter the student responses into the Data Entry Interface (DEI) and return the test booklets to the testing vendor. The tests submitted via the DEI are then scored.

The total number of students who took paper-pencil tests is presented in Table 3. Please note that students who took the paper-pencil tests took the test in the classroom.

Table 3. Number of Students Who Took Paper-Pencil Tests in the 2021–2022 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
ELA/L	3	4		8	6	6	27
Mathematics	2	4		7	6	6	25

2.2.4 Braille Test Administration

The adaptive braille test was available with the same test blueprint in English in both ELA/L and mathematics. In the 2017–2018 test administration, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a CAT segment, and a fixed-form PT. The fixed-form segment includes items with tactile graphics which can be embossed at the testing location or received as a package of pre-embossed materials through the CSDE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described as follows:

- The braille interface includes a text-to-speech component for mathematics consistent with the read-aloud assessment accommodation. The Job Access with Speech (JAWS) screen-reading software provided by Freedom Scientific is an essential component that students use with the braille interface.

- Mathematics items are presented to students in the Nemeth Braille Code for Mathematics via a braille embosser through the online CAT and a fixed-form PT.
- Students taking the summative ELA/L assessment can emboss both reading passages and items as they progress through the assessment. If a student has an RBD, a 40-cell RBD is recommended. The summative ELA/L is presented to the student with items in either contracted or un-contracted literary braille (for items containing only text) and via a braille embosser (for items with tactile or spatial components that cannot be read by an RBD).

Before administering the online summative assessments using the braille interface, TEs or TAs must ensure that the technical requirements are met. These requirements apply to the student’s computer, the TE’s or TA’s computer, and any supporting braille technologies used in conjunction with the braille interface.

2.3 TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

All DAs, DTCs, and STCs oversee all aspects of testing at their schools and serve as the main points of contact, and TEs and TAs administer the online assessments. The online CAI TA Certification Course, webinars, user guides, manuals, and training sites are used to train TEs and TAs about the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration are provided online.

2.3.1 Online Training

Multiple online training opportunities are offered to key staff.

TA Certification Course

CAI’s online TA Certification Course is available as an optional course to any user in TIDE. This web-based course is about 30–45 minutes long and covers information on testing policies and steps for administering a test session in the online system. The course is interactive, requiring participants to start test sessions under different scenarios. Throughout the training and at the end of the course, participants are required to answer multiple-choice questions about the information provided.

Office Hour Webinars

During the testing window, the CSDE and CAI held office hours every Thursday from 3:00 p.m.–4:00 p.m. During office hours, the CSDE and CAI staff provided brief, weekly assessment updates and were available for phone support to answer any questions from districts. All office hour sessions were recorded, and the recordings were posted to the portal.

Practice and Training Test Site

In January 2015, separate practice and training sites were opened for TEs/TAs and students, and these sites were refreshed before the 2021–2022 school year. TEs and TAs can practice administering assessments and starting and ending test sessions on the TA Training Site. Students can practice taking an online assessment on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments for ELA/L and mathematics. Each test provides students with a grade-specific testing experience, including a variety of item types and levels of difficulty (approximately 30 items each in ELA/L and mathematics), as well as an opportunity to practice the PT.

The training tests are designed to provide students and teachers with opportunities to quickly familiarize themselves with the software and navigational tools they will use for the upcoming Smarter Balanced assessments for ELA/L and mathematics. Training tests are available for both ELA/L and mathematics, and the tests are organized by grade bands (grades 3–5, grades 6–8, and grade 11), with each test containing 5–10 items.

A student can log in directly to the practice and training test site as a guest without a TA-generated test session ID, or the student can log in through a training test session created by the TE or TA in the TA Training Site. The student training test includes all item types in the operational item pool, including multiple-choice items, grid items, and natural language items. Teachers can also use these training tests to help students become familiar with the online platform and item types.

Manuals and User Guides

The following manuals and user guides are available on the Connecticut portal: <https://ct.portal.cambiumast.com/>.

The *Test Coordinator Manual* provides information for DCs and STCs regarding policies and procedures for the 2021 Smarter Balanced assessments in ELA/L and mathematics.

The *Smarter Balanced Summative Assessment Test Administration Manual* provides information for TEs and TAs administering the Smarter Balanced online summative assessments in ELA/L and mathematics. It includes screen captures and step-by-step instructions on how to administer the online tests.

The *Assistive Technology Manual* provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with specific accessibility needs complete online tests in the Test Delivery System (TDS). It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration in order to work with TDS.

The technology resource manuals contain technology requirements and instructions that will assist technology coordinators in preparing computers and devices for online testing. A guide is created for each of the approved operating systems (Windows, Mac, iPad, Linux, ChromeOS).

The *Centralized Reporting System User Guide* provides information about the reporting system, including instructions for viewing score reports, accessing test management resources, creating and editing rosters, and searching for students for interim and summative assessments.

The *Test Administrator User Guide* is designed to help users navigate the TDS, including the Student Interface and the TA Interface, and help TEs/TAs manage and administer online testing for students.

The *Assessment Viewing Application User Guide* provides an overview of how to access and use the Assessment Viewing Application (AVA). AVA allows teachers to view items on the Smarter Balanced interim assessments.

The *Test Information Distribution Engine (TIDE) User Guide* is designed to help users navigate TIDE. Users can find information on managing user account information, managing student test settings, appeals, and rosters.

All manuals and user guides pertaining to the 2021–2022 online testing are available on the portal, and DAs, DTCs, and STCs used the manuals and user guides to train TAs and TEs in test administration policies and procedures.

Brochures and Quick Guides

The following brochures and quick guides are available on the Connecticut portal, <https://ct.portal.cambiumast.com/>.

Accessing Participation Reports: This brochure provides instructions for how to extract participation reports for the Smarter Balanced assessments.

Accessing TIDE: This brochure provides a brief overview of user management in TIDE and how to log in to the system. School personnel will need to use TIDE account credentials to access all secure online systems used to administer Connecticut Comprehensive Assessment Program online assessments.

Embedded and Non-Embedded Designated Supports for English Learners: This brochure provides recommendations for students who are English learners (ELs) on what supports they may benefit from when participating in the Connecticut statewide assessments. These designated supports are intended as a language support for students who have limited English language skills, whether or not they are identified in the Public School Information System (PSIS) as EL or EL with a disability. The use of these supports may result in the student needing additional overall time to complete the assessment.

How to Access the Data Entry Interface (DEI): This brochure describes how to access the DEI to submit the Smarter Balanced paper-pencil tests.

How to Activate a Test Session for the Interim Assessments: This document provides a step-by-step guide on how to start a test session for the Smarter Balanced interim assessments, including the interim assessment blocks (IABs). It includes a complete list of all interim test labels as they appear in the TA Interface.

Managing Student Test Settings Brochure: This brochure provides a brief overview on how to manage student test settings in TIDE. Students' embedded accommodations, non-embedded accommodations, and designated supports must be set in TIDE prior to test administration for these settings to be reflected in the TDS.

Monitoring Test Progress: Test Status Code Report and Test Completion Rates: This brochure contains instructions for generating Test Status Code Reports and Test Completion Rates in TIDE. These are excellent tools that should be used to track test completion for students at both the district and school level.

User Role Permissions for Online Systems Brochure: This brochure outlines the user roles and permissions for each secure online testing system used to administer the online assessments for the Connecticut Comprehensive Assessment Program. These systems include TIDE, the Centralized Reporting System (CRS), TA Interface, DEI, Teacher Hand Scoring System (THSS), and AVA.

Understanding and Creating Rosters: This document provides instructions for how to create, view, and modify rosters in TIDE and in the CRS. Rosters are groups of students associated with a teacher in a particular school. Rosters typically represent entire classrooms in lower grades, or individual classroom periods in upper grades.

2.3.2 District Test Coordinator Training Workshops

DTC training workshops were held on January 19–21, 2022. Five remote training sessions were held during this period. Training was provided for the administration of the Smarter Balanced assessments for ELA/L and mathematics. During the training, DTCs were provided with information to support training of the STCs, TEs, and TAs.

2.4 TEST SECURITY

All test items, test materials, and student-level testing information are considered secure materials for all assessments. The importance of maintaining test security and the integrity of test items is stressed throughout the webinar trainings and in the user guides, modules, and manuals. Features in the testing system also protect test security. This section describes system security, student confidentiality, and policies on testing improprieties.

2.4.1 Student-Level Testing Confidentiality

All secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal laws. Secure transmission and password-protected access are basic features of the current system and permit authorized data access only. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. In addition, CAI's systems use role-based security models that ensure that users access only the data to which they are entitled and may edit data according to their user rights only.

There are three dimensions related to identifying that students are accessing appropriate test content:

1. *Test eligibility* refers to the assignment of a test to a particular student.
2. *Test accommodation* refers to the assignment of a test setting to specific students based on needs.
3. *Test session* refers to the authentication process of a TE/TA creating and managing a test session, the TE/TA reviewing and approving a test (and its settings) for every student, and the student signing on to take the test.

FERPA prohibits public disclosure of student information or test results. The following are examples of prohibited practices:

- Providing login information (username and password) to other authorized TIDE users or to unauthorized individuals
- Sending a student's name and SSID number together in an email message; if information must be sent via email or fax, include only the SSID number, not the student's name
- Having students log in and test under another student's SSID number

Test materials and score reports should not be exposed to identify student names with test scores except by authorized individuals with an appropriate need to know.

All students, including home-schooled students, must be enrolled or registered at their testing schools in order to take the online, paper-pencil, or braille assessments. Student enrollment information, including

demographic data, is generated using a CSDE file and uploaded nightly via a secure file transfer site to the online testing system during the testing period.

Students log in to the online assessment using their legal first name, SSID number, and a test session ID. Only students can log in to an online test session. TEs/TAs, proctors, or other personnel are not permitted to log in to the system on behalf of students, although they are permitted to assist students who need help logging in. For the paper-pencil versions of the assessments, TEs and TAs are required to affix the student label to the student's answer document.

After a test session, only staff with the administrative roles of DA, DTC, STC, or TE can view their students' scores. TAs do not have access to student scores.

2.4.2 System Security

The objective of system security is to ensure that all data are protected and accessed appropriately by the designated user groups. It is about protecting data and maintaining data and system integrity as intended, including ensuring that all personal information is secured, that transferred data (whether sent or received) is not altered in any way, that the data source is known, and that any service can only be performed by a specific, designated user.

A hierarchy of control: As described in Section 2.2, all DAs, DTCs, STCs, TAs, and TEs have defined roles and levels of access to the testing system. When the TIDE testing window opens, the CSDE provides a verified list of DAs to the testing contractor, who uploads the information into TIDE. DAs are then responsible for selecting and entering the DTCs' and STCs' information into TIDE, and the STC is responsible for entering TA and TE information into TIDE. Throughout the year, the DA, DTC, and STC are also expected to delete information in TIDE for any staff members who have transferred to other schools, resigned, or no longer serve as TAs or TEs.

Password protection: All access points by different roles at the state, district, school principal, and school staff levels require a password to log in to the system. Newly added STCs, TAs, and TEs receive separate passwords through their personal email addresses assigned by the school.

CAI Secure Browser: A key role of the TC is to ensure that the CAI Secure Browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the CAI Secure Browser prevents students from accessing other computers or Internet applications and from copying test information. The CAI Secure Browser suppresses access to commonly used browsers, such as Internet Explorer and Firefox, and prevents students from searching for answers on the Internet or communicating with other students. The assessments can be accessed only through the CAI Secure Browser and not by other Internet browsers.

2.4.3 Security of the Testing Environment

The STCs, TEs, and TAs work together to determine appropriate testing schedules based on the number of computers available, the number of students in each tested grade, and the average amount of time needed to complete each assessment.

Testing personnel are reminded in the online training and user manuals that assessments should be administered in testing rooms that do not crowd students. Good lighting, ventilation, and freedom from noise and interruption are important factors to consider when selecting testing rooms.

TEs and TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish more quickly than others. If students are allowed to leave the testing room when they finish, TEs or TAs are required to explain the procedures for leaving and where students are expected to report once they leave without disrupting others. If students are expected to remain in the testing room until the end of the session, TEs or TAs are encouraged to prepare some quiet work for students to do after they finish the assessment.

If a student needs to leave the room for a brief time during testing, the TAs or TEs are required to pause the student’s assessment. For the CAT, if the pause lasts longer than 20 minutes, the student can continue with the rest of the assessment in a new test session, but the system will not allow the student to return to the items answered before the pause. This measure is implemented to prevent students from using the time outside of the testing room to look up answers.

Room Preparation

The room should be prepared prior to the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to help answer test items should be removed or covered. This rule applies to rubrics, vocabulary charts, student work, posters, graphs, content-area strategies charts, and other materials. The cell phones of both testing personnel and students must be turned off and stored in the testing room out of sight. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances in order to promote optimum testing conditions; they should also post “TESTING—DO NOT DISTURB” signs on the doors of testing rooms.

Seating Arrangements

TEs and TAs should provide adequate space between students’ seats. Students should be seated so that they will not be tempted to look at the answers of others. Because the online CAT is adaptive, it is unlikely that students will see the same test items as other students; however, through appropriate seating arrangements, students should be discouraged from communicating with each other. For the PTs, different forms are distributed throughout a classroom so that students receive different forms of the PTs.

After the Test

At the end of the test session, TEs or TAs must walk through the classroom to pick up any scratch paper that students used and any papers that display students’ SSID numbers and names together. These materials should be securely shredded or stored in a locked area immediately. The printed reading passages and items for any content-area assessment provided for a student allowed to use this accommodation in an individual setting must also be shredded immediately after a test session ends.

For the paper-pencil versions, specific instructions on how to package and secure the test booklets to be returned to the testing contractor’s office are provided in the *Paper and Pencil Test Administration Manual*.

2.4.4 Test Security Violations

Everyone who administers or proctors the assessments is responsible for understanding the security procedures for administering them. Prohibited practices as detailed in the *Smarter Balanced Online Summative Test Administration Manual* are categorized into three groups:

Impropriety: This is a test security incident that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (e.g., students leaving the testing room without authorization).

Irregularity: This is a test security incident that impacts an individual or group of students who are testing and may potentially affect student performance on the test, test security, or test validity. These circumstances can be contained at the local level (e.g., disruption during the test session, such as a fire drill).

Breach: This is a test security incident that poses a threat to the validity of the test. Breaches require immediate attention and escalation to the CSDE. Examples may include such situations as exposure of secure materials or a repeatable security/system risk. These circumstances have external implications (e.g., administrators modifying student answers or students sharing test items through social media).

District and school personnel are required to document all test security incidents in the test security incident log. The log serves as the document of record for all test security incidents and should be maintained at the district level and submitted to the CSDE at the end of testing.

2.5 STUDENT PARTICIPATION

All students (including retained students) currently enrolled in grades 3–8 at public schools in Connecticut are required to participate in the Smarter Balanced assessments. Students must be tested in the enrolled grade assessment; out-of-grade-level testing is not allowed for the administration of Smarter Balanced assessments.

2.5.1 Home-Schooled Students

Students who are home-schooled may participate in the Smarter Balanced assessments at the request of their parent or guardian. Schools must provide these students with one testing opportunity for each relevant content area, if requested.

2.5.2 Exempt Students

Students who have a significant medical emergency are exempt from participating in the Smarter Balanced assessments.

2.6 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

The Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines* (UAA Guidelines) are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) and Section 504 Plan teams, as they prepare for and implement the Smarter Balanced assessments. The UAA Guidelines provide information for classroom teachers, English language development educators, special education teachers, and instructional assistants to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The UAA Guidelines are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The *Connecticut Assessment Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. They focus on universal tools, designated supports, and accommodations

for the Smarter Balanced assessments of ELA/L and mathematics. At the same time, the UAA Guidelines support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

The summative assessments contain universal tools, designated supports, and accommodations in both embedded and non-embedded versions. Embedded resources are part of the computer administration system, whereas non-embedded resources are provided outside of that system.

State-level users, DTCs, and STCs have the ability to set embedded and non-embedded designated supports and accommodations based on their specific user role. Designated supports and accommodations must be set in TIDE before starting a test session.

All embedded and non-embedded universal tools will be activated for use by all students during a test session. One or more of the pre-selected universal tools can be deactivated by a TE/TA in the TA Interface of the testing system for a student who may be distracted by the ability to access a specific tool during a test session.

For additional information about the availability of designated supports and accommodations, refer to the Connecticut’s Assessment Guidelines for complete information at this URL: <https://ct.portal.cambiumast.com/resources/guides/csde-assessment-guidelines>.

2.6.1 Online Universal Tools for All Students

Universal tools are access features of an assessment or exam that are embedded or non-embedded components of the test administration system. Universal tools are available to all students based on their preference and selection and have been pre-set in TIDE. In the 2021–2022 test administration, the following features of universal tools were available for *all* students to access. For specific information on how to access and use these features, refer to the *Test Administrator User Guide* at this URL: <https://ct.portal.cambiumast.com>.

Embedded Universal Tools

Breaks: The student can pause and resume the assessment. However, if an assessment is paused for more than 20 minutes, students will not be allowed to return to previous test items.

Calculator: An embedded on-screen digital calculator can be accessed for calculator-allowed items when students click the calculator button. This tool is available only with the specific items for which the Smarter Balanced item specifications indicate that it would be appropriate.

Digital Notepad: This tool is used for making notes about an item. The digital notepad is item-specific and available through the end of the test segment. Notes are not saved when the student moves on to the next segment or after a break of more than 20 minutes.

English Glossary: Grade- and context-appropriate definitions of specific construct-irrelevant terms are shown in English on the screen via a pop-up window. The student can access the embedded glossary by clicking on any of the pre-selected terms.

Expandable Passages/Stimuli/Items: Each passage or stimulus can be expanded so that it takes up a larger portion of the screen.

Highlighter: This tool is used to highlight passages or sections of passages and test items.

Keyboard Commands: Navigation throughout text can be accomplished by using a keyboard.

Line Reader: Students can use the line reader tool to assist in reading by raising and lowering the tool for each line of text on the screen.

Mark for Review: Students can mark an item to return to later during testing. However, for the CAT, if the assessment is paused for more than 20 minutes, students will not be allowed to return to marked test items.

Mathematics Tools: These digital tools (e.g., embedded ruler, embedded protractor) are used for measurements related to mathematics items. They are available only with the specific items for which the Smarter Balanced item specifications indicate that one or more of these tools would be appropriate.

Strikethrough: This tool allows users to cross out response options. If the response option is an image, a strikethrough line will not appear, but the image will be grayed out.

Writing Tools (for interim ELA/L performance tasks): Selected writing tools (e.g., bold, italic, bullets, undo/redo) are available for all student-generated responses.

Zoom: Students can zoom in and zoom out on test items, text, or graphics.

Non-Embedded Universal Tools

Breaks: Breaks may be given at predetermined intervals or after completion of sections of the assessment for students taking a paper-pencil test. Sometimes, students are allowed to take breaks when individually needed in order to reduce cognitive fatigue when they experience heavy assessment demands. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Scratch Paper/White Board with Marker: Scratch paper to make notes, write computations, or record responses may be made available. Only plain paper or lined paper is appropriate for ELA/L. Graph paper is required beginning in grade 6 and can be used on all mathematics assessments. A student can use an assistive technology device for scratch paper as long as the device is consistent with the child's IEP and acceptable to the CSDE.

2.6.2 Designated Supports and Accommodations

Designated supports for the Smarter Balanced assessments are features that are available for use by any student for whom the need has been indicated by an educator (or team of educators with parent/guardian and student). Scores achieved by students using designated supports will be included for federal accountability purposes. It is recommended that a consistent process be used to determine these supports for individual students. All educators making these decisions should be trained on the process and should understand the range of designated supports available. Smarter Balanced Assessment Consortium members have identified digitally embedded and non-embedded designated supports for students for whom an adult or team has indicated a need for the support.

Accommodations are changes in procedures or materials that increase equitable access during the Smarter Balanced assessments. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations are available for students with documented IEPs or Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standard, or intended outcome of the assessments.

Embedded Designated Supports

Color Contrast: Students can adjust screen background or font color, based on student needs or preferences. This may include reversing the colors for the entire interface or choosing the color of font and background. Black on white, reverse contrast, black on rose, medium gray on light gray, and yellow on blue were offered for the online assessments.

Illustration Glossary: The illustration glossaries are provided for selected construct-irrelevant terms for math. Illustrations for these terms appear on the computer screen when students select them. Students with the illustration glossary setting enabled can view the illustration glossary. Students can also adjust the size of the illustration and move it around the screen.

Masking: Masking involves blocking off content that is not of immediate need or that may be distracting to the student. Students can focus their attention on a specific part of a test item by using the masking feature.

Mouse Pointer: This embedded support allows the mouse pointer to be set to a larger size and/or for the color of the mouse pointer to be changed. A TA sets the size and color of the mouse pointer prior to testing.

Print Size Online: This tool allows the font size viewed by the student in the TDS to be pre-set for the entire test. This support is generally most beneficial for students with visual disabilities. Selections are entered in the TIDE system prior to testing.

Streamline: This accommodation provides a streamlined interface of the test in an alternate, simplified format in which the items are displayed below the stimuli.

Text-to-Speech (for mathematics stimuli items and ELA/L items): Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed of the voice and raise or lower the volume of the voice via a volume control.

Translated Test Directions (for mathematics): Translation of test directions is a language support available prior to beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically part of the stacked translation designated support.

Translations (glossaries) (for mathematics): Translated glossaries are a language support provided for selected construct-irrelevant terms for mathematics. Translations for these terms appear on the computer screen when students click on them. The following language glossaries were offered: Arabic, Burmese, Cantonese, Filipino, Hmong, Korean, Mandarin, Punjabi, Russian, Somali, Spanish, Ukrainian, and Vietnamese.

Translations (Spanish-stacked) (for mathematics): Stacked translations are a language support available for some students. They provide the full translation of each test item above the original item in English.

Turn Off Any Universal Tools: Teachers can disable any universal tools that might be distracting, that students do not need to use, or that students are unable to use.

Non-Embedded Designated Supports

Amplification: The student adjusts the volume control beyond the computer’s built-in settings using headphones or other non-embedded devices.

Color Contrast: Test content of online items may be printed with different colors.

Color Overlay: Color transparencies may be placed over a paper-pencil assessment.

Illustration Glossary: The illustration glossaries are a language support provided for selected construct-irrelevant terms for math. Illustrations for these terms appear in a supplement to the paper/pencil test and are identified by item number.

Magnification: The size of specific areas of the screen (e.g., text, formulas, tables, graphics, navigation buttons) may be adjusted by the student with an assistive technology device. Magnification allows the student to increase the size of test content to a level not allowed by the zoom universal tool.

Noise Buffers: These include ear mufflers, white noise, and/or other equipment to reduce environmental noise.

Read-Aloud (for mathematics items and ELA/L items but not reading passages): Text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual* and the *Guidelines for Read Aloud, Test Reader*. All or portions of the content may be read aloud.

Read-Aloud in Spanish (for mathematics): Spanish text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Test Administration Manual* and the read-aloud guidelines. All or portions of the content may be read aloud.

Separate Setting: Test location is altered so that the student is tested in a setting different from that which is available for most students.

Simplified Test Directions: The TA simplifies or paraphrases the test directions found in the *Test Administration Manual* according to the Simplified Test Directions guidelines.

Translated Test Directions: The TA uses a PDF file of directions translated in each of the languages currently supported. A bilingual adult can read the file to the student.

Translations (glossaries) (for mathematics paper-pencil tests): Translated glossaries are a language support provided for selected construct-irrelevant terms for mathematics. Glossary terms are listed by item and include the English term and its translated equivalent.

Embedded Accommodations

American Sign Language (ASL) (for ELA/L listening items and mathematics items): Test content is translated into ASL video. An ASL human signer and the signed test content are viewed on the same screen. Students may view portions of the ASL video as often as needed.

Braille: This is a raised-dot code that individuals read with their fingertips. Graphic material (e.g., maps, charts, graphs, diagrams, illustrations) is presented in a raised format (paper or thermoform). Contracted and non-contracted braille is available, and Nemeth Code is available for mathematics.

Braille Transcript: A braille transcript of the closed captioning is available for the listening passages of the ELA/L assessment in the following braille codes: English Braille, American Edition (EBAE) uncontracted; and EBAE contracted.

Closed Captioning (for ELA/L listening stimuli items): This is printed text that appears on the computer screen as audio materials are presented.

Speech-to-Text: This tool allows students to dictate their responses into an open text box.

Text-to-Speech (ELA/L reading passages): Text is read aloud to the student via embedded text-to-speech technology. The student can control the speed of the voice and raise or lower the volume of the voice via a volume control.

Non-Embedded Accommodations

100s Number Table: This is a paper-based list of all the digits from 1 through 100 in table format.

Abacus: This tool may be used in place of scratch paper for students who typically use an abacus.

Alternate Response Options: Alternate response options include but are not limited to an adapted keyboard, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touch screen, head wand, and switches.

Calculator (for grades 6-8 mathematics tests): When the embedded Desmos calculator or specialized calculator is inaccessible, the provision of a hand-held calculator may be appropriate: either a basic calculator (grade 6) or a scientific calculator (grades 7-8).

Human Signer: This sign language accommodation allows a qualified test administrator to sign or provide visual language support for the test directions, test content, and/or reading passages to a student who is deaf or hard of hearing.

Math Manipulatives: These tools are available to allow eligible students to use concrete mathematical tools strategically to support their decision making. Students eligible for this accommodation typically have visual or math-related disabilities.

Multiplication Table: This is a paper-based single digit (1–9) multiplication table students use for reference.

Paper Tests (large print and braille): Paper tests are available in large print and braille for students who need these accommodations in paper format.

Print-on-Demand: Paper copies of passages, stimuli, and/or items are printed for students. For those students who need a paper copy of a passage or stimulus, permission for the students to request printing must first be set in TIDE.

Read-Aloud (for ELA/L passages): Text is read aloud to the student via an external screen reader or by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual* and *Read Aloud Guidelines*. All or portions of the content may be read aloud. Members can refer to the *Guidelines for Choosing the Read Aloud Accommodation* when deciding if this accommodation is appropriate for a student.

Scribe: Students dictate their responses to a human who records what they dictate verbatim. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

Specialized Calculator (for grades 6–8 mathematics tests): A non-embedded calculator may be provided for students who need a special calculator, such as a braille calculator or a talking calculator that is currently unavailable within the assessment platform.

Speech-to-Text: Voice recognition allows students to use their voices as devices to input information into the computer to dictate responses or give commands (e.g., opening application programs, pulling down menus, saving work). Voice recognition software generally can recognize speech up to 160 words per minute. Students may use their own assistive technology devices.

Table 4 presents a list of universal tools, designated supports, and accommodations that were offered in the 2021–2022 administration. Tables 5–10 provide the number of students who utilized any of the offered accommodations and designated supports.

Table 4. 2021–2022 Universal Tools, Designated Supports, and Accommodations

Universal Tools	Designated Supports	Accommodations
Embedded		
Breaks Calculator ¹ Digital Notepad English Glossary Expandable Passages/Stimuli/Items Highlighter Keyboard Commands Line Reader Mark for Review Mathematics Tools ² Strikethrough Writing Tools ³ Zoom	Color Contrast Illustration Glossary ⁴ Masking Mouse Pointer Print Size Online Streamline Text-to-Speech ⁵ Translated Test Directions ⁴ Translations (Glossary) ⁴ Translations (Spanish-Stacked) ⁶ Turn Off Any Universal Tools	American Sign Language ⁷ Braille Braille Transcripts ⁸ Closed Captioning ⁸ Speech-to-Text Text-to-Speech ⁹
Non-Embedded		
Breaks Scratch Paper/White Board	Amplification Color Contrast Color Overlay Illustration Glossary ⁴ Magnification Noise Buffers Read Aloud ¹⁰ Read Aloud in Spanish ⁵ Separate Setting Simplified Test Directions Translated Test Directions Translations (Glossary) ⁵	100s Number Table Abacus Alternate Response Options ¹¹ Calculator ¹ Human Signer Math Manipulatives Multiplication Table ⁵ Paper Tests (Large Print and Braille) Print-on-Demand Read Aloud ¹² Scribe Specialized Calculator ¹ Speech-to-Text

Note: Items shown are available for ELA/L and mathematics unless otherwise noted.

¹ For calculator-allowed items only in grades 6–8

² Includes embedded ruler, embedded protractor

³ For interim ELA/L performance tasks; includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo

⁴ For mathematics items

⁵ For ELA/L PT stimuli, ELA/L PT and CAT items (not ELA/L CAT reading passages), and mathematics stimuli and items; must be set in TIDE before test begins

⁶ For mathematics test

⁷ For ELA/L listening items and mathematics items

⁸ For ELA/L listening items

⁹ For ELA/L reading passages; must be set in TIDE by state-level user

¹⁰ For ELA/L items (not ELA/L reading passages) and mathematics items

¹¹ Includes adapted keyboards, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touch screen, head wand, and switches

¹² For ELA/L reading passages, all grades

Table 5. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	5	8	3	9	5	5
Braille	1	1	1	1	2	
Braille Transcripts	3	3	1	1	1	
Closed Captioning	30	36	39	55	37	41
Speech-to-Text	295	448	373	268	182	143
Text-to-Speech: Passages and Items	1,378	1,353	1,422	1,257	1,157	1,118
Non-Embedded Accommodations						
Alternate Response Options	7	10	9	5	5	3
Braille Paper				1		1
Large Print Paper				1		1
Speech-to-Text	71	119	123	105	83	68

Table 6. ELA/L Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	8	7	5	14	7	5
	LEP	1					
	IDEA	4	4	2	7	4	2
Masking	Overall	171	205	178	201	164	182
	LEP	54	52	31	52	48	53
	IDEA	125	146	131	110	123	124
Mouse Pointer	Overall	1	1	3	1		3
	LEP	1		1			1
	IDEA		1	1	1		2
Print Size Online	Overall	80	82	87	41	29	35
	LEP	44	45	40	2	2	5
	IDEA	32	31	31	28	16	19
Streamline	Overall	175	159	167	189	142	126
	LEP	30	30	24	23	15	16
	IDEA	135	125	132	127	123	107
Text-to-Speech: Items	Overall	7,560	7,746	7,336	6,142	5,913	5,562
	LEP	3,173	3,123	2,680	1,736	1,770	1,519
	IDEA	2,227	2,675	2,674	2,619	2,448	2,244

Table 7. ELA/L Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	3	6	3	4	4	5
	LEP				1		1
	IDEA	3	1	3	3	3	3
Color Overlay	Overall	1	6	6	5	4	4
	LEP	1					
	IDEA	1	6	6	4	4	2
Magnification	Overall	13	6	14	19	17	14
	LEP		2	2	5	2	1
	IDEA	5	5	9	14	15	10
Noise Buffers	Overall	6	5	7	13	9	14
	LEP		1		1		
	IDEA	4	2	5	9	5	9
Read-Aloud Items	Overall	136	93	98	38	62	79
	LEP	66	52	49	21	36	36
	IDEA	85	55	54	20	35	48
Separate Setting	Overall	3,759	4,243	4,400	3,763	3,693	3,718
	LEP	814	848	804	519	488	486
	IDEA	2,840	3,230	3,365	3,039	2,993	3,004
Simplified Test Directions	Overall	1,067	1,123	993	825	897	775
	LEP	393	431	343	260	321	258
	IDEA	706	816	789	684	738	656
Translated Test Directions	Overall	71	97	112	161	187	180
	LEP	71	93	110	160	186	178
	IDEA	7	9	16	19	20	16

Table 8. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
Embedded Accommodations						
American Sign Language	5	7	2	8	5	5
Braille	1	1	1	1	2	
Speech-to-Text	282	419	360	247	159	134
Non-Embedded Accommodations						
100s Number Table	1,566	1,457	1,186	802	547	365
Abacus	2	2	2	1	5	1
Alternate Response Options	6	11	8	4	5	3
Braille Paper				1		1
Large Print Paper						1
Multiplication Table		2,595	3,089	3,065	2,850	2,615
Specialized Calculator	3	9	14	52	66	99
Speech-to-Text	65	110	121	92	73	62

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	9	6	6	14	6	5
	LEP	1					
	IDEA	5	3	3	7	3	2
Illustration Glossary	Overall	1,466	1,441	1,152	1,113	1,021	875
	LEP	1,437	1,423	1,129	1,092	1,011	864
	IDEA	154	202	151	147	153	144
Masking	Overall	165	204	178	174	170	175
	LEP	51	52	30	52	48	52
	IDEA	122	146	130	110	120	124
Mouse Pointer	Overall	1	1	3	1		3
	LEP	1		1			1
	IDEA		1	1	1		2
Print Size Online	Overall	78	81	86	42	29	37
	LEP	44	45	40	2	2	5
	IDEA	31	29	30	29	16	21
Streamline	Overall	172	158	166	142	136	122
	LEP	32	31	23	22	15	16
	IDEA	131	123	131	121	118	103
Text-to-Speech: Stimuli and Items	Overall	9,340	9,497	9,143	7,481	7,130	6,706
	LEP	3,470	3,412	3,006	2,031	2,051	1,769
	IDEA	3,613	4,018	4,105	3,856	3,579	3,324
Translations (Glossary): Spanish	Overall	878	890	811	911	887	762
	LEP	871	884	796	895	874	753
	IDEA	87	101	86	136	126	111
Translations (Glossary): Other Languages	Overall	62	64	39	49	35	38
	LEP	59	64	38	48	34	37
	IDEA		5	1	4		2
Translations (Spanish-Stacked)	Overall	522	503	531	555	587	617
	LEP	517	493	521	548	582	612
	IDEA	27	40	35	34	30	45

Table 10. Mathematics Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	4	5	3	3	5	5
	LEP	1					1
	IDEA	3		3	3	4	3
Color Overlay	Overall	1	5	6	5	4	4
	LEP	1					
	IDEA	1	5	6	4	4	2
Illustration Glossary	Overall	59	51	52	35	30	28
	LEP	59	51	52	34	30	28
	IDEA	6	8	8	7	5	5
Magnification	Overall	13	5	14	15	18	14
	LEP		2	2	5	2	2
	IDEA	5	4	10	10	16	10
Noise Buffers	Overall	6	5	7	14	9	13
	LEP	1	1		1		
	IDEA	4	2	5	9	5	8
Read Aloud Stimuli and Items	Overall	174	137	113	45	63	76
	LEP	87	67	57	25	29	30
	IDEA	91	68	65	26	42	48
Read Aloud Stimuli and Items (Spanish)	Overall	45	40	26	22	44	22
	LEP	41	38	25	20	40	22
	IDEA	10	8	6	7	15	6
Separate Setting	Overall	3,795	4,297	4,466	3,814	3,707	3,714
	LEP	816	856	808	533	499	480
	IDEA	2,878	3,269	3,421	3,080	3,018	3,008
Simplified Test Directions	Overall	1,106	1,156	1,032	876	929	814
	LEP	386	433	350	272	324	271
	IDEA	754	850	827	725	773	686
Translated Test Directions	Overall	75	88	92	161	164	176
	LEP	75	84	91	160	163	174
	IDEA	7	9	15	19	19	16
Translations (Glossary): Spanish	Overall	71	94	98	95	94	104
	LEP	70	92	85	95	94	104
	IDEA	7	13	6	17	9	9
Translations (Glossary): Other Languages	Overall	9	8	6	21	6	12
	LEP	8	8	4	20	5	12
	IDEA	1			1		

2.7 TESTING TIME

The online environment also allows item response time to be captured as the item page time (the time each item page is presented) in milliseconds. For discrete items, each item appears on the screen one item at a time, whereas stimulus-based items appear on the screen together. The page time is the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items.

For each student, the total time taken to complete the test is computed by adding up the page time for all items and item groups (stimulus-based items).

The Smarter Balanced summative assessments are not timed, and an individual student may need more or less testing time overall. The length of a test session is determined by TEs/TAs who are knowledgeable about the class periods in the school’s instructional schedule and the timing needs associated with the assessments. Students should be allowed extra time if they need it, but TEs/TAs must use their best professional judgment when allowing students extra time. Students should be actively engaged in responding productively to test items.

Tables 11 and 12 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 11. ELA/L Testing Times

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time by Percentile (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test (CAT Component)							
3	1:39	0:53	1:58	2:07	2:18	2:34	3:08
4	1:45	0:54	2:06	2:15	2:26	2:42	3:14
5	1:43	0:47	2:04	2:13	2:23	2:39	3:07
6	1:48	0:48	2:10	2:19	2:31	2:46	3:14
7	1:41	0:46	2:03	2:11	2:22	2:37	3:03
8	1:36	0:42	1:56	2:03	2:13	2:27	2:50

Table 12. Mathematics Testing Times

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time by Percentile (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	1:55	0:58	2:22	2:33	2:47	3:07	3:42
4	2:02	1:02	2:30	2:42	2:58	3:19	3:57
5	2:11	1:05	2:42	2:55	3:11	3:33	4:12
6	2:03	0:58	2:31	2:41	2:55	3:16	3:50
7	1:50	0:52	2:15	2:24	2:36	2:54	3:25
8	1:51	0:52	2:18	2:28	2:40	2:56	3:24
CAT Component							
3	1:21	0:43	1:41	1:49	1:59	2:14	2:42
4	1:29	0:47	1:50	1:59	2:11	2:28	2:57
5	1:29	0:44	1:51	1:59	2:11	2:25	2:52
6	1:25	0:41	1:44	1:52	2:01	2:16	2:41
7	1:24	0:41	1:44	1:51	2:01	2:15	2:40
8	1:24	0:41	1:45	1:52	2:01	2:15	2:36
PT Component							
3	0:34	0:21	0:43	0:48	0:53	1:00	1:13
4	0:33	0:21	0:42	0:46	0:51	0:59	1:12
5	0:42	0:29	0:53	0:59	1:06	1:17	1:35
6	0:38	0:25	0:48	0:53	0:59	1:07	1:23
7	0:25	0:18	0:33	0:36	0:40	0:47	0:58
8	0:27	0:18	0:35	0:39	0:43	0:49	1:00

2.8 DATA FORENSICS PROGRAM

The validity of test scores depends critically on the integrity of the test administrations. Any irregularities in test administration could cast doubt on the validity of the inferences based on those test scores. Multiple facets ensure that tests are administered properly, including clear test administration policies, effective TA training, and tools to identify possible irregularities in test administrations.

For online administrations, a set of quality assurance (QA) reports is generated during and after the testing window. One of the QA reports focuses on flagging possible testing anomalies. Testing anomalies are analyzed by examining changes in student performance from year to year, test taking time, item response patterns using a person-fit index, and item response change analyses.

Analyses are performed at the student level and summarized for each aggregate unit, including testing session, TA, and school. Flagging criteria used for these analyses are described below and are configurable by an authorized user. When the aggregate unit size is small, the aggregate unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is 5 or fewer students but this value is configurable. For each aggregate unit, small groups are identified based on the number of tests included in the aggregate unit from that analysis. Thus, a small unit identified in one analysis may not be a small unit in another analysis. The QA reports are provided to state clients to monitor testing anomalies throughout the testing window.

2.8.1 Changes in Student Performance

Changes in student scores between administration years are examined using a regression model to check for outliers. For these between-year comparisons, students' current-year scores are regressed on their test scores from the previous year and on the number of days between the two years' test-end dates (to control for the instruction time between the two test scores).

A large score gain or loss in student scores between administration years is detected by examining the residuals for outliers. The residuals are computed as the observed value minus the regression model's predicted value. To detect unusual residuals, the studentized residuals are computed. An unusual increase or decrease in student scores between administration years is flagged when the absolute value of the studentized residual is greater than 3.

The residuals of students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations and/or years based on the average of the residuals in the aggregate unit (e.g., testing session, TA, school). For each aggregate unit, a t value is computed and flagged when $|t|$ is greater than 3,

$$t = \frac{\sum_{i=1}^n \hat{e}_i / n}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^n \sigma^2(1 - h_{ii})}{n^2}}}$$

where s is the standard deviation of residuals in an aggregate unit; n is the number of students in an aggregate unit (e.g., testing session, TA, school), σ^2 is the MSE from the regression, and \hat{e}_i is the residual for the i th student.

The variance of average residuals in the denominator is estimated in two components, conditioning on true residual e_i , $var(E(\hat{e}_i|e_i)) = s^2$ and $E(var(\hat{e}_i|e_i)) = \sigma^2(1 - h_{ii})$. Following the law of total variance (Billingsley, 1995, p. 456),

$$var(\hat{e}_i) = var(E(\hat{e}_i|e_i)) + E(var(\hat{e}_i|e_i)) = s^2 + \sigma^2(1 - h_{ii}), \text{ hence,}$$

$$var\left(\frac{\sum_{i=1}^n \hat{e}_i}{n}\right) = \frac{\sum_{i=1}^n (s^2 + \sigma^2(1 - h_{ii}))}{n^2} = \frac{s^2}{n} + \frac{\sum_{i=1}^n (\sigma^2(1 - h_{ii}))}{n^2}.$$

2.8.2 Test-Taking Time

The summative assessments are not timed, and thus an individual student's test taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking time may indicate irregularities in test administration. An example of an unusual test-taking time is a test record for an individual who scores very well on the test even though the average time spent is far less than that required of students statewide. If students already know the answers to the items, the test-taking time may be much shorter than the test-taking time for those who have no prior knowledge of the item content. Conversely, if a TA helps students by coaching them to change their responses during the test, the testing time could be longer than expected.

The state average testing time and standard deviation are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the test-taking time is different from the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

2.8.3 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify test-takers whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test-taker has prior knowledge of some test items (or is provided answers during the exam), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. However, if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response time index might flag such a student.

The person-fit index is based on all item responses in a test. An unlikely response to a single test item may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session, TA, and school.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and Sotaridona, Pornel, and Vallejo (2003), aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the “asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05” (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using l_z for systematic flagging of aberrant response patterns. Students with l_z values less than -3 are flagged. Aggregate units are flagged with t less than -3,

$$t = \frac{\text{Average } l_z \text{ values}}{\sqrt{s^2/n}},$$

where s = standard deviation of l_z values in an aggregate unit and n = number of students in an aggregate unit. The QA report includes a list of the flagged aggregate units.

2.8.4 Item Response Change

Students are allowed to revisit items as many times as they wish within a session and may also mark items to be revisited prior to completing the session. However, excessively high rates of response change, especially high rates of item score increases (i.e., response changes from wrong to right), may indicate irregularities in test administration. For example, test administrators (TAs) could review students’ responses and either coach them to modify their responses or keep the session active and change responses themselves.

To identify irregular patterns of response change, the item score for the final response to each item and the penultimate response if one exists are examined, and the number of instances in which the item score increases are counted.

The average and standard deviation of positive item score changes are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the number of

positive item score changes is larger than the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

2.9 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

CAI is continuously improving its ability to protect testing systems from interruptions. CAI's TDS is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. The CAI architecture, described in the following paragraphs, is designed to recover from a failure of any component with little interruption. Each system is redundant, and critical student response data are transferred to a different data center each night.

CAI has developed a unique monitoring system that is extremely sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. The CAI system does, too, but it also provides warnings when any given server performs differently from its performance over the few hours prior or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing CAI to detect of potential problems, investigate them, and mitigate them. This system has enabled CAI to make adjustments and replace equipment on multiple occasions before any problems occurred.

CAI has also implemented an escalation procedure to alert clients within minutes of any disruption. The emergency alert system notifies CAI's executive and technical staff by text message, who then immediately join a call to identify and address the problem.

The following subsection describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.9.1 High-Level System Architecture

CAI's architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. The general approach, which Smarter Balanced has adopted as standard policy, is pragmatic and well supported by the system architecture.

CAI posits that any system built around an expectation of the flawless performance of computers or networks within schools and districts is bound to fail. Therefore, the system is designed to ensure that the testing results and experience respond robustly to such inevitable failures. CAI's TDS is designed to protect data integrity and prevent student data loss at every point throughout the test administration process. Fault tolerance and automated recovery are built into every component of the system.

Fault tolerance and automated recovery are built into every component of the system. The key elements of the testing system, including the data integrity processes at work at each point in the system, are described as follows.

Student Machine

Student responses are conveyed to CAI's servers in real time as students respond. Long responses, such as essays, are saved automatically at configurable intervals (usually set to one minute) so that student work is not at risk during testing.

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated

time (usually set to 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning at a later time. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying the save.
- If the system fails completely, upon logging back in the system, the student returns to the item at which the failure occurred.

In short, data integrity is preserved by confirmed saves to CAI servers and prevention of further testing if confirmation is not received.

Test Delivery Satellites

The test delivery satellites communicate with the student machines to deliver items and receive responses. Each satellite is a collection of web and database servers. Each satellite is equipped with redundant array of independent disks (RAID) systems to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server serves as a backup hub for every four satellites. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon malfunction, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described in the next subsection), with backup copies remaining on the drive arrays of the disabled satellite.

If a satellite fails, students will exit the system. The automatic recovery system enables them to log in again within seconds or minutes of the failure, without data loss. This process is managed by the hub. Data will remain on the satellites until the satellite receives notice from the demographic and history servers that the data are safely stored on those disks.

Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store that data as described earlier. This real-time backup copy remains on the hub until the hub receives a notification from the demographic and history servers that the data have reached the designated storage location.

Demographic and History Servers

The demographic and history servers store student data for the duration of the testing window. They are clustered database servers, also with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of the storage of the information, these servers notify the hub and satellites that it is safe to delete student data.

Quality Assurance System

The QA system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QA system, and any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) are flagged, and a notification immediately goes out to CAI’s psychometricians and project team.

Database of Record

The Database of Record (DOR) is the final storage location for the student data. These clustered database servers with RAID systems hold the completed student data.

2.9.2 Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

2.9.3 Other Disruption Prevention and Recovery Systems

These testing systems are designed to be extremely fault-tolerant. The systems can withstand failure of any component with little or no service interruption. This robustness is archived through redundancy. Key redundant systems are as follows:

- The system’s hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from the system’s data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level are redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching in all server cabinets.
- Data are protected by nightly backups. A full weekly backup and incremental nightly backups protect data. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they need to rerun it.

The system’s TDS is hosted in an industry-leading facility with redundant power, cooling, state-of-the-art security, and other features that protect the system from failure. The system is redundant at every component, and in the event of failure, the unique design ensures that data are always stored in at least two locations. The engineering that led to this system protects student responses from loss.

3. SUMMARY OF 2021–2022 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All Connecticut students enrolled in grades 3–8 in all public schools are required to participate in the Smarter Balanced English language arts/literacy (ELA/L) and mathematics assessments. Before the testing window opens, the state or districts send Cambium Assessment, Inc. (CAI) a student enrollment file to load into the Test Information Distribution Engine (TIDE). Using this enrollment file, the participation rates are calculated as the percentage of students who attempted the tests. Tables 13 and 14 present the participation rates in percentages for all students and by subgroups who attempted the tests. Tables 15 and 16 present the number of Connecticut students who meet attemptedness requirements for the Smarter Balanced summative scoring and reporting.

Table 13. Participation Rates in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	97.9	97.9	97.9	97.6	97.4	96.7
Female	98.5	98.4	98.5	98.3	97.9	97.1
Male	97.3	97.4	97.3	96.9	96.9	96.4
Black or African American	97.2	97.0	97.2	96.9	96.8	95.9
AmerIndian/Alaskan	98.3	100.0	98.7	99.0	98.0	96.6
Asian	98.1	97.4	98.0	97.3	98.7	98.1
Hispanic or Latino	97.6	97.8	97.7	97.2	97.2	96.4
Pacific Islander	100.0	95.9	100.0	100.0	100.0	100.0
White	98.3	98.2	98.2	98.0	97.6	97.0
Multi-Racial	98.1	97.8	97.8	97.0	97.6	96.6
LEP	97.6	97.8	97.9	96.5	96.0	95.3
IDEA	89.0	89.5	90.1	89.5	89.7	87.9

Note: AmerIndian/Alaskan = American Indian or Alaska Native; Pacific Islander = Native Hawaiian or Other Pacific Islander; LEP = Limited English Proficiency Status; IDEA= Individuals with Disabilities Education Act.

Table 14. Participation Rates in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	97.7	97.7	97.7	97.1	96.7	96.0
Female	98.3	98.2	98.2	97.8	97.3	96.3
Male	97.1	97.2	97.1	96.3	96.2	95.8
Black or African American	96.8	96.7	96.7	96.2	95.7	95.1
AmerIndian/Alaskan	98.3	100.0	98.7	98.0	97.0	95.7
Asian	98.2	97.4	97.7	97.1	98.5	97.8
Hispanic or Latino	97.4	97.6	97.5	96.6	96.4	95.4
Pacific Islander	100.0	95.9	100.0	100.0	100.0	100.0
White	98.0	98.0	98.0	97.7	97.1	96.4
Multi-Racial	98.0	97.7	97.5	96.3	96.6	96.1
LEP	97.5	97.7	97.7	95.8	95.5	94.6
IDEA	88.2	89.1	89.7	88.5	88.3	86.4

Table 15. Number of Students in ELA/L Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	35,315	35,940	36,300	36,627	37,794	38,522
Female	17,296	17,683	17,859	18,057	18,435	18,684
Male	18,017	18,253	18,432	18,558	19,321	19,797
Black or African American	4,217	4,392	4,561	4,676	4,917	5,007
AmerIndian/Alaskan	114	82	75	99	98	113
Asian	1,911	1,975	1,884	1,872	1,926	1,941
Hispanic or Latino	10,855	10,663	10,822	10,838	11,184	11,037
Pacific Islander	29	47	35	31	42	31
White	16,485	17,071	17,262	17,548	18,115	18,847
Multi-Racial	1,704	1,710	1,661	1,563	1,512	1,546
LEP	4,710	4,590	3,926	3,259	3,126	2,764
IDEA	5,153	5,642	5,795	5,835	5,997	6,063

Table 16. Number of Students in Mathematics Summative Assessment

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	35,220	35,860	36,200	36,426	37,528	38,238
Female	17,252	17,640	17,808	17,966	18,315	18,527
Male	17,966	18,216	18,383	18,448	19,176	19,669
Black or African American	4,195	4,375	4,537	4,636	4,858	4,965
AmerIndian/Alaskan	114	82	75	98	97	112
Asian	1,911	1,975	1,878	1,868	1,923	1,935
Hispanic or Latino	10,820	10,636	10,780	10,760	11,090	10,918
Pacific Islander	29	47	35	31	42	31
White	16,448	17,039	17,238	17,482	18,022	18,739
Multi-Racial	1,703	1,706	1,657	1,551	1,496	1,538
LEP	4,710	4,582	3,912	3,244	3,109	2,743
IDEA	5,146	5,630	5,781	5,777	5,911	5,971

3.2 SUMMARY OF STUDENT PERFORMANCE

Tables 17–20 summarize overall student performance in the 2021–2022 summative test for all students and by subgroups, including the average and the standard deviation of overall scale scores, the percentage of students in each achievement level, and the percentage of proficient students.

Figures 1 and 2 show the percentage of proficient students over the past six years for all students in ELA/L and over the past seven years for all students in mathematics (cohort comparisons). Figures 3 and 4 show the average scale scores over the past six years for all students in ELA/L and over the past seven years for all students in mathematics. In ELA/L, student performance is compared for six years because ELA/L scores in 2014–2015 were based on both computer-adaptive test (CAT) and performance task (PT) components while ELA/L scores from 2015–2016 were based on the CAT component only. In Figures 1–4, the 2019–2020 performance is not included because the testing was cancelled due to the COVID-19 pandemic. The average and the standard deviation of scale scores, as well as the percentage of proficient students for each test administration across four years, are provided in Appendix B.

Table 17. Descriptive Statistics and Percentage of Students in Achievement Levels
for Overall and by Subgroup: ELA/L (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	35,315	2419	96	31	22	21	26	47
Female	17,296	2426	95	29	22	21	28	49
Male	18,017	2413	96	34	22	20	24	44
Black or African American	4,217	2375	86	49	24	16	11	27
AmerIndian/Alaskan	114	2390	105	46	18	17	19	36
Asian	1,911	2466	93	16	16	22	45	67
Hispanic or Latino	10,855	2374	89	50	24	15	11	27
Pacific Islander	29	2405	86	38	24	21	17	38
White	16,485	2454	86	17	21	25	37	62
Multi-Racial	1,704	2436	94	25	21	23	31	54
LEP	4,710	2347	80	63	22	10	5	15
IDEA	5,153	2347	82	64	20	10	6	16
Grade 4								
All Students	35,940	2463	102	34	17	22	27	49
Female	17,683	2470	100	31	18	22	29	51
Male	18,253	2457	103	36	17	22	26	47
Black or African American	4,392	2415	93	54	18	15	12	27
AmerIndian/Alaskan	82	2445	91	44	18	17	21	38
Asian	1,975	2522	94	14	14	22	49	72
Hispanic or Latino	10,663	2415	95	53	18	17	12	29
Pacific Islander	47	2458	98	36	19	19	26	45
White	17,071	2498	90	19	17	27	38	64
Multi-Racial	1,710	2481	100	26	17	23	33	56
LEP	4,590	2383	87	66	17	12	5	17
IDEA	5,642	2384	90	68	14	12	6	18
Grade 5								
All Students	36,300	2502	106	30	19	27	25	52
Female	17,859	2510	104	26	19	27	27	55
Male	18,432	2494	108	33	18	26	23	49
Black or African American	4,561	2449	96	49	21	19	10	29
AmerIndian/Alaskan	75	2467	98	40	17	32	11	43
Asian	1,884	2562	99	12	13	27	48	75
Hispanic or Latino	10,822	2453	101	47	21	21	12	32
Pacific Islander	35	2481	106	34	17	34	14	49
White	17,262	2539	94	16	17	32	35	67
Multi-Racial	1,661	2512	105	26	18	28	28	56
LEP	3,926	2401	83	68	19	11	1	12
IDEA	5,795	2411	92	66	17	12	5	17

Note: The percentage of each achievement level may not add up to 100% due to rounding.

Table 18. Descriptive Statistics and Percentage of Students in Achievement Levels
for Overall and by Subgroup: ELA/L (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	36,627	2521	102	28	25	30	18	48
Female	18,057	2529	100	24	25	31	20	51
Male	18,558	2512	103	31	24	28	17	45
Black or African American	4,676	2476	91	44	28	21	7	28
AmerIndian/Alaskan	99	2495	83	32	34	29	4	33
Asian	1,872	2584	98	11	15	32	42	74
Hispanic or Latino	10,838	2474	96	44	27	21	8	29
Pacific Islander	31	2494	83	39	32	19	10	29
White	17,548	2554	91	15	23	37	25	62
Multi-Racial	1,563	2532	102	25	23	30	22	53
LEP	3,259	2406	71	76	20	4	0	5
IDEA	5,835	2431	84	65	22	11	2	13
Grade 7								
All Students	37,794	2541	109	28	22	33	17	50
Female	18,435	2553	105	24	23	34	19	54
Male	19,321	2529	112	32	22	31	15	46
Black or African American	4,917	2496	102	43	27	23	7	30
AmerIndian/Alaskan	98	2517	105	35	28	28	10	38
Asian	1,926	2611	97	10	14	36	40	76
Hispanic or Latino	11,184	2489	106	46	24	23	7	30
Pacific Islander	42	2542	103	24	29	33	14	48
White	18,115	2576	96	15	21	41	23	64
Multi-Racial	1,512	2552	107	24	22	35	19	54
LEP	3,126	2408	82	79	16	5	0	5
IDEA	5,997	2443	96	65	21	12	2	14
Grade 8								
All Students	38,522	2558	109	27	24	32	17	49
Female	18,684	2572	106	22	24	34	20	54
Male	19,797	2544	109	31	24	31	14	45
Black or African American	5,007	2512	101	43	27	23	7	30
AmerIndian/Alaskan	113	2523	92	35	34	26	6	32
Asian	1,941	2626	101	10	15	36	39	75
Hispanic or Latino	11,037	2510	104	42	27	24	7	31
Pacific Islander	31	2561	126	29	19	35	16	52
White	18,847	2590	97	15	23	39	23	62
Multi-Racial	1,546	2568	111	26	20	33	21	54
LEP	2,764	2422	74	81	15	3	0	4
IDEA	6,063	2463	91	63	23	11	3	14

Note: The percentage of each achievement level may not add up to 100% due to rounding.

Table 19. Descriptive Statistics and Percentage of Students in Achievement Levels
for Overall and by Subgroup: Mathematics (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	35,220	2426	92	31	22	25	22	47
Female	17,252	2421	89	33	22	26	20	45
Male	17,966	2430	95	29	21	25	24	50
Black or African American	4,195	2374	83	53	24	17	6	23
AmerIndian/Alaskan	114	2406	101	44	16	25	16	40
Asian	1,911	2486	91	13	13	26	48	74
Hispanic or Latino	10,820	2382	84	50	24	18	9	27
Pacific Islander	29	2424	93	34	21	17	28	45
White	16,448	2459	80	16	21	32	32	64
Multi-Racial	1,703	2437	91	25	22	27	25	53
LEP	4,710	2367	81	57	23	14	6	20
IDEA	5,146	2351	89	64	18	12	6	18
Grade 4								
All Students	35,860	2469	94	28	27	24	21	45
Female	17,640	2465	90	29	28	25	19	43
Male	18,216	2473	98	27	26	23	24	47
Black or African American	4,375	2413	83	51	29	13	6	20
AmerIndian/Alaskan	82	2442	88	37	35	15	13	28
Asian	1,975	2537	86	8	18	25	50	75
Hispanic or Latino	10,636	2423	86	46	30	16	8	24
Pacific Islander	47	2466	83	28	34	23	15	38
White	17,039	2503	81	13	26	32	30	61
Multi-Racial	1,706	2482	95	24	25	26	26	51
LEP	4,582	2405	82	54	28	12	5	17
IDEA	5,630	2390	88	62	24	10	5	15
Grade 5								
All Students	36,200	2493	98	36	25	18	21	39
Female	17,808	2489	94	38	26	18	18	36
Male	18,383	2497	102	35	24	18	24	41
Black or African American	4,537	2433	83	62	23	9	5	14
AmerIndian/Alaskan	75	2456	80	49	32	13	5	19
Asian	1,878	2566	93	14	18	19	49	68
Hispanic or Latino	10,780	2446	88	56	24	12	8	20
Pacific Islander	35	2459	92	46	31	14	9	23
White	17,238	2529	86	20	26	23	30	54
Multi-Racial	1,657	2501	99	34	25	18	24	41
LEP	3,912	2415	76	72	19	6	2	9
IDEA	5,781	2411	86	72	17	7	4	11

Note: The percentage of each achievement level may not add up to 100% due to rounding.

Table 20. Descriptive Statistics and Percentage of Students in Achievement Levels
for Overall and by Subgroup: Mathematics (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	36,426	2506	115	36	27	18	19	37
Female	17,966	2504	112	36	28	18	17	36
Male	18,448	2508	119	36	26	18	20	38
Black or African American	4,636	2440	102	61	25	9	5	14
AmerIndian/Alaskan	98	2486	88	43	33	16	8	24
Asian	1,868	2594	112	14	16	19	51	69
Hispanic or Latino	10,760	2450	107	57	26	11	7	18
Pacific Islander	31	2481	94	48	32	16	3	19
White	17,482	2548	98	20	29	25	27	51
Multi-Racial	1,551	2515	118	34	26	19	22	41
LEP	3,244	2391	88	81	15	3	1	4
IDEA	5,777	2401	104	75	17	5	3	8
Grade 7								
All Students	37,528	2524	116	37	25	19	19	38
Female	18,315	2522	112	37	26	19	17	36
Male	19,176	2526	120	37	23	19	20	39
Black or African American	4,858	2459	100	62	24	10	5	15
AmerIndian/Alaskan	97	2497	107	44	34	9	12	22
Asian	1,923	2621	111	13	16	20	51	71
Hispanic or Latino	11,090	2466	103	58	24	12	6	18
Pacific Islander	42	2524	96	38	26	19	17	36
White	18,022	2566	102	21	27	26	27	53
Multi-Racial	1,496	2536	117	34	24	20	22	42
LEP	3,109	2404	81	84	13	2	1	3
IDEA	5,911	2421	98	76	15	6	3	8
Grade 8								
All Students	38,238	2532	124	42	23	16	18	34
Female	18,527	2534	119	42	24	17	17	34
Male	19,669	2531	128	43	22	16	19	35
Black or African American	4,965	2463	106	67	20	8	5	13
AmerIndian/Alaskan	112	2493	104	53	31	10	6	16
Asian	1,935	2632	123	15	17	21	47	67
Hispanic or Latino	10,918	2472	107	64	21	9	6	15
Pacific Islander	31	2545	142	42	13	19	26	45
White	18,739	2574	111	26	26	22	26	47
Multi-Racial	1,538	2542	130	40	22	15	22	37
LEP	2,743	2404	80	90	8	2	1	2
IDEA	5,971	2426	97	80	13	4	2	6

Note: The percentage of each achievement level may not add up to 100% due to rounding.

Figure 1. ELA/L Percent Proficient Across Years

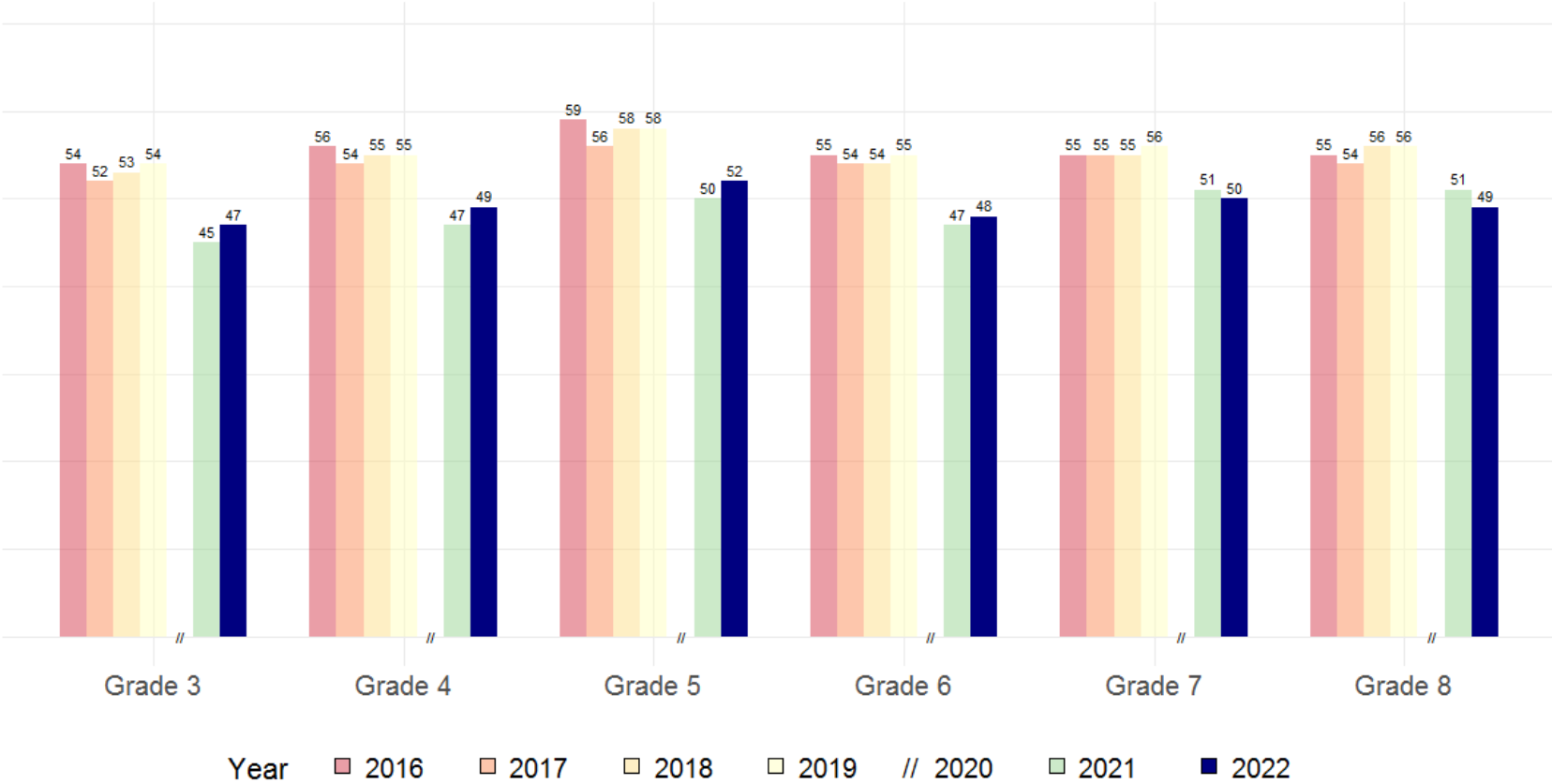


Figure 2. Mathematics Percent Proficient Across Years

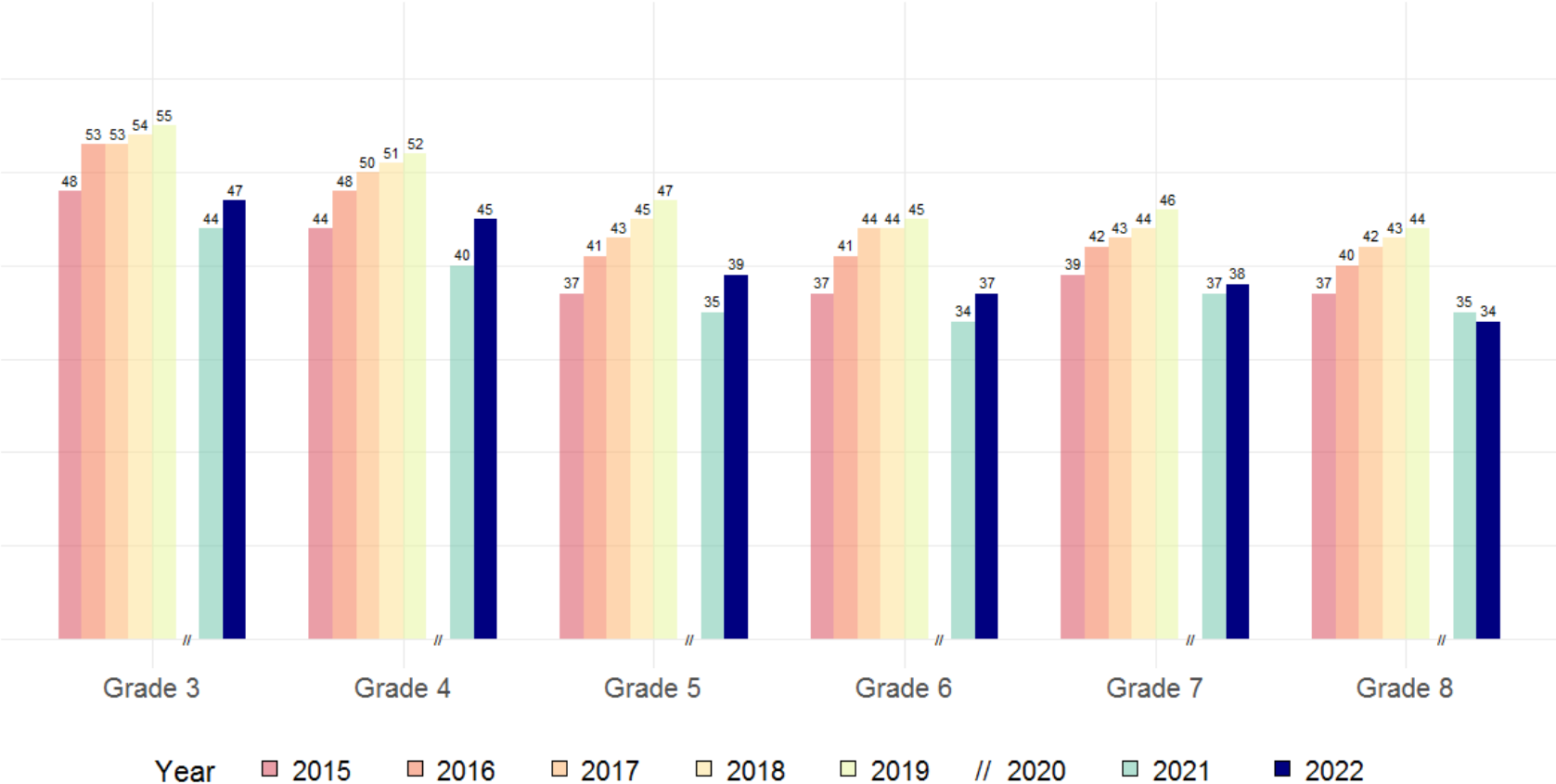


Figure 3. ELA/L Average Scale Score Across Years

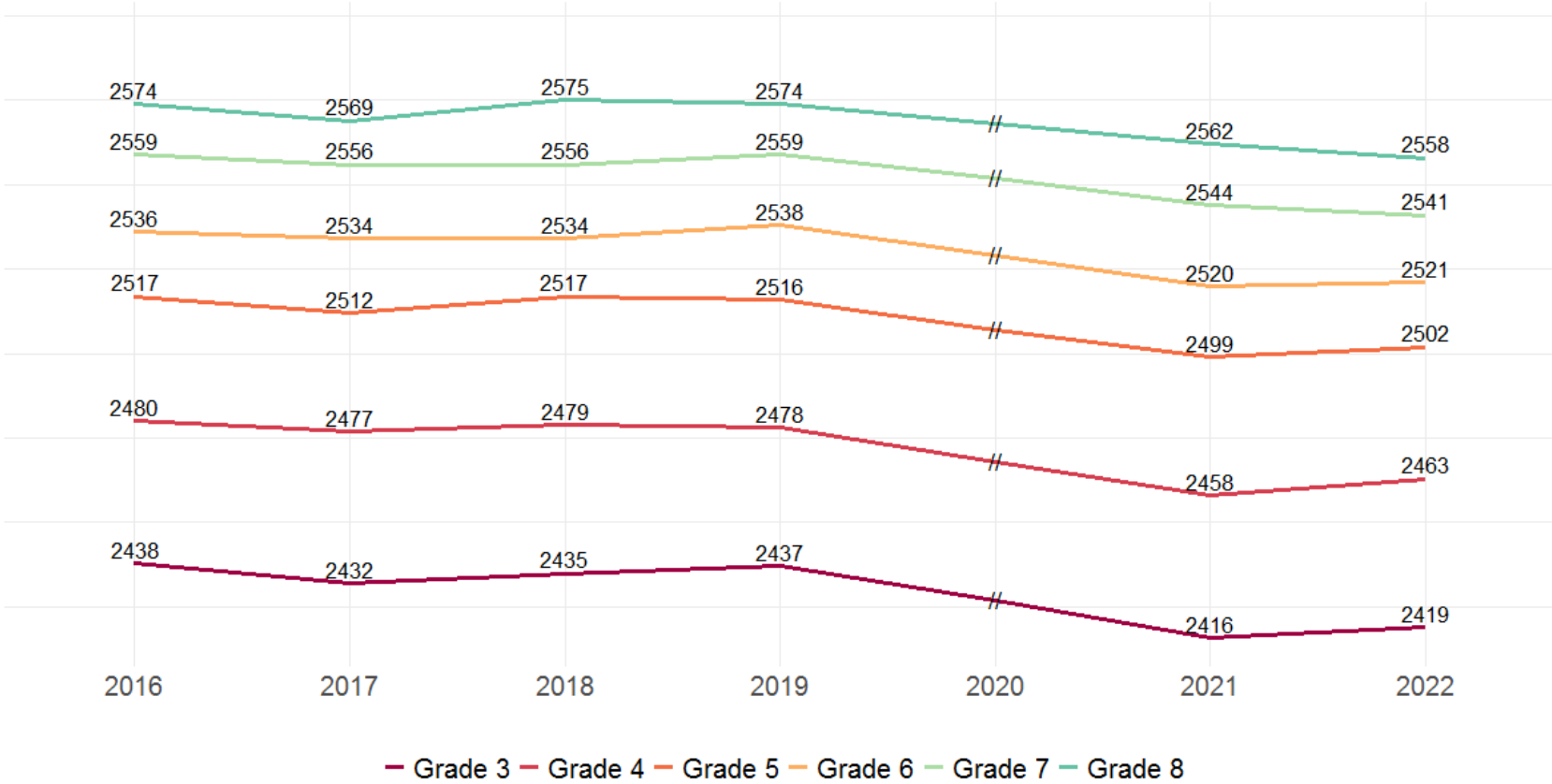
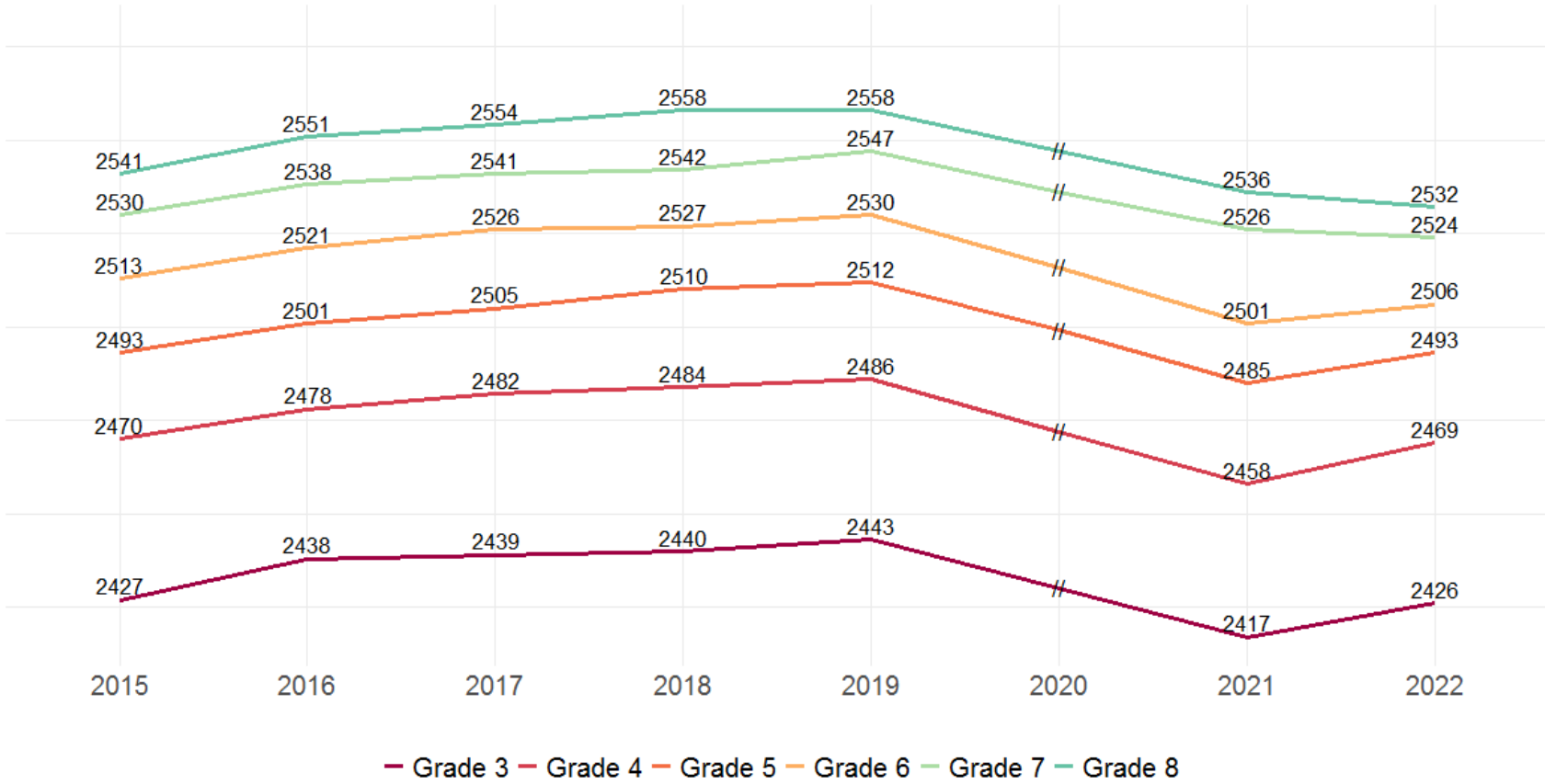


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three performance categories, considering the standard error of measurement (SEM) of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. Tables 21 and 22 present the distribution of performance categories for each claim. The number of claims is three in both ELA/L and mathematics, combining claims 2 and 4.

Table 21. ELA/L Percentage of Students in Performance Categories for Claims

Grade	Performance Category	Claim 1 Reading	Claims 2 and 4: Writing and Research	Claim 3 Listening
3	Below	32	37	17
	At/Near	41	39	62
	Above	26	24	21
4	Below	29	33	18
	At/Near	44	43	59
	Above	27	24	23
5	Below	28	32	18
	At/Near	44	38	59
	Above	29	30	23
6	Below	31	33	20
	At/Near	46	43	62
	Above	24	24	18
7	Below	28	32	21
	At/Near	45	45	60
	Above	27	24	18
8	Below	28	32	19
	At/Near	44	44	60
	Above	28	24	21

Table 22. Mathematics Percentage of Students in Performance Categories for Claims

Grade	Performance Category	Claim 1	Claims 2 and 4	Claim 3
3	Below	37	31	28
	At/Near	30	42	44
	Above	33	28	27
4	Below	39	35	33
	At/Near	29	41	40
	Above	32	24	26
5	Below	44	38	37
	At/Near	29	42	43
	Above	27	20	20
6	Below	45	40	36
	At/Near	31	42	45
	Above	24	19	19
7	Below	45	36	33
	At/Near	30	44	47
	Above	25	20	20
8	Below	47	39	36
	At/Near	31	41	48
	Above	23	20	16

Legend:

Claim 1: Concepts and Procedures;

Claims 2 and 4: Problem Solving and Modeling and Data Analysis;

Claim 3: Communicating Reasoning

3.3 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 5–10 display the empirical distribution of the Connecticut student scale scores in the 2021–2022 test administration and the distribution of the administered summative item difficulty parameters for each grade for overall and by claim. For overall, the student ability distribution is to the left in all grades and subjects, a pattern more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. At the claim, the student ability distribution is shifted to the left for all claims except for claims 2 and 4 in grades 4–5 and claim 3 in grades 4–8 in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in grades 3–4. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth of Knowledge [DOK], item type, and item difficulties) to better measure low-performing students.

Figure 5. Student Ability—Item Difficulty Distribution for ELA/L

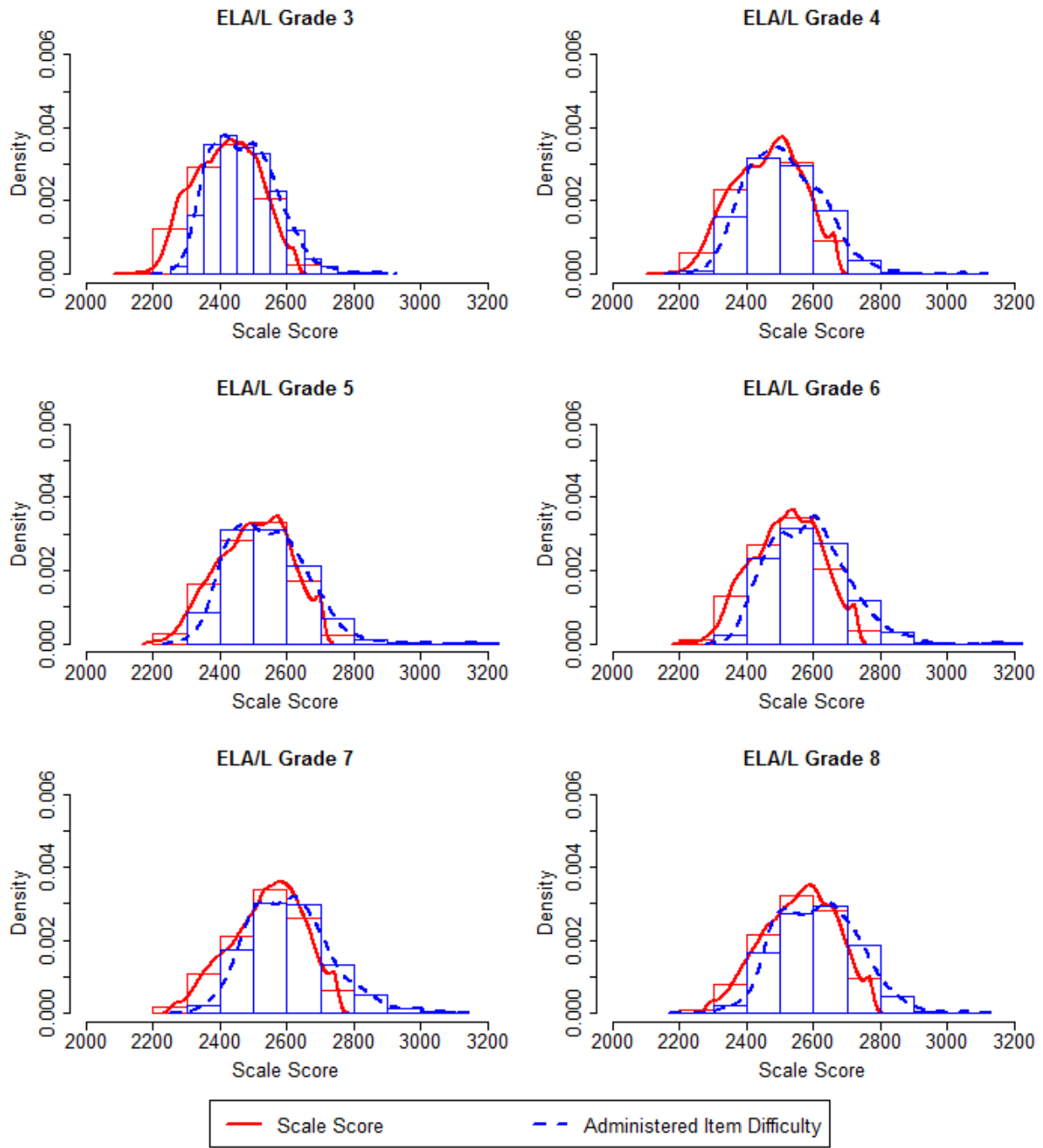


Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)

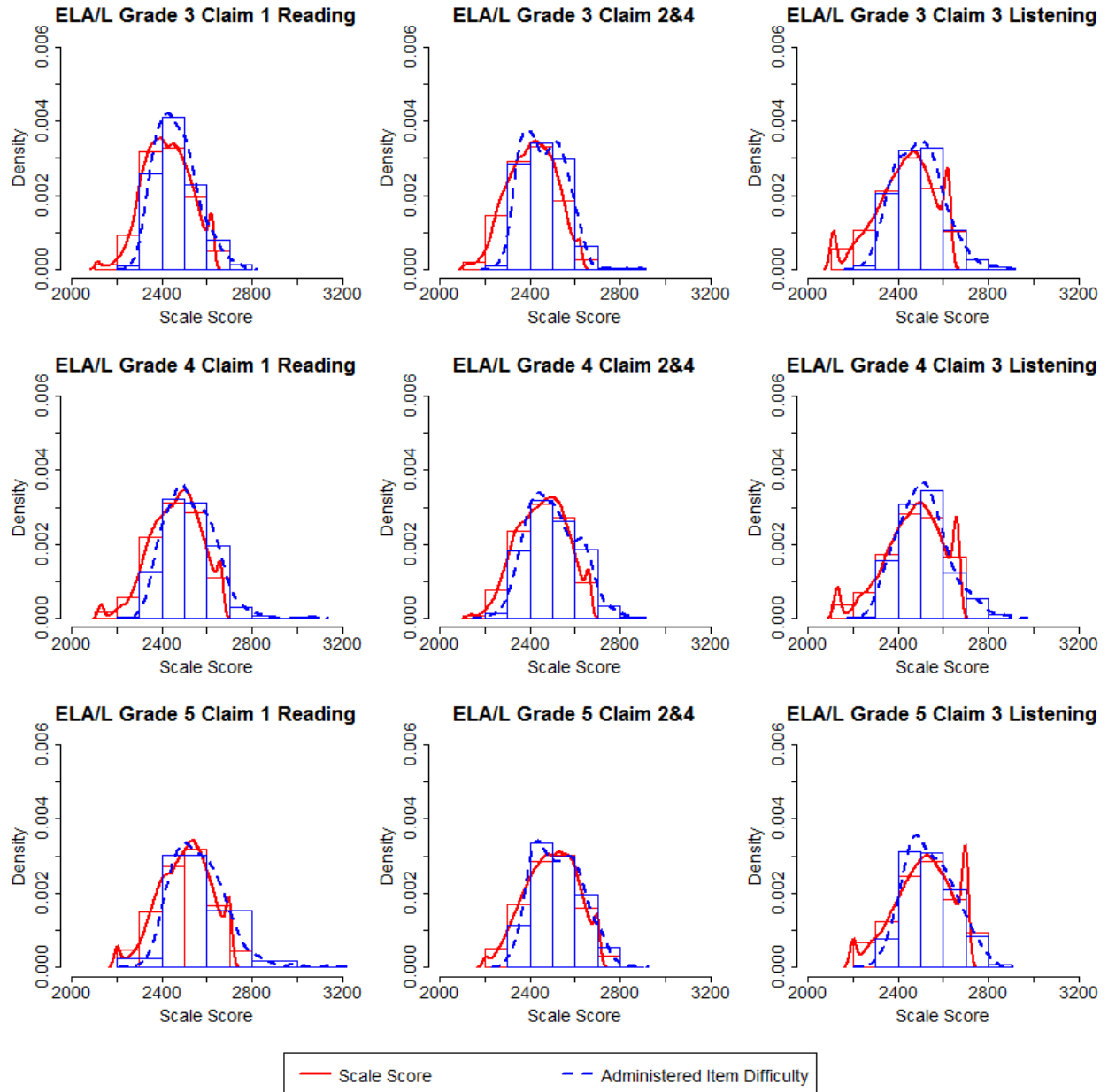


Figure 7. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 6–8)

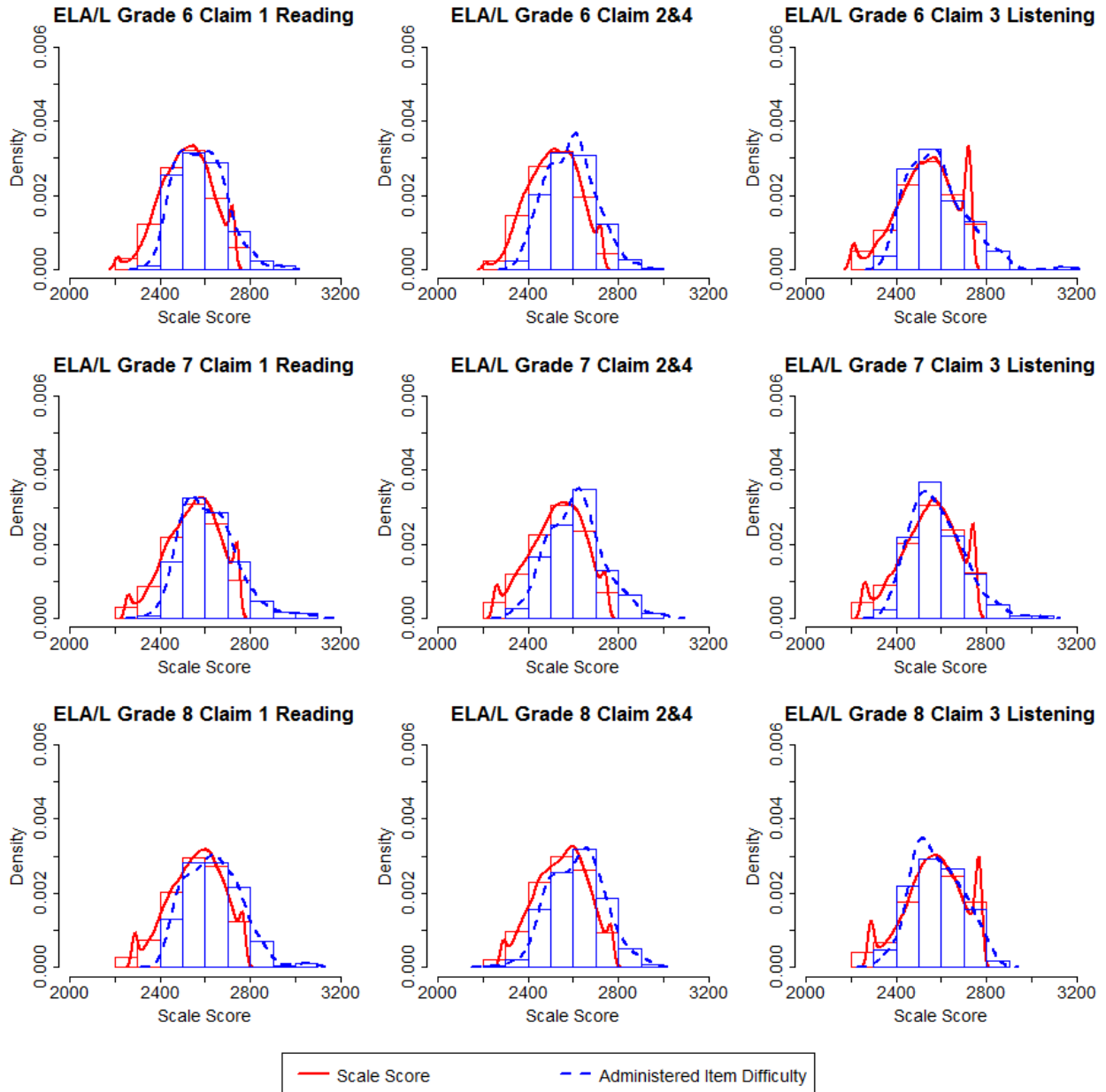


Figure 8. Student Ability—Item Difficulty Distribution for Mathematics

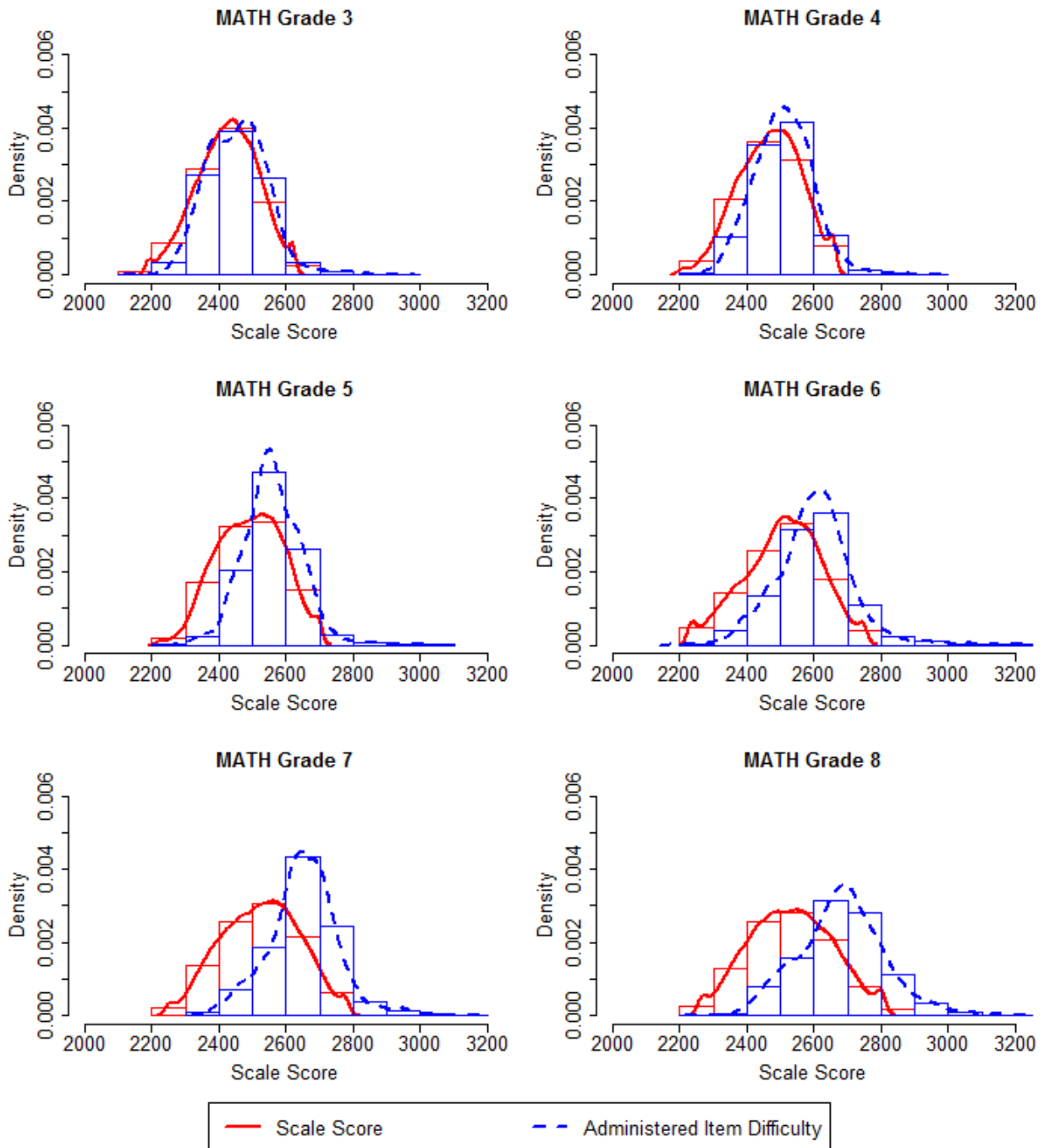


Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

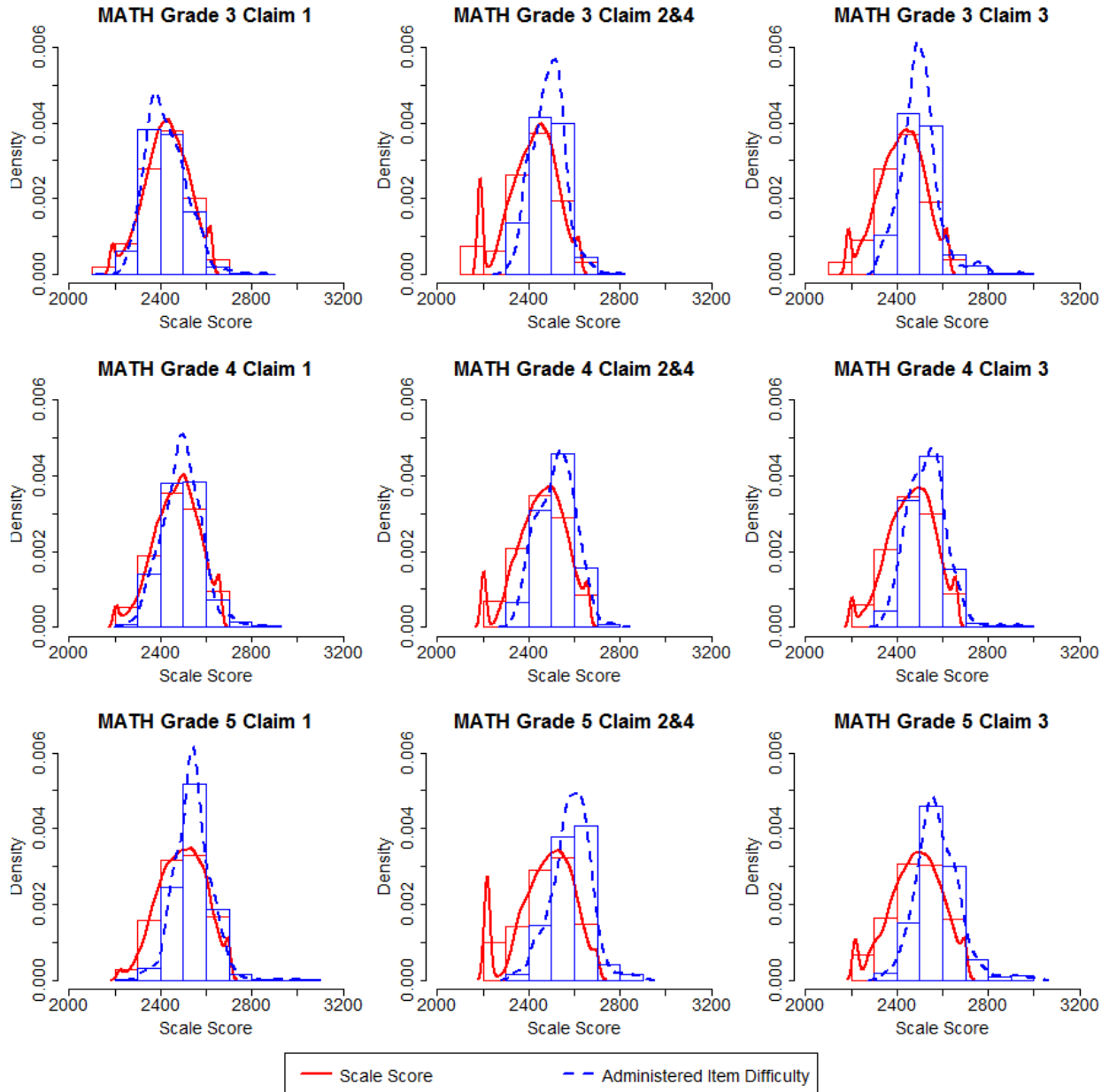
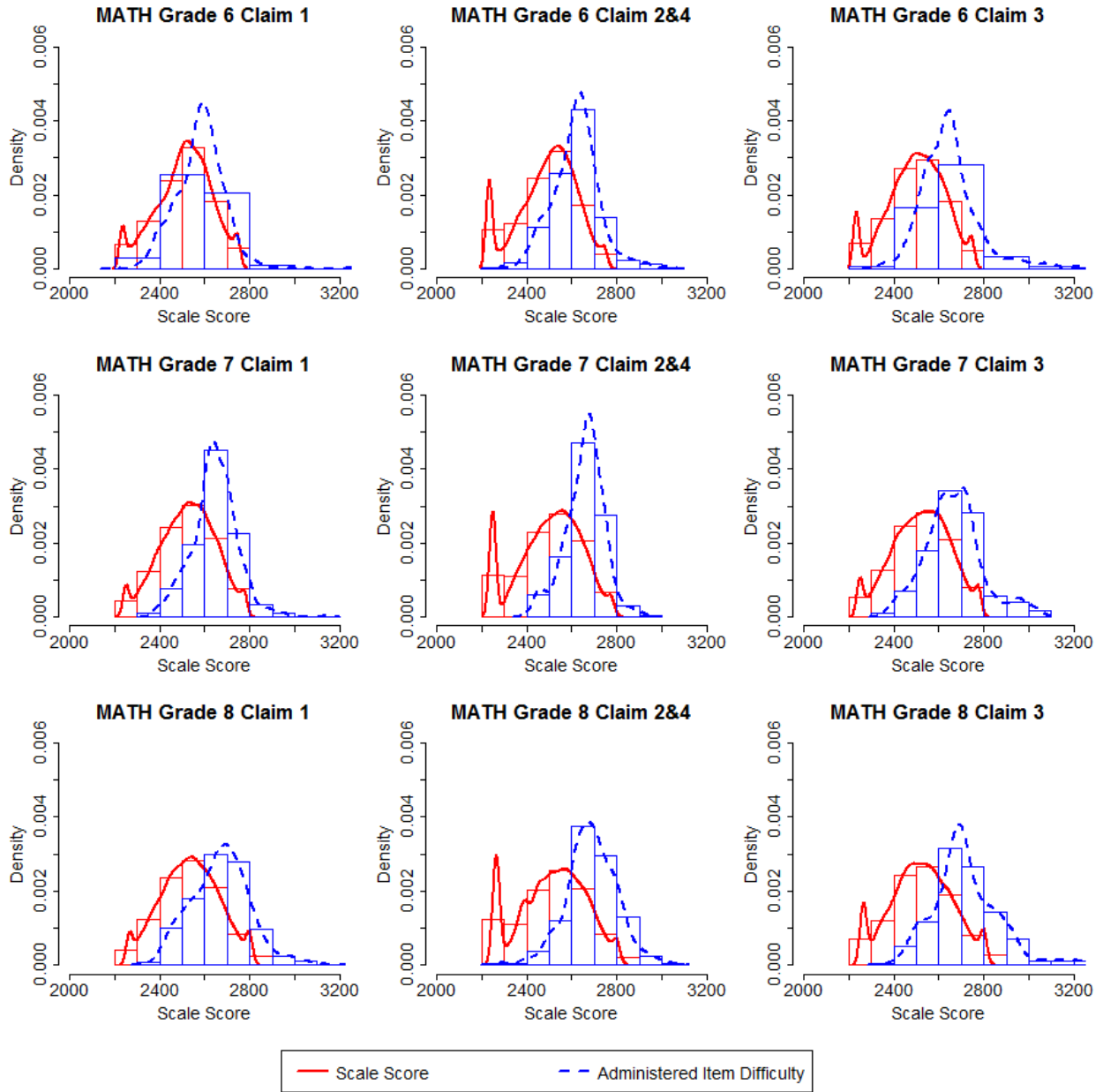


Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8)



4. VALIDITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), validity refers to the degree to which evidence and theory support the interpretations of test scores as described by the intended uses of assessments. The validity of an intended interpretation of test scores relies on all the evidence accrued about the technical quality of a testing system, including test development and construction procedures, test score reliability, accurate scaling and equating, procedures for setting meaningful achievement standards, standardized test administration and scoring procedures, and attention to fairness for all test takers. The appropriateness and usefulness of the Smarter Balanced summative assessments depends on the assessments meeting the relevant standards of validity.

Validity evidence provided in this chapter is as follows:

- Test content
- Internal structure

Evidence on test content validity is provided with the blueprint match rates for the delivered tests. Evidence on internal structure is examined in the results of inter-correlations among claim scores.

Some of the evidence on standardized test administration, scoring procedures, and attention to fairness for all test takers is provided in other chapters.

4.1 EVIDENCE ON TEST CONTENT

The Smarter Balanced summative assessment includes two components: the computer-adaptive test (CAT) and the performance task (PT). For the CAT, each student receives a different set of items adapted to his or her ability. For the PT, each student is administered with a fixed-form test. The content coverage in all PT forms is the same.

In the adaptive item selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints specify a range of items to be administered in each claim, content domain/standards, and/or targets. Moreover, blueprints constrain the Depth of Knowledge (DOK) and item and passage types. For DOK constraints, the Smarter Balanced blueprint specifies either the minimum or maximum number of items, not both the minimum and maximum. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In English language arts/literacy (ELA/L), the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 23 and 24 present the percentages of tests aligned with the ELA/L test blueprint constraints for items in claims, targets, DOK, and passages in claims 1 and 3. For the passage constraints, four passages in claim 1 reading and three to four passages in claim 3 listening are required. The composition of four reading passages in claim 1 is two literary text passages (one long and one short passage) and two informational text passages (one long and one short passage) in grades 3–5 and one literary text passage (long passage) and three informational text passages (one long and two short passages) in grades 6–8.

In ELA/L, all tests met the blueprint requirements except some targets in claim 1, which administered a few items more or less than the item requirement. The violations in claim 1 reading targets appeared in all grades due to the uneven distribution of items across targets and DOKs within and across passages.

Tables 25 and 26 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT. In mathematics, the tests met all blueprint requirements except a target set of E and F in Claim 1 for grade 7. Violations involved administering one item more or fewer than the blueprint requirements.

Table 23. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target:
ELA/L Grades 3–5

Claim	Content Category/Target	Required Items/Passages	% Blueprint Match		
			Grade 3	Grade 4	Grade 5
1	Literary Text	7–8	100	100	100
	Target 2: Central Ideas	1–2	100	100	100
	Target 4: Reasoning and Evidence	1–2	100	100	99
	Targets 1, 3, 5, 6, and 7	3–6	100	100	100
	Target 2 or 4 Short Text	0–1	100	100	100
	Long Literary Text Passage	≥ 1	100	100	100
	Short Literary Text Passage	≤ 2	100	100	100
	Informational Text	7–8	100	100	100
	Target 9: Central Ideas	1–2	99	99	99
	Target 11: Reasoning and Evidence	1–2	100	100	100
	Targets 8, 10, 12, 13, and 14	3–6	100	100	100
	Target 9 or 11 Short Text	0–1	100	100	100
	Long Informational Text Passage	≥ 1	100	100	100
	Short Informational Text Passage	≤ 2	100	100	100
	DOK 2	≥ 7	100	100	100
DOK 3 or Higher	≥ 2	100	100	100	
2	Writing	10	100	100	100
	Target 1, 3, or 6: Organization/Purpose ^a	3	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration ^a	3	100	100	100
	Target 8: Language and Vocabulary Use	2	100	100	100
	Target 9: Edit/Clarify	5	100	100	100
	DOK 2	≥ 4	100	100	100
	DOK 3 or 4	1	100	100	100
Brief-Write	1	100	100	100	
3	Listening	8–9	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100
	DOK 2 or Higher	≥ 3	100	100	100
	Listening Passage	3–4	100	100	100
4	Research	6	100	100	100
	Target 2: Interpret and Integrate Information	6	100	100	100
	Target 3: Analyze Information/Sources	6	100	100	100
	Target 4: Use Evidence	6	100	100	100

^a Each student will receive a total of three items, with at least one item in Organization/Purpose and at least one item in Evidence/Elaboration.

Table 24. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target:
ELA/L Grades 6–8

Claim	Content Category/Target	Required Items/Passages	% Blueprint Match		
			Grade 6	Grade 7	Grade 8
1	Literary Text	4–7	100	100	100
	Target 2: Central Ideas	1	99	100	100
	Target 4: Reasoning and Evidence	1	99	100	98
	Targets 1, 3, 5, 6, and 7	2–5	100	100	100
	Target 2 or 4 Short Text	0–1	100	100	100
	Long Literary Text Passage	≥ 1	100	100	100
	Informational Text	10–12 ^a	100	100	100
	Targets 9 and 11	2–5	100	99	100
	Targets 8, 10, 12, 13, and 14	7–10	100	99	100
	Target 9 or 11 Short Text	0–1	100	100	100
	Long Informational Text Passage	≥ 1	100	100	100
	Short Informational Text Passage	≤ 2	100	100	100
	DOK 1	≤ 5	100	100	100
	DOK 3 or Higher	≥ 2	100	100	100
2	Writing	10	100	100	100
	Target 1, 3, or 6: Organization/Purpose ^b	3	100	100	100
	Target 1, 3, or 6: Evidence/Elaboration ^b	3	100	100	100
	Target 8: Language and Vocabulary Use	2	100	100	100
	Target 9: Edit/Clarify	5	100	100	100
	DOK 2	≥ 4	100	100	100
	DOK 3 or 4	1	100	100	100
	Brief-Write	1	100	100	100
3	Listening	8–9	100	100	100
	Target 4: Listen/Interpret	8–9	100	100	100
	DOK 2 or Higher	≥ 3	100	100	100
	Listening Passage	3–4	100	100	100
4	Research	6	100	100	100
	Target 2: Analyze/Integrate Information	6	100	100	100
	Target 3: Evaluate Information/Sources	6	100	100	100
	Target 4: Use Evidence	6	100	100	100

^a Required items for Informational Text are 10–12 in grades 6 and 7, and 12 in grade 8.

^b Each student will receive a total of three items, with at least one item in Organization/Purpose and at least one item in Evidence/Elaboration.

Table 25. Percentage of Delivered Tests Meeting Blueprint Requirement for Each Claim and Target:
Mathematics Grades 3–5

Claim	Content Domain	Grade 3		Grade 4		Grade 5	
		Required Items	% Blueprint Match	Required Items	% Blueprint Match	Required Items	% Blueprint Match
1	Overall	17–20	100	17–20	100	17–20	100
	DOK 2 or Higher	≥ 7	100	≥ 7	100	≥ 7	100
	<i>Priority Cluster</i>	13–15	100				
	Targets B, C, G, I	5–6	100				
	Targets D, F	5–6	100				
	Target A	2–3	100				
	<i>Supporting Cluster</i>	4–5	100				
	Targets E, J, K	3–4	100				
	Target H	1	100				
	<i>Priority Cluster</i>			13–15	100		
	Targets A, E, F			8–9	100		
	Target G			2–3	100		
	Target D			1–2	100		
	Target H			1	100		
	<i>Supporting Cluster</i>			4–5	100		
	Targets I, K			2–3	100		
	Targets B, C, J			1	100		
	Target L			1	100		
	<i>Priority Cluster</i>					13–15	100
Targets E, I					5–6	100	
Target F					4–5	100	
Targets C, D					3–4	100	
<i>Supporting Cluster</i>					4–5	100	
Targets J, K					2–3	100	
Targets A, B, G, H					2	100	
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3	Overall	8	100	8	100	8	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	3	100	3	100	3	100
	Targets B, E	3	100	3	100	3	100
	Targets C, F	2	100	2	100	2	100

Table 26. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target:
Mathematics Grades 6–8

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	% Blueprint Match	Required Items	% Blueprint Match	Required Items	% Blueprint Match
1	Overall	16–20	100	16–20	100	16–20	100
	DOK 2 or Higher	≥ 7	100	≥ 7	100	≥ 7	100
	<i>Priority Cluster</i>	12–15	100				
	Targets E, F	5–6	100				
	Target A	3–4	100				
	Targets B, G	2	100				
	Target D	2	100				
	<i>Supporting Cluster</i>	4–5	100				
	Targets C, H, I, J	4–5	100				
	<i>Priority Cluster</i>			12–15	100		
	Targets A, D			8–9	100		
	Targets B, C			5–6	100		
	<i>Supporting Cluster</i>			4–5	99		
	Targets E, F			2–3	99		
	Targets G, H, I			1–2	100		
<i>Priority Cluster</i>					12–15	100	
Targets C, D					5–6	100	
Targets B, E, G					5–6	100	
Targets F, H					2–3	100	
<i>Supporting Cluster</i>					4–5	100	
Targets A, I, J					4–5	100	
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
4. Targets C, F	1	100	1	100	1	100	
3-Calc	Overall	7	100	8	100	8	100
	DOK 3 or Higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	2–3	100	3	100	3	100
	Targets B, E	2–3	100	3	100	3	100
	Targets C, F, G	1–2	100	2	100	2	100
3-No Calc	Overall	1	100				

Table 27 summarizes the target coverage, the average, and the range of the numbers of unique targets administered in each delivered CAT test by claim. Because the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered varies across tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level across all tests combined.

Table 27. Average and the Range of the Number of Unique Targets Assessed within Each Claim Across All Delivered CAT Components

Grade	Total Targets in Blueprint				Average				Range (Minimum – Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
ELA/L												
3	14	5	1	3	11	5	1	3	8–14	4–5	1–1	3–3
4	14	5	1	3	12	5	1	3	8–14	4–5	1–1	3–3
5	14	5	1	3	11	5	1	3	8–14	4–5	1–1	3–3
6	14	5	1	3	11	5	1	3	8–11	5–5	1–1	3–3
7	14	5	1	3	10	5	1	3	9–11	4–5	1–1	3–3
8	14	5	1	3	11	5	1	3	8–11	4–5	1–1	3–3
Mathematics												
3	11	4	6	6	11	2	5	3	10–11	2–2	3–6	3–4
4	12	4	6	6	10	2	5	3	9–10	2–2	3–6	3–3
5	11	4	6	6	9	2	5	3	8–9	2–2	3–6	2–4
6	10	4	7	6	10	2	5	3	9–10	2–2	3–7	3–3
7	9	3	7	6	8	2	5	3	6–8	1–2	3–6	2–4
8	10	4	7	6	10	2	5	3	10–10	2–2	3–6	2–4

An adaptive testing algorithm constructs a test form unique to each student, targeting the student’s level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty) across individual students, but test scores from the individual tests are comparable since all test forms measure the same content, albeit with a different set of test items. Although each form is unique with respect to its items, all forms align with the same curricular expectations outlined in the test blueprints.

4.2 EVIDENCE ON INTERNAL STRUCTURE

The measurement model used in the Smarter Balanced assessments assumes a single underlying latent trait in student ability estimates, which supports the reporting of a single total ability score. During the test construction phase, the test blueprint was designed to cover multiple claims under each subject. The item selection algorithm prioritizes blueprint matching to ensure that each test contains an appropriate combination of items from each claim. Assessing the relationship between these different claim scores is a measure of internal validity according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The presence of high correlations among claim scores is evidence that the Smarter Balanced assessment measures a single underlying ability and the claim scores are related to each other.

The correlations among claim scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 28 and 29. The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates. The observed correlation between two claim scores with measurement errors can be corrected for attenuation as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$, where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y , r_{xx} is the reliability coefficient for x , and r_{yy} is the reliability coefficient for y .

When corrected for attenuation (above diagonal), the correlations among claim scores are higher than observed correlations. The disattenuated correlations are quite high, especially in mathematics. The correction for attenuation is large in mathematics, because the marginal reliabilities of claims 2 and 4 and claim 3 scores are low. The low reliabilities are due to large standard errors among lower scores because of a shortage of easy items in the item pool.

Because the reliability for claim scores is low, the performance of all the claim scores is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 21 and 22, Section 3.2. Scale scores are not reported for claims.

Table 28. Correlations among Claims for ELA/L

Grade	Claim	Observed and Disattenuated Correlations		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1: Reading		0.97	0.98
	Claims 2 & 4: Writing & Research	0.79		0.99
	Claim 3: Listening	0.66	0.69	
4	Claim 1: Reading		0.98	1
	Claims 2 & 4: Writing & Research	0.77		1
	Claim 3: Listening	0.68	0.70	
5	Claim 1: Reading		0.99	1
	Claims 2 & 4: Writing & Research	0.79		0.99
	Claim 3: Listening	0.69	0.72	
6	Claim 1: Reading		0.98	1
	Claims 2 & 4: Writing & Research	0.77		1
	Claim 3: Listening	0.68	0.70	
7	Claim 1: Reading		1	1
	Claims 2 & 4: Writing & Research	0.79		1
	Claim 3: Listening	0.68	0.71	
8	Claim 1: Reading		0.99	1
	Claims 2 & 4: Writing & Research	0.78		1
	Claim 3: Listening	0.70	0.70	

Table 29. Correlations among Claims for Mathematics

Grade	Claim	Observed and Disattenuated Correlations		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1		1	0.97
	Claims 2 & 4	0.81		1
	Claim 3	0.80	0.76	
4	Claim 1		1	0.98
	Claims 2 & 4	0.83		1
	Claim 3	0.83	0.78	
5	Claim 1		1	0.99
	Claims 2 & 4	0.79		1
	Claim 3	0.80	0.74	
6	Claim 1		1	1
	Claims 2 & 4	0.83		1
	Claim 3	0.79	0.75	
7	Claim 1		1	1
	Claims 2 & 4	0.81		1
	Claim 3	0.80	0.74	
8	Claim 1		1	1
	Claims 2 & 4	0.81		1
	Claim 3	0.78	0.72	

Legend:

Claim 1: Concepts and Procedures;

Claims 2 & 4: Problem Solving & Modeling and Data Analysis;

Claim 3: Communicating Reasoning

5. RELIABILITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), reliability refers to the consistency of test scores across replications of a testing procedure. Reliability is related to the precision of measurement for a test and is evaluated, in part, in terms of the scores' standard error of measurement (SEM). In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores, and reliability coefficients are the correlation between scores on two equivalent forms of the test. Within the item response theory (IRT) framework, measurement error is conditional on ability and varies across the ability scale. The amount of precision in estimating achievement can be determined by the test information function, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is the inverse of measurement error; the larger the measurement error, the less test information is being provided. In computer-adaptive tests (CATs), items administered vary among students, so the amount of measurement error differs from one test to another, which yields conditional standard error of measurement (CSEM).

The reliability evidence of the Smarter Balanced summative assessments is provided with marginal reliability, CSEM, and classification accuracy and consistency in each achievement level.

5.1 MARGINAL RELIABILITY

The marginal reliability was computed for the scale scores, considering the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average CSEM, estimated at different points on the ability scale, for all students.

The marginal reliability ($\bar{\rho}$) is defined as

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^N CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i$ is the CSEM of the scale score for student i , and σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Another way to examine test reliability is with the SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that make up the test. In CATs, items administered vary among all students, so the SEM also can vary among students, which yields CSEM. The average CSEM can be computed as

$$\text{Average CSEM} = \sigma\sqrt{1 - \bar{\rho}} = \sqrt{\sum_{i=1}^N CSEM_i^2 / N}.$$

The smaller the value of the average CSEM, the greater accuracy of test scores.

Table 30 presents the marginal reliability coefficients and the average CSEM for the total scale scores.

Table 30. Marginal Reliability for ELA/L and Mathematics

Grade	N	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
ELA/L						
3	35,315	38–41	0.92	2419	96	27
4	35,940	38–41	0.91	2463	102	30
5	36,300	38–41	0.92	2502	106	30
6	36,627	38–42	0.91	2521	102	31
7	37,794	38–42	0.91	2541	109	33
8	38,522	38–42	0.91	2558	109	33
Mathematics						
3	35,220	39–40	0.95	2426	92	20
4	35,860	37–40	0.95	2469	94	21
5	36,200	38–40	0.94	2493	98	24
6	36,426	38–39	0.94	2506	115	29
7	37,528	38–40	0.93	2524	116	30
8	38,238	38–40	0.93	2532	124	33

5.2 STANDARD ERROR CURVES

Figures 11 and 12 present plots of the CSEM of scale scores across the range of abilities. The vertical lines indicate the three cut scores for the four achievement levels. For most of the ability range, the selection algorithm matched items to each student’s ability and to the test blueprints with similar precision. Because the item pool is finite and has fewer items located at the extremes of the ability scale, the selection algorithm had to prioritize meeting blueprint requirements over matching items to ability level for those students with very high or very low abilities. This results in higher standard errors for students with very high or very low abilities compared to students with abilities around and between the three cut scores.

Given that classifying students into achievement levels, especially into proficient or not proficient levels based on the Level 3 cut, is a high-stakes decision for schools, it is important that ability levels near and between the cut scores are measured with as much precision as possible. This increased precision near and between the cut scores is achieved by having more items in the item pool for abilities across the middle of the scale, where the cut scores are located.

A consequence of the selection algorithm’s prioritization of meeting blueprint requirements is that student ability near the low and high extremes of the scale is measured with relatively less precision. This produces the expected u-curve shape for the CSEM plots in Figures 11 and 12. An adaptive test with an infinitely large item pool and a selection algorithm that focused on maximizing information over blueprint requirements would produce flatter CSEM curves. The Smarter Balanced assessments focus on increasing precision where it is most needed, ability scores near and in between the cut scores. It is worth noting that larger standard errors are observed at the lower ends of the score distribution, relative to the higher ends. This occurs because the item pools currently have a shortage of very easy items that are better targeted toward these lower-achieving students. Content experts use this information to consider how to further target and populate item pools.

Figure 11. Conditional Standard Errors of Measurement for ELA/L

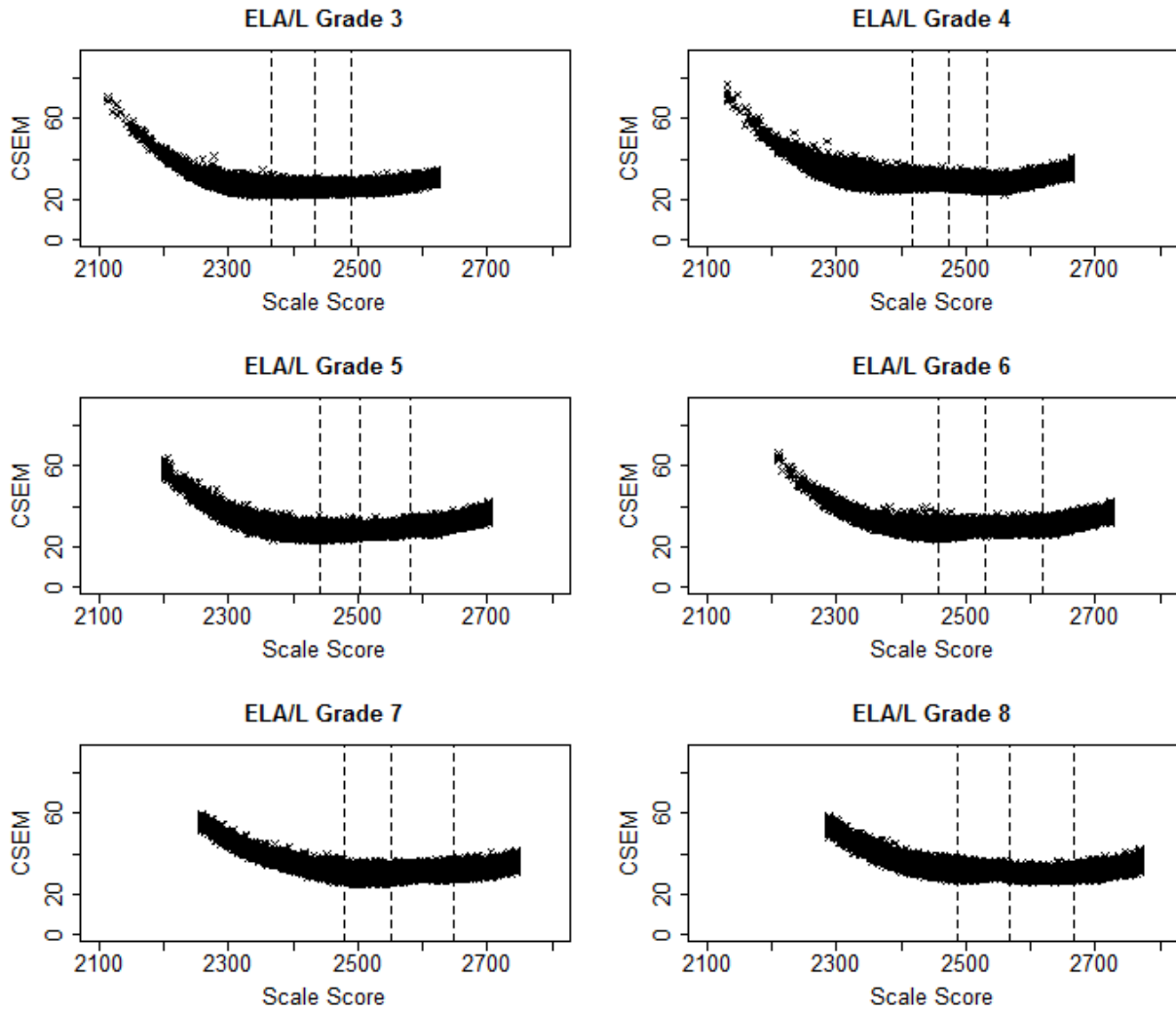
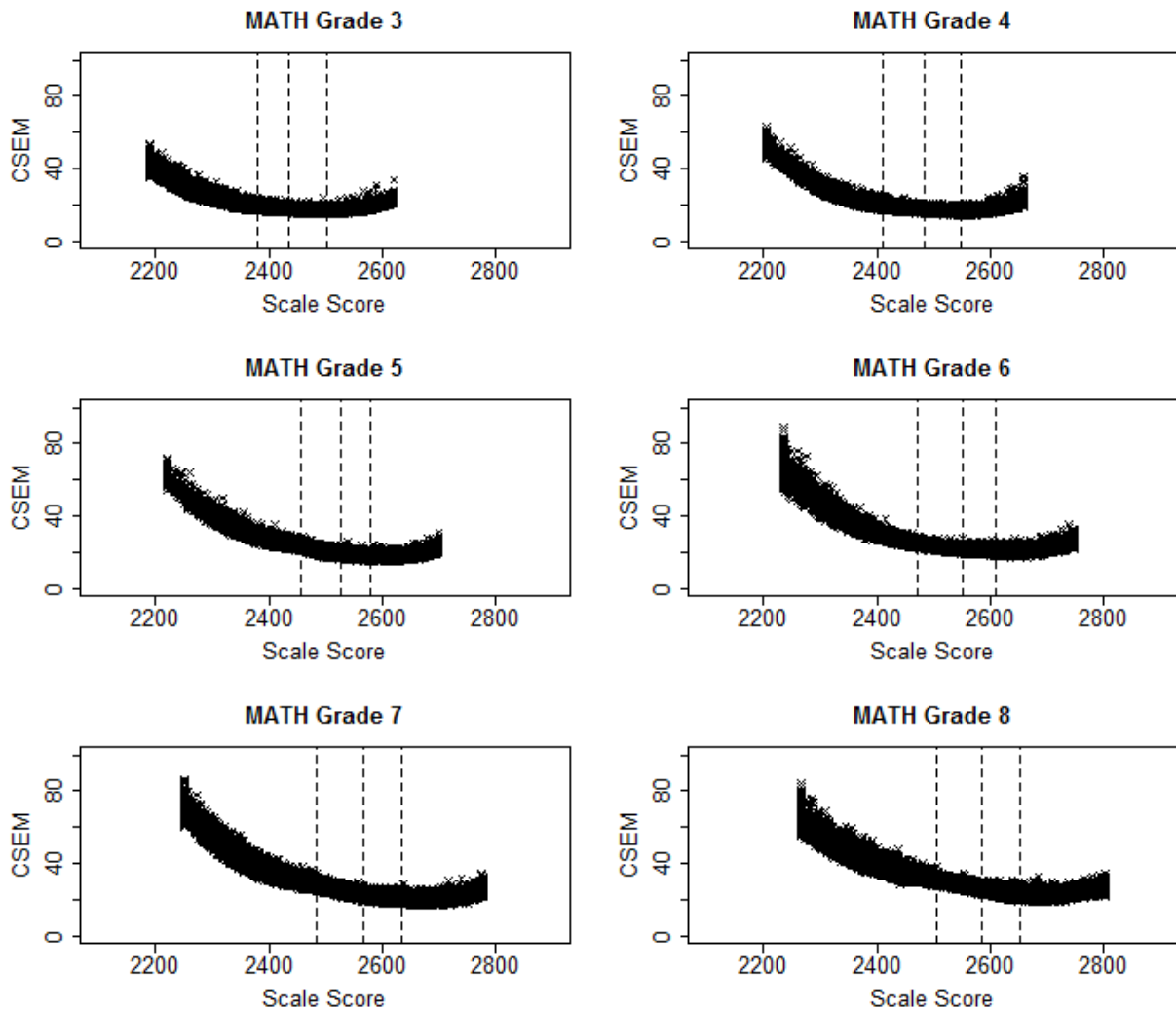


Figure 12. Conditional Standard Errors of Measurement for Mathematics



The CSEMs presented in Figures 11 and 12 are summarized in Tables 31 and 32. Table 31 provides the average CSEM for all scores and by achievement level. Table 32 presents the average CSEMs at each cut score and the difference in average CSEMs between two cut scores. As shown in Figures 11 and 12, the greatest average CSEM is in Level 1 in both English language arts/literacy (ELA/L) and mathematics. Average CSEMs at all cut scores are similar in ELA/L, but they are larger in Level 2 cut scores in mathematics.

Table 31. Average Conditional Standard Errors of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	30	25	26	27	27
4	32	29	28	30	30
5	31	27	28	31	30
6	33	29	30	32	31
7	38	30	31	32	33
8	38	31	30	32	33
Mathematics					
3	25	19	17	19	20
4	26	18	17	18	21
5	30	21	18	19	24
6	37	23	21	22	29
7	39	25	22	21	30
8	40	29	25	23	33

Table 32. Average Conditional Standard Errors of Measurement at Each Achievement Level Cut and Difference of the Standard Errors of Measurement between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	L2-L3	L3-L4	L2-L4
ELA/L						
3	25	25	26	0	0	0
4	29	28	28	0	1	1
5	27	28	29	1	1	2
6	28	29	30	1	1	2
7	30	30	31	0	1	1
8	31	31	30	1	0	1
Mathematics						
3	20	18	17	2	1	3
4	19	17	17	2	1	3
5	23	19	18	4	1	6
6	25	22	21	3	1	4
7	28	23	21	6	2	8
8	31	26	23	5	3	8

5.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When student performance is reported in terms of achievement levels, a reliability of achievement classification is computed in terms of the probabilities of accurate and consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The indices consider the accuracy and consistency of classifications.

For a fixed-form test, the accuracy and consistency of classifications are estimated on a single form's test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the computer-adaptive test (CAT), because the adaptive

testing algorithm constructs a test form unique to each student, the classification indices are computed based on all sets of items administered across students using an IRT-based method (Guo, 2006).

The classification index can be examined in terms of the classification accuracy and the classification consistency. Classification accuracy refers to the agreement between the classifications based on the form taken and the classifications that would be made based on the test takers' true scores if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who are consistently classified in the same achievement levels on two equivalent test forms.

In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, the classification accuracy and the classification consistency are estimated based on students' item scores, the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

For the i th student, the student's estimated ability is $\hat{\theta}_i$ with a SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$, assuming a normal distribution, where θ_i is the unknown true ability of the i th student and Φ the cumulative distribution function of the standard normal distribution. The probability of the true score at achievement level l based on the cut scores c_{l-1} and c_l is estimated as

$$\begin{aligned} p_{il} &= P(c_{l-1} \leq \theta_i < c_l) = P\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \leq \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) = P\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \leq \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) \\ &= \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right). \end{aligned}$$

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$, the above probabilities can be estimated directly using the likelihood function.

The likelihood function of theta, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, a probability of being at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and one minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, the various classification probabilities can be defined.

The probability of the i th student being classified at achievement level l ($l = 1, 2, \dots, L$) based on the cut scores cut_{l-1} and cut_l , given the student's item scores $\mathbf{z}_i = (z_{i1}, \dots, z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_J)$ and using the J administered items, can be estimated as

$$\begin{aligned} p_{il} &= P(cut_{l-1} \leq \theta_i < cut_l | \mathbf{z}, \mathbf{b}) = \frac{\int_{cut_{l-1}}^{cut_l} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta} \text{ for } l = 2, \dots, L - 1, \\ p_{i1} &= P(-\infty < \theta_i < cut_1 | \mathbf{z}, \mathbf{b}) = \frac{\int_{-\infty}^{cut_1} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta} \end{aligned}$$

$$p_{iL} = P(\text{cut}_{L-1} \leq \theta_i < \infty | \mathbf{z}, \mathbf{b}) = \frac{\int_{\text{cut}_{L-1}}^{\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta},$$

where the likelihood function based on general IRT models is

$$L(\theta | \mathbf{z}_i, \mathbf{b}) = \prod_{j \in d} \left(z_{ij} c_j + \frac{(1-c_j) \exp(z_{ij} D a_j (\theta - b_j))}{1 + \exp(D a_j (\theta - b_j))} \right) \prod_{j \in p} \left(\frac{\exp(D a_j (z_{ij} \theta - \sum_{k=1}^{K_j} b_{jk}))}{1 + \sum_{m=1}^{K_j} \exp(D a_j (\sum_{k=1}^m (\theta - b_{jk})))} \right),$$

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (a_j, b_j, c_j)$ if the j th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \dots, b_{jK_j})$ if the j th item is a polytomous item; a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and two-parameter logistic [2PL] models, $c_j = 0$), and D is 1.7 for non-Rasch models and 1 for Rasch model.

Classification Accuracy

Using p_{il} , a $L \times L$ table can be constructed as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix},$$

where $n_{alm} = \sum_{pl_i=l} p_{im} \cdot n_{alm}$ is the expected count of students at achievement level lm , pl_i is the i th student's achievement level, and p_{im} are the probabilities of the i th student being classified at achievement level m . In the above table, the row represents the observed level and the column represents the expected level.

The classification accuracy (CA) at level l ($l = 1, \dots, L$) is estimated by

$$CA_l = \frac{n_{all}}{\sum_{m=1}^L n_{alm}},$$

and the overall classification accuracy is estimated by

$$CA = \frac{\sum_{l=1}^L n_{all}}{N},$$

where N is the total number of students. Because classifying students as proficient or not proficient is such a high stakes decision, classification accuracy is also considered at the proficiency level by repeating the process for overall classification accuracy of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

Classification Consistency

Using p_{il} , which is similar to accuracy, another $L \times L$ table can be constructed by assuming the test is administered twice independently to the same student group

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix},$$

where $n_{clm} = \sum_{i=1}^N p_{il} p_{im} \cdot p_{il}$ and p_{im} are the probabilities of the i th student being classified at achievement level l and m , respectively based on observed scores and hypothetical scores from the equivalent test form.

The classification consistency (CC) at level l ($l = 1, \dots, L$) is estimated by

$$CC_l = \frac{n_{cll}}{\sum_{m=1}^L n_{clm}},$$

and the overall classification consistency is

$$CC = \frac{\sum_{l=1}^L n_{cll}}{N}.$$

As with classification accuracy, classification consistency is also considered at the proficiency level by repeating the process for overall classification consistency of achievement levels but with the four achievement levels collapsed into two proficiency categories: proficient (achievement levels 3 and 4) and not proficient (achievement levels 1 and 2).

The analysis of the classification index is performed based on overall scale scores. Table 33 provides the proportion of classification accuracy and consistency for overall, by achievement level, and at proficiency cut score.

The overall classification index ranged from 79% to 85% for the accuracy and from 70% to 79% for the consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher accuracy at L1 and L4 is due to the fact that the intervals used to compute the classification probabilities for students in L1 and L4 $[-\infty, L2 \text{ cut}; L4 \text{ cut}, \infty]$ are wider than the intervals used to compute the classification probabilities for students in L2 and L3 $[L2 \text{ cut}, L3 \text{ cut}; L3 \text{ cut}, L4 \text{ cut}]$. The misclassification probability tends to be higher for narrower intervals. Classification accuracy and classification consistency at the proficiency cut scores were high, ranging from 92% to 95% for accuracy and from 88% to 92% for consistency.

Accuracy of classifications is higher than the consistency of classifications in all achievement levels. The accuracy is higher than the consistency because the accuracy is based on one test with a measurement error and the true score while the consistency is based on two tests with measurement errors. The classification indices by subgroup are provided in Appendix C.

Table 33. Classification Accuracy and Consistency

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	80	72	83	76
	L1	91	86	89	85
	L2	69	58	73	62
	L3	65	54	78	70
	L4	87	82	90	85
	Proficiency Cut	93	90	94	92
4	Overall	79	71	85	79
	L1	90	86	91	87
	L2	60	48	80	72
	L3	63	52	79	71
	L4	88	81	90	85
	Proficiency Cut	92	89	95	92
5	Overall	80	72	84	77
	L1	91	86	92	88
	L2	64	52	77	67
	L3	72	62	71	61
	L4	86	80	90	85
	Proficiency Cut	93	90	95	92
6	Overall	79	70	84	77
	L1	90	84	93	89
	L2	69	58	77	69
	L3	72	63	71	60
	L4	84	75	89	84
	Proficiency Cut	92	88	94	91
7	Overall	79	71	84	78
	L1	90	85	92	88
	L2	67	56	76	66
	L3	75	67	74	65
	L4	83	74	90	85
	Proficiency Cut	92	88	94	91
8	Overall	79	71	83	77
	L1	89	83	91	88
	L2	69	58	72	61
	L3	76	68	71	59
	L4	84	75	90	85
	Proficiency Cut	92	89	94	92

5.4 RELIABILITY FOR SUBGROUPS

Tables 34 through 39 present the marginal reliability coefficients by the subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for limited English proficiency (LEP) and the Individuals with Disabilities Education Act (IDEA) subgroups. A large percentage of students in these subgroups received Level 1 with large SEMs.

Table 34. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 3–4)

Subgroup	Grade 3				Grade 4			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.92	2419	96	27	0.91	2463	102	30
Female	0.92	2426	95	27	0.91	2470	100	30
Male	0.92	2413	96	27	0.91	2457	103	30
Black or African American	0.90	2375	86	28	0.89	2415	93	30
AmerIndian/Alaskan	0.93	2390	105	27	0.90	2445	91	29
Asian	0.92	2466	93	27	0.90	2522	94	30
Hispanic or Latino	0.90	2374	89	28	0.90	2415	95	31
Pacific Islander	0.91	2405	86	26	0.91	2458	98	29
White	0.90	2454	86	27	0.89	2498	90	30
Multi-Racial	0.92	2436	94	27	0.91	2481	100	30
LEP	0.87	2347	80	28	0.87	2383	87	31
IDEA	0.87	2347	82	30	0.87	2384	90	32

Table 35. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 5–6)

Subgroup	Grade 5				Grade 6			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.92	2502	106	30	0.91	2521	102	31
Female	0.92	2510	104	30	0.91	2529	100	31
Male	0.92	2494	108	30	0.91	2512	103	31
Black or African American	0.90	2449	96	30	0.89	2476	91	31
AmerIndian/Alaskan	0.91	2467	98	29	0.87	2495	83	30
Asian	0.91	2562	99	30	0.90	2584	98	31
Hispanic or Latino	0.91	2453	101	30	0.90	2474	96	31
Pacific Islander	0.92	2481	106	30	0.88	2494	83	29
White	0.90	2539	94	30	0.89	2554	91	31
Multi-Racial	0.92	2512	105	29	0.91	2532	102	31
LEP	0.86	2401	83	31	0.79	2406	71	32
IDEA	0.88	2411	92	31	0.85	2431	84	32

Table 36. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L (Grades 7–8)

Subgroup	Grade 7				Grade 8			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.91	2541	109	33	0.91	2558	109	33
Female	0.90	2553	105	33	0.91	2572	106	33
Male	0.91	2529	112	33	0.91	2544	109	33
Black or African American	0.89	2496	102	34	0.89	2512	101	34
AmerIndian/Alaskan	0.90	2517	105	32	0.88	2523	92	32
Asian	0.89	2611	97	32	0.90	2626	101	32
Hispanic or Latino	0.90	2489	106	34	0.89	2510	104	34
Pacific Islander	0.90	2542	103	32	0.90	2561	126	39
White	0.89	2576	96	32	0.89	2590	97	32
Multi-Racial	0.91	2552	107	33	0.92	2568	111	32
LEP	0.79	2408	82	38	0.74	2422	74	38
IDEA	0.85	2443	96	37	0.84	2463	91	37

Table 37. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 3–4)

Subgroup	Grade 3				Grade 4			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.95	2426	92	20	0.95	2469	94	21
Female	0.95	2421	89	20	0.95	2465	90	20
Male	0.95	2430	95	21	0.95	2473	98	21
Black or African American	0.93	2374	83	22	0.93	2413	83	22
AmerIndian/Alaskan	0.95	2406	101	21	0.94	2442	88	21
Asian	0.95	2486	91	20	0.95	2537	86	19
Hispanic or Latino	0.93	2382	84	22	0.93	2423	86	22
Pacific Islander	0.95	2424	93	21	0.94	2466	83	20
White	0.94	2459	80	19	0.95	2503	81	19
Multi-Racial	0.95	2437	91	20	0.95	2482	95	20
LEP	0.92	2367	81	22	0.92	2405	82	23
IDEA	0.92	2351	89	25	0.92	2390	88	25

Table 38. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 5–6)

Subgroup	Grade 5				Grade 6			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.94	2493	98	24	0.94	2506	115	29
Female	0.94	2489	94	24	0.94	2504	112	28
Male	0.94	2497	102	24	0.94	2508	119	29
Black or African American	0.89	2433	83	27	0.90	2440	102	33
AmerIndian/Alaskan	0.90	2456	80	26	0.91	2486	88	27
Asian	0.95	2566	93	21	0.95	2594	112	25
Hispanic or Latino	0.91	2446	88	27	0.91	2450	107	33
Pacific Islander	0.92	2459	92	25	0.90	2481	94	29
White	0.94	2529	86	22	0.94	2548	98	25
Multi-Racial	0.94	2501	99	24	0.94	2515	118	28
LEP	0.85	2415	76	29	0.81	2391	88	38
IDEA	0.88	2411	86	30	0.86	2401	104	38

Table 39. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics (Grades 7–8)

Subgroup	Grade 7				Grade 8			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.93	2524	116	30	0.93	2532	124	33
Female	0.93	2522	112	30	0.93	2534	119	32
Male	0.94	2526	120	30	0.93	2531	128	33
Black or African American	0.88	2459	100	35	0.88	2463	106	37
AmerIndian/Alaskan	0.91	2497	107	32	0.89	2493	104	34
Asian	0.95	2621	111	25	0.95	2632	123	28
Hispanic or Latino	0.89	2466	103	34	0.88	2472	107	37
Pacific Islander	0.91	2524	96	29	0.94	2545	142	35
White	0.93	2566	102	26	0.93	2574	111	29
Multi-Racial	0.93	2536	117	30	0.94	2542	130	32
LEP	0.75	2404	81	40	0.70	2404	80	44
IDEA	0.84	2421	98	40	0.83	2426	97	41

5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability coefficients and the measurement errors are also computed for the claim scores. In both ELA/L and mathematics, claims 2 and 4 are combined to generate a score. Because the precision of scores in claims is insufficient to report scores given a small number of items, the scores on each claim are reported using one of the three achievement categories, considering the SEM of the claim score: (1) Below Standard, (2) At/Near Standard, or (3) Above Standard. Tables 40 and 41 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 40. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1: Reading	14–16	0.78	2423	103	48
	Claims 2 & 4: Writing & Research	16	0.85	2409	104	41
	Claim 3: Listening	8–9	0.57	2429	129	85
4	Claim 1: Reading	14–16	0.76	2462	112	55
	Claims 2 & 4: Writing & Research	16	0.82	2455	110	47
	Claim 3: Listening	8–9	0.60	2468	130	82
5	Claim 1: Reading	14–16	0.76	2502	115	57
	Claims 2 & 4: Writing & Research	16	0.85	2494	114	44
	Claim 3: Listening	8–9	0.61	2512	128	80
6	Claim 1: Reading	14–17	0.77	2519	112	54
	Claims 2 & 4: Writing & Research	16	0.81	2513	109	48
	Claim 3: Listening	8–9	0.56	2534	128	85
7	Claim 1: Reading	14–17	0.76	2548	118	58
	Claims 2 & 4: Writing & Research	16	0.81	2528	120	53
	Claim 3: Listening	8–9	0.62	2546	125	77
8	Claim 1: Reading	14–17	0.77	2560	119	57
	Claims 2 & 4: Writing & Research	16	0.80	2548	116	52
	Claim 3: Listening	8–9	0.62	2569	128	80

Table 41. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
3	Claim 1	20	0.92	2427	98	28
	Claims 2 & 4	9–11	0.69	2417	106	60
	Claim 3	9–10	0.75	2421	101	51
4	Claim 1	20	0.92	2470	100	29
	Claims 2 & 4	8–10	0.76	2459	106	52
	Claim 3	9–10	0.77	2464	103	49
5	Claim 1	20	0.90	2496	103	33
	Claims 2 & 4	8–10	0.64	2476	121	73
	Claim 3	9–10	0.72	2483	114	60
6	Claim 1	19	0.89	2508	124	40
	Claims 2 & 4	9–10	0.70	2493	130	70
	Claim 3	9–11	0.68	2502	124	71
7	Claim 1	20	0.88	2525	123	42
	Claims 2 & 4	9–10	0.62	2505	138	85
	Claim 3	8–10	0.70	2522	128	70
8	Claim 1	20	0.88	2535	130	44
	Claims 2 & 4	8–10	0.64	2516	145	87
	Claim 3	9–10	0.65	2524	137	81

Legend:

Claim 1: Concepts and Procedures

Claims 2 & 4: Problem Solving & Modeling and Data Analysis

Claim 3: Communicating Reasoning

6. SCORING

The Smarter Balanced Assessment Consortium provided the vertically scaled item parameters by linking across all grades using common items in adjacent grades. All scores are estimated based on these item parameters. Each student received an overall scale score, an overall achievement level, and a performance category for each claim. This section describes the rules used in generating scores, as well as the handscoring procedure.

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The Smarter Balanced assessments are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of item types.

Indexing items by i , the likelihood function based on the j th person's score pattern for I items is

$$L_j(\theta_j | \mathbf{z}_j, \mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_k) = \prod_{i=1}^I p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}),$$

where the vector $\mathbf{b}_i = (b_{i,1}, \dots, b_{i,m_i})$ for the i th item's step parameters, m_i is the maximum possible score of this item, a_i is the discrimination parameter for item i , z_{ij} is the observed item score for the person j , and k indices the step of the item i .

Depending on the item score points, the probability $p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of a two-parameter logistic (2PL) model for items with one point or the form based on the generalized partial credit model (GPCM) for items with two or more points.

In the case of items with one score point, $m_i = 1$,

$$p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \left\{ \begin{array}{l} \frac{\exp(Da_i(\theta_j - b_{i,1}))}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = p_{ij}, \text{ if } z_{ij} = 1 \\ \frac{1}{1 + \exp(Da_i(\theta_j - b_{i,1}))} = 1 - p_{ij}, \text{ if } z_{ij} = 0 \end{array} \right\};$$

in the case of items with two or more points,

$$p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \left\{ \begin{array}{l} \frac{\exp(\sum_{k=1}^{z_{ij}} Da_i(\theta_j - b_{i,k}))}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, \text{ if } z_{ij} > 0 \\ \frac{1}{s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})}, \text{ if } z_{ij} = 0 \end{array} \right\},$$

where $s_{ij}(\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = 1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\theta_j - b_{i,k}))$, and $D = 1.7$.

Standard Error of Measurement

With MLE, the standard error (SE) for student j is:

$$SE(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}},$$

where $I(\theta_j)$ is the test information for student j , calculated as

$$I(\theta_j) = \sum_{i=1}^I D^2 a_i^2 \left(\frac{\sum_{l=1}^{m_i} l^2 \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))} - \left(\frac{\sum_{l=1}^{m_i} l \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))} \right)^2 \right),$$

where m_i is the maximum possible score point (starting from 0) for the i th item, and D is the scale factor, 1.7. The SE is calculated based only on the answered items for both complete and incomplete tests. The upper bound of the SE is set to 2.5 on theta metric. Any value larger than 2.5 is truncated at 2.5 on theta metric.

The algorithm allows previously answered items to be changed; however, it does not allow items to be skipped. Item selection requires iteratively updating the estimate of the overall and strand ability estimates after each item is answered. When a previously answered item is changed, the proficiency estimate is adjusted to account for the changed responses when the next new item is selected. While the update of the ability estimates is performed at each iteration, the overall and claim scores are recalculated using all data at the end of the assessment for the final score.

6.2 RULES FOR TRANSFORMING THETA TO VERTICAL SCALE SCORES

The student’s performance in each subject is summarized in an overall test score referred to as a *scale score*. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced Assessment Consortium. Table 42 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 42. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA/L	3–8	85.8	2508.2
Mathematics	3–8	79.3	2514.9

Standard errors of the MLEs are transformed to be placed onto the reporting scale. This transformation is:

$$SE_{SS} = a * SE_{\theta},$$

where SE_{SS} is the standard error of the ability estimate on the reporting scale, SE_{θ} is the standard error of the ability estimate on the Θ scale, and a is the slope of the scaling constant that transforms Θ into the reporting scale.

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 43 provides three achievement standards for each grade and content area.

Table 43. Cut Scores in Scale Scores

Grade	ELA/L			Mathematics		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	2367	2432	2490	2381	2436	2501
4	2416	2473	2533	2411	2485	2549
5	2442	2502	2582	2455	2528	2579
6	2457	2531	2618	2473	2552	2610
7	2479	2552	2649	2484	2567	2635
8	2487	2567	2668	2504	2586	2653

6.3 LOWEST/HIGHEST OBTAINABLE SCORES (LOSS/HOSS)

Although the observed score is measured more precisely in an adaptive test than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include easy or difficult items to measure low- and high-performing students, the standard error could be large at the low and high ends of the ability range. The Smarter Balanced Assessment Consortium decided to truncate extreme unreliable student ability estimates. Table 44 presents the lowest obtainable theta score (LOT) or LOSS and the highest obtainable theta score (HOT) or HOSS. Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values and are assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and claim scores). The standard errors for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 44. Lowest and Highest Obtainable Scores

Subject	Grade	Theta Metric		Scale Score Metric	
		LOT	HOT	LOSS	HOSS
ELA/L	3	-4.5941	1.3374	2114	2623
	4	-4.3962	1.8014	2131	2663
	5	-3.5763	2.2498	2201	2701
	6	-3.4785	2.5140	2210	2724
	7	-2.9114	2.7547	2258	2745
	8	-2.5677	3.0430	2288	2769
Mathematics	3	-4.1132	1.3335	2189	2621
	4	-3.9204	1.8191	2204	2659
	5	-3.7276	2.3290	2219	2700
	6	-3.5348	2.9455	2235	2748
	7	-3.3420	3.3238	2250	2778
	8	-3.1492	3.6254	2265	2802

6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

In the item response theory (IRT) maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned.

6.5 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR CLAIM SCORES

In both English language arts/literacy (ELA/L) and mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim score, three performance categories' relative strengths and weaknesses are produced. The difference between the proficiency cut score and the claim score plus or minus 1.5 times standard error of the claim is used to determine the relative strengths and weaknesses.

For summative tests, the specific rules are as follows:

- Below Standard (Code = 1): if $\text{round}(SS_{rc} + 1.5 * SE(SS_{rc}), 0) < SS_p$
- At/Near Standard (Code = 2): if $\text{round}(SS_{rc} + 1.5 * SE(SS_{rc}), 0) \geq SS_p$ and $\text{round}(SS_{rc} - 1.5 * SE(SS_{rc}), 0) < SS_p$, a strength or weakness is indeterminable
- Above Standard (Code = 3): if $\text{round}(SS_{rc} - 1.5 * SE(SS_{rc}), 0) \geq SS_p$

where SS_{rc} is the student's scale score on a claim; SS_p is the proficiency scale score cut (Level 3 cut); and $SE(SS_{rc})$ is the standard error of the student's scale score on the claim. HOSS and LOSS are automatically assigned to *Above Standard* and *Below Standard*, respectively.

6.6 TARGET SCORES

The target-level reports are impossible to produce for a fixed-form test because the number of items included per target is too small to produce a reliable score at the target level. A typical fixed-form test includes only one or two items per target. Even when aggregated, these data narrowly reflect the benchmark because they reflect only one or two ways of measuring the target. However, an adaptive test offers a tremendous opportunity for target-level data at the class-, school-, and district-area levels. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given target. Target scores are computed for attempted tests based on the responded items. Target scores are computed in each of the four claims in ELA/L and claim 1 for mathematics.

Target scores are computed in two ways: (1) target scores relative to a student's overall estimated ability (θ), and (2) target scores relative to the proficiency standard (level 3 cut).

6.6.1 Target Scores Relative to Student's Overall Estimated Ability

By defining $p_{ij} = p(z_{ij} = 1)$, representing the probability that student j responds correctly to item i , z_{ij} represents the j th student's score on the i th item. For items with one score point, the 2PL IRT model is used to calculate the expected score on item i for student j with estimated ability $\hat{\theta}_j$ as:

$$E(z_{ij}) = \frac{\exp(Da_i(\hat{\theta}_j - b_i))}{1 + \exp(Da_i(\hat{\theta}_j - b_i))}$$

For items with two or more score points, using the GPCM, the expected score for student j with estimated ability $\hat{\theta}_j$ on an item i with a maximum possible score of m_i is calculated as:

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{\exp(\sum_{k=1}^l Da_i(\hat{\theta}_j - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\hat{\theta}_j - b_{i,k}))}$$

For each item i , the residual between observed and expected score for each student is defined as:

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, T .

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}$$

For an aggregate unit, a target score is computed by averaging the individual student target scores for that target across all students in the aggregate unit.

$$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} (\delta_{jT} - \bar{\delta}_{Tg})^2},$$

where n_g is the number of students who responded to any of the items that belong to the target T for an aggregate unit g . If a student did not happen to see any items on a particular target, the student is NOT included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates may indicate that a roster, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$ is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target.

Direct reporting of the statistic $\bar{\delta}_{Tg}$ is not suggested. Instead reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target is recommended. In some cases, insufficient information will be available, and that will be indicated, as well.

For target-level strengths and weaknesses, the following are reported:

- If $\bar{\delta}_{Tg} - se(\bar{\delta}_{Tg}) \geq 0.07$, then performance is better than on the overall test.
- If $\bar{\delta}_{Tg} + se(\bar{\delta}_{Tg}) \leq -0.07$, then performance is worse than on the overall test.
- Otherwise, performance is similar to performance on the overall test.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

6.6.2 Target Scores Relative to Proficiency Standard (Level 3 Cut)

By defining $p_{ij} = p(z_{ij} = 1)$, indicating the probability that student j responds correctly to item i . z_{ij} represents the j th student's score on the i th item. For items with one score point the 2PL IRT model is used to calculate the expected score on item i for student j with $\theta_{Level\ 3\ cut}$ as:

$$E(z_{ij}) = \frac{\exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}{1 + \exp(Da_i(\theta_{Level\ 3\ cut} - b_i))}$$

For items with two or more score points, using the GPCM, the expected score for student j with *Level 3 cut* on an item i with a maximum possible score of m_i is calculated as:

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{\exp(\sum_{k=1}^l Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l Da_i(\theta_{Level\ 3\ cut} - b_{i,k}))}$$

For each item i , the residual between observed and expected score for each student is defined as:

$$\delta_{ij} = z_{ij} - E(z_{ij})$$

Residuals are summed for items within a target. The sum of residuals is divided by the total number of points possible for items within the target, T .

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}$$

For an aggregate unit, a target score is computed by averaging the individual student target scores for that target across all students in the aggregate unit.

$$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}, \text{ and } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} (\delta_{jT} - \bar{\delta}_{Tg})^2},$$

where n_g is the number of students who responded to any of the items that belong to the target T for an aggregate unit g . If a student did not happen to see any items on a particular target, the student is NOT included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates may indicate that a class, teacher, school, or district is more effective (if $\bar{\delta}_{Tg}$ is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given target.

Direct reporting of the statistic $\bar{\delta}_{Tg}$ is not suggested. Instead reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target is recommended. In some cases, insufficient information will be available, and that will be indicated, as well.

For target-level strengths and weaknesses, the following are reported:

- If $\bar{\delta}_{Tg} - se(\bar{\delta}_{Tg}) \geq 0.07$, then performance is *above* the Proficiency Standard.
- If $\bar{\delta}_{Tg} + se(\bar{\delta}_{Tg}) \leq -0.07$, then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

6.7 HAND-SCORING

Constructed response short-answer (SA) items in both English language arts/literacy (ELA/L) and mathematics for the summative assessments administered by Cambium Assessment Inc. (CAI) are routed to Measurement Incorporated (MI) for scoring. MI provides hand-scoring using human raters (called handscoring). For the 2021-2022 Connecticut summative operational item pool, there were a total of 326 SA items in ELA/L and 300 SA items in mathematics. Table 45 shows the number of SA items by grade and subject.

Table 45. Number of Hand-Scored Items in 2021–2022 Connecticut Summative Item Pool, by Grade and Subject

Grade	ELA/L	Mathematics
3	54	46
4	58	52
5	68	74
6	53	52
7	43	35
8	50	41
Total	326	300

All guidelines for hand-scoring responses were specified by Smarter Balanced. Outlined below is the hand-scoring process MI followed in spring 2022 in accordance with the Smarter Balanced guidelines. This process applied to the scoring of all student constructed response SA items for ELA/L and mathematics.

6.7.1 Rater Selection

MI has developed a pool of over three thousand raters experienced in scoring the Smarter Balanced assessments. MI first recruited qualified raters who had experience scoring these assessments. Recent advancements in rater evaluation practices have allowed MI to estimate rater accuracy parameters for experienced Smarter Balanced raters; these data were used to recruit the most historically accurate raters. Once recruited, experienced raters were assigned to the content area and grade band(s) with which they were most experienced.

To supplement this pool, MI also recruited raters with experience successfully scoring other large-scale assessments. MI assigned those raters to the grade level, subject area, and item type for which they were most qualified based on their performance on similar projects. Returning raters were selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. Finally, MI targeted recruitment of new raters as needed, in an effort to continue to identify talent across the country that will best fulfill the hand-scoring requirements.

All raters possessed, at a minimum, a four-year college degree. MI collected proof of degree for all raters as a condition of employment. All raters resided in the United States, and properly completed Form I-9 to verify their identity and employment authorization. Raters' I-9 forms are retained on file as required by law and made available for inspection by authorized government officers as needed. MI is an equal-opportunity employer, and believes that a diverse work force is of the utmost importance. When hiring, MI strives to ensure the work force is diverse across age, ethnicity, gender, and other demographic groups.

In selecting team leaders who will monitor the raters, MI scoring leadership reviewed records of all returning staff. They looked for people who were experienced team leaders with a record of good performance on previous projects, and they also considered raters who had been recommended for promotion to the team leader position.

MI requires all hand-scoring project staff (scoring directors, team leaders, raters, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or viewing any secure project

materials. The employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or the scoring methods to any person.

6.7.2 Rater Training and Scoring

All raters hired to score the Smarter Balanced assessments were trained using the rubric(s), anchor sets, and training/qualifying sets provided by Smarter Balanced. These sets were created during the original field-test scoring in 2014 and approved by Smarter Balanced. The same anchor sets are used each year. Additionally, MI conducts an annual review of the rater agreement and scoring materials in order to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the subsequent operational administration.

Once hired, raters were assigned to a scoring group that corresponds to the subject/grade that they were deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Raters were trained to score a specific item group of SA (research, brief write, reading, and mathematics) items. Within each item group, raters were divided into teams supervised by team leaders and a scoring director. Each scoring director, team leader, and rater was assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual rater scores was minimized to allow the rater to quickly develop experience scoring responses to a given set of items.

All raters, regardless of experience, were required to train on all anchor and training sets. Following training, all raters were required to pass the qualification sets in order to prove that they understood and could apply the criteria accurately. Until a rater had trained and qualified successfully, the rater was not permitted to score any student responses. MI carefully orchestrated training so that raters understood that all scoring decisions must be grounded in the training materials. In addition, raters learned how to navigate the anchor set, developed the knowledge and flexibility needed to evaluate or escalate a variety of responses, and retained the necessary consistency to score all responses accurately.

In order to begin working, all scoring personnel logged in to MI's secure Scoring Resource Center (SRC). SRC includes all online training modules, serves as the portal to MI's Virtual Scoring Center (VSC) interface, and maintains the data repository of all scoring reports used for rater monitoring. MI's training system (VSC Train) provides a remote, secure application for training both team leaders and raters. VSC Train provided each trainee with a training lesson for each item that allowed the trainee to complete the following steps:

- 1) Review the anchor set(s)
- 2) Score the practice set(s)
- 3) Review an annotated version of the practice set(s) after submitting scores
- 4) Score the qualification sets

Training design varied slightly depending on Smarter Balanced item type:

- ELA/L brief write, reading, and research SA: Raters trained and qualified on a baseline lesson within a specific grade band and target. Qualification on the baseline lesson qualified the rater to score all items in that grade band and target.

- Mathematics SA: Raters trained and qualified on baseline lessons within a specific grade band. Qualification on a baseline lesson qualified the rater to score that item and all items associated with it; for items with no associated items, training was for the specific item.

Rater training time varied by grade and content area. Training for ELA/L brief write, ELA/L reading, research SA, and mathematics SA items could typically be accomplished in one day. Raters generally worked 6.5 hours per day, excluding breaks. Evening shift raters worked 3.75 hours, excluding breaks.

In addition to item-specific information, a variety of substantive procedural and policy information was provided to each trainee during training. This included information about “alert” responses and non-scorable responses, as well as instructions for how to communicate with leadership during hand scoring. This ensured that raters were fully prepared to hand-score responses and were also aware of all responsibilities and scoring requirements before they were allowed to begin scoring.

Each trainee’s practice and qualification results were reported to the team leaders and scoring director. Scoring leadership reviewed each trainee’s results, paying particular attention to frequently mis-scored responses.

Following training, all training materials remained available to raters throughout scoring via the VSC Score Resource Library. This library included the item and rubric, the annotated anchor and practice sets, and any supplemental materials that were required to ensure accurate completion of the scoring effort.

When scoring, raters had access only to those items for which they had successfully trained and qualified. The hand-scoring system sorts individual student responses into small sets of 5-10, grouped by item. When a rater is qualified to score multiple items, this approach eases cognitive load by presenting the rater with a scoring set in which all responses relate to the same item.

Raters were trained to recognize non-scorable responses, and these responses were systematically routed to scoring supervisors for final condition-code assignment per Smarter Balanced requirements. For some item types, condition-code responses were scored by scoring experts trained to specialize in the scoring of these types of responses.

An “alerts” procedure was explained to raters during training sessions, where raters are trained to recognize “alerts” in their various forms, including those for suicide, criminal activity, alcohol or drug use, extreme depression, violence, rape, sexual or physical abuse, self-harm, intent to harm others, and neglect.

Multiple strategies were employed to minimize rater bias during scoring. First, raters did not have access to any student identifiers. Unless the students signed their names, wrote about their hometowns, or in some way provided other identifying information as part of their response, the raters had no knowledge of student characteristics. Second, all raters were trained using Smarter Balanced–provided materials, which were approved as unbiased examples of responses at the various score points. Training involved constant comparisons with the rubric and anchor papers so that raters’ judgments were based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback was maintained to identify any issues. Specifically, raters were closely monitored during scoring, and any instances of raters making scoring decisions based on anything except the criteria were discussed with the raters. After this feedback had been provided, raters were further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback, they were dismissed.

Finally, a series of automated score verifications were implemented to further ensure the accuracy of scores. For example, a blank check was conducted, which reset scores when a condition code of “blank” was

assigned to a response that had one or more characters in the response string (e.g., a response comprised of spaces or tabs). In this case, only after three independent raters had assigned a condition code of “blank” to a response that appeared blank, but which included characters in the response string, was the score recorded. A similar check was run when a score or condition code other than “blank” was assigned to a response that included no characters in the response string. Automatic resetting of double-scored responses when two raters assign non-adjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provided an additional score verification. In addition to automatically resetting and rescored these responses, the raters’ information was captured in a report and reviewed by scoring directors, one of many tools used to determine retraining needs.

6.7.3 Rater Statistics and Monitoring

At a minimum, 10-15% (depending on state contractual requirements) of the hand-scored responses received blind double reads. Additionally, 5% of the responses scored comprised pre-approved validity responses. MI’s VSC system automatically and randomly routed the requisite number of responses to raters for second reads and validity in an inconspicuous manner. Raters had no means of discerning whether they were scoring a first read, a second read, or a validity response. This system also prohibited raters from being eligible to score second reads for responses they had already scored.

MI’s VSC scoring system randomly seeds validity responses among operational responses during scoring. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The “true” scores for these responses are entered into a validity database. Validity responses are indistinguishable from operational responses.

VSC reports provided real-time reports throughout the scoring effort. These reports were available for access by hand-scoring management. Inter-rater reliability reports provide the percentage of exact, adjacent, and non-adjacent agreement for scorable responses. Validity performance reports provide the percentage of exact, adjacent, and non-adjacent agreement for validity responses and were used to monitor drift. Score point frequency distribution reports provide the percentage per score point and include the mean and standard deviation for each item.

Years of Smarter Balanced hand-scoring has allowed MI to amass a longitudinal dataset of rater performance data. MI’s rater monitoring system uses validity responses calibrated to fit a unidimensional Item Response Theory (IRT) model for each content area/item type. Extensive metrics (inter-rater reliability, calibrated validity, and sub-pools for monitoring drift) calculated by the monitoring system were used to ensure accuracy and productivity throughout the hand-scoring of a project. The system generated automated measures of rater performance drawing on validity, IRR, and other performance data. Raters and scoring managers received daily, automated messages summarizing raters’ performance, ensuring all hand-scoring staff were aware of current performance and any issues that required attention. Additional outputs were also provided in manager-level reports and used to identify raters who required retraining and/or removal due to issues with accuracy and/or production. These data allowed scoring management to direct scoring leaders in review of specific VSC reports in order to determine the specific areas of attention required for any raters.

The monitoring system afforded the objective, dynamic identification of the most accurate and productive raters, referred to as “advanced raters.” Advanced rater status changed daily based on current rater performance to ensure that any rater drift did not negatively impact scoring accuracy. Advanced rater status was a precondition for conducting second readings.

Team leaders spot-checked (i.e., read behind) raters’ scoring to ensure that the raters were on target, and conducted one-on-one retraining sessions to address any problems found. At the beginning of the project, team leaders read behind every rater every day; they became more selective about the frequency and number of read-behinds as raters became more proficient at scoring.

6.7.4 Rater Retraining and Dismissal

Retraining was an ongoing process once scoring is underway. Daily analysis of the rater status reports enabled management personnel to identify individual or group retraining needs. When it became apparent that a whole team or group as having difficulty with a particular type of response, large group training sessions were conducted.

When read-behinds or daily statistics identified a rater who could not maintain acceptable agreement rates, the rater was retrained and monitored by scoring leadership personnel. Raters are released from the project if retraining is unsuccessful. In these situations, all items scored by a rater during the timeframe in question were identified, reset, and released back into the scoring pool. The aberrant rater’s scores were deleted, and the responses were redistributed to other qualified raters for rescoring.

6.7.5 Rater Agreement

Rater inter-rater reliability (IRR) was computed based only on scorable responses (numeric scores) scored by two independent raters. Non-scorable responses (e.g., off-topic, off-purpose, or foreign-language responses) that were scored by scoring leadership—and not by two independent raters—were excluded from IRR computations. For the hand-scored items, the human-human agreement was computed based on 2021–2022 Connecticut summative assessments.

All ELA/L SA items were scored using a 0–2 rubric. Mathematics SA items were scored using 0–1, 0–2, or 0–3 rubrics. Condition codes are scored as zero.

Tables 46 through 47 provide a summary of the human-human IRR based on items with a sample size greater than 50. The IRR is presented with mean of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and adjacent agreement, and the mean, minimum and maximum quadratic weighted kappa (QWK). The average number of responses, as well as minimum and maximum number of responses to a given item are presented as well.

Table 46. Inter-Rater Agreement for ELA/L Short-Answer Items

Grade	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK		
		Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	17	277	52	694	75	65	84	100	0.63	0.42	0.78
4	21	246	61	853	73	54	84	100	0.63	0.33	0.78
5	22	234	54	665	70	54	83	100	0.65	0.37	0.79
6	18	276	76	605	70	59	79	100	0.60	0.43	0.79
7	22	289	57	845	73	55	86	100	0.67	0.46	0.81
8	29	231	54	768	71	53	82	100	0.64	0.44	0.85

Table 47. Inter-Rater Agreement for Mathematics Items

Grade	Score Point Range	Number of Items	Number of Responses			%Exact			% (Exact+ Adjacent)	QWK		
			Mean	Min	Max	Mean	Min	Max		Mean	Min	Max
3	0–1	8	265	200	316	94	88	98	100	0.86	0.68	0.95
4	0–1	10	250	230	301	88	81	95	100	0.70	0.55	0.90
5	0–1	9	227	217	239	92	86	98	100	0.73	0.51	0.94
6	0–1	12	227	148	340	97	95	100	100	0.81	0.62	1.00
7	0–1	10	296	188	373	96	88	100	100	0.78	0.25	1.00
8	0–1	15	353	334	382	92	84	98	100	0.78	0.51	0.96
3	0–2	32	277	70	350	90	79	100	100	0.92	0.84	1.00
4	0–2	38	239	58	301	90	76	100	100	0.89	0.47	0.99
5	0–2	57	229	78	267	89	78	98	100	0.87	0.53	0.98
6	0–2	40	331	311	373	88	71	98	100	0.86	0.71	0.98
7	0–2	24	308	282	361	91	82	98	100	0.86	0.59	0.97
8	0–2	26	345	325	381	90	83	99	100	0.86	0.74	0.99
3	0–3	6	203	136	300	91	88	96	100	0.96	0.93	0.99
4	0–3	4	250	230	297	85	80	88	100	0.93	0.92	0.94
5	0–3	8	215	156	258	89	84	100	100	0.89	0.75	1.00
7	0–3	1	313	313	313	86	86	86	100	0.85	0.85	0.85

7. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and any handscored items are scored. Because the score reports on students' performance are updated each time that students complete tests and handscored items are scored, authorized users (e.g., school principals, teachers) can quickly access information on students' performance and use it to improve student learning. In addition to individual students' score reports, the CRS also produces aggregate score reports by class, school, and district. The timely accessibility of aggregate score reports helps users monitor students' performance in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year.

This section contains a description of the types of scores reported in the CRS and a description of the ways to interpret and use these scores in detail.

7.1 CENTRALIZED REPORTING SYSTEM

The CRS is designed to help educators and students answer questions about how well students have performed on English language arts/literacy (ELA/L) and mathematics assessments. The CRS is the online tool that provides all stakeholders with timely, relevant score reports. The CRS for the Smarter Balanced assessments has been designed such that score reports are easy to read and understand for all stakeholders. This is achieved by using plain, non-technical language to facilitate review by parents and the general public. The CRS is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar elements.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 48 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Centralized Reporting System User Guide*, located via a Help button on the CRS.

Table 48. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
District School Teacher Roster	<ul style="list-style-type: none"> • Number of students tested and percentage of students with Level 3 or 4 (for overall students and by subgroup) • Average scale score and standard error of average scale score (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test and by claims (for overall students and by subgroup) • Performance category in each target (overall students) • On-demand student roster report
Student	<ul style="list-style-type: none"> • Total scale score and standard error of measurement (SEM) • Achievement level on overall and claim scores with achievement-level descriptors • Reported Lexile® Measure and Quantile® Measure • Average scale scores and standard errors of average scale scores for student’s school, and district

Aggregate score reports at a selected aggregate level are provided for students overall and by subgroup. Users can see student assessment results in any of the subgroups. Table 49 presents the types of subgroups and subgroup category provided in the CRS.

Table 49. Types of Subgroups

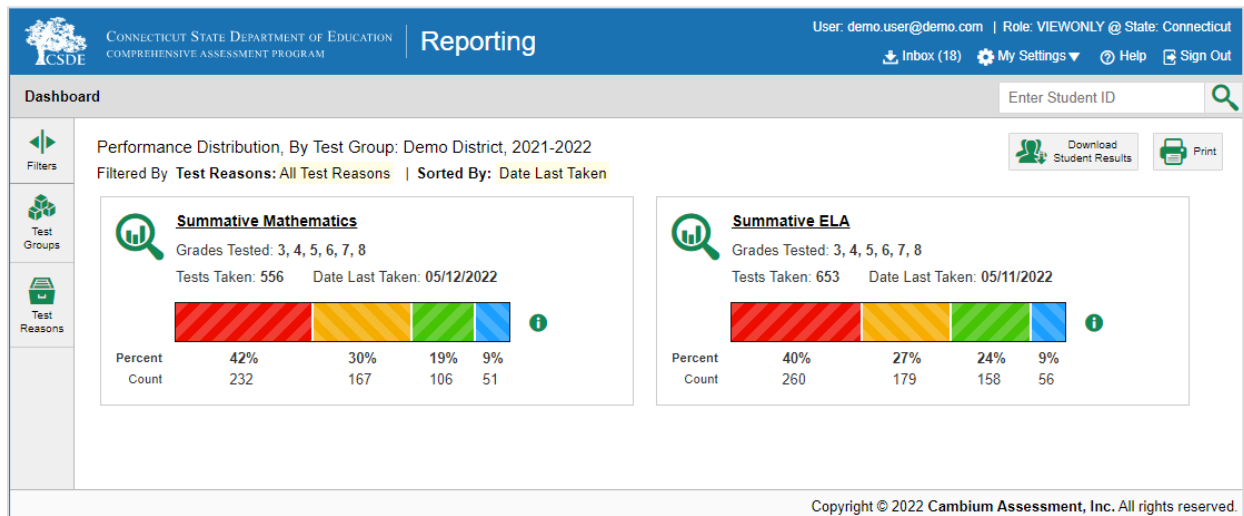
Subgroup	Subgroup Category
Gender	Male Female
IDEA Indicator	Special Education Not Special Education Unknown
Limited English Proficiency (LEP) Status	Yes No Unknown
Ethnicity/Race	American Indian or Alaskan Native Asian Black or African American Hispanic or Latino Native Hawaiian or Other Pacific Islander White Multi-Racial

7.1.1 Dashboard

Once authorized users in the district, school, and teacher level log in to the CRS, the dashboard page shows overall test results for all tests that the students have taken grouped by test family (e.g., Smarter Balanced Summative ELA/L). The dashboard summarizes students’ performance by test family for both ELA/L and mathematics across all grades, including (1) the grades of the students who have tested, (2) the number of tests taken, (3) the test date last taken, and (4) the percentage and counts of students at each achievement level. District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students.

Exhibit 1 presents an example dashboard page at the district level.

Exhibit 1. Dashboard: District Level



Once the user clicks the test family that he or she wants to explore further, it will take the user to the detailed dashboard, where the results are shown by test (e.g., Grade 3 ELA/L). The detailed dashboard page will appear by test in each grade. The detailed dashboard summarizes students’ performance by test in each grade, including (1) student count, (2) average scale score and standard error of the average scale score, (3) the percentage and counts of students at each achievement level, and (4) test date last taken.

Exhibit 2 presents an example detailed dashboard page for summative ELA/L at the district level.

Exhibit 2. Detailed Dashboard: District Level

Average Score and Performance Distribution, by Assessment: Demo District, 2021-2022
 Filtered By **School:** All Schools | **Test Reasons:** All Test Reasons |

Assessment Name	Test Group	Test Grade	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 7 ELA - Summative	Summative	7	Spring 2022 (Smarter Summative)	1406	2485 ± 3	 Percent: 47% 27% 20% 6% Count: 665 377 277 87	06/03/2022
Grade 5 ELA - Summative	Summative	5	Spring 2022 (Smarter Summative)	1353	2425 ± 3	 Percent: 51% 18% 12% 9% Count: 623 240 167 123	06/03/2022
Grade 8 ELA - Summative	Summative	8	Spring 2022 (Smarter Summative)	1481	2509 ± 3	 Percent: 41% 30% 21% 7% Count: 613 442 317 109	06/02/2022
Grade 4 ELA - Summative	Summative	4	Spring 2022 (Smarter Summative)	1269	2395 ± 3	 Percent: 62% 17% 12% 8% Count: 792 216 156 105	06/02/2022
Grade 3 ELA - Summative	Summative	3	Spring 2022 (Smarter Summative)	1307	2349 ± 2	 Percent: 63% 21% 11% 6% Count: 820 269 142 76	06/02/2022
Grade 6 ELA - Summative	Summative	6	Spring 2022 (Smarter Summative)	1376	2470 ± 3	 Percent: 47% 26% 20% 7% Count: 644 360 271 101	05/31/2022

7.1.2 Aggregate Score Reports: Overall Performance

Student performance for each grade in a subject area for a selected aggregate level is presented when users select a specific assessment name. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the aggregate unit both above and below the selected aggregate. For example, if a school is selected, the summary results of the district that the school belongs to are provided as well as the school summary results so that school performance can be compared with the other aggregate levels.

The aggregated summary report provides the summaries on a specific grade in a subject, including (1) student count, (2) the average scale score and standard error of the average scale score, (3) the percentage and counts of students in each achievement level, and (4) the percentage of proficient students. The summaries are also presented for students overall and by subgroup.

Exhibit 3 presents an example overall performance summary result for grade 3 ELA/L at the district level, and Exhibit 4 presents an example summary by gender.

Exhibit 3. Overall Performance Summary Results for Grade 3 ELA/L: District Level

Average Score and Performance Distribution for **Grade 3 ELA - Summative** (Spring 2022 (Smarter Summative)), by School and Reporting Category: Demo District, 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

School	Total	Total				Listening	Reading	Writing and Research/Inquiry
		Student Count	Average Scale Score	Performance Distribution	Percent At and Above Proficient			
District		65	2363 ± 10	Percent Count: 55% (36), 18% (12), 25% (16), 2% (1)	26%			
Demo School 1		12	2377 ± 21	Percent Count: 33% (4), 33% (4), 33% (4)	33%			
Demo School 2		22	2382 ± 19	Percent Count: 55% (12), 5% (1), 36% (8), 5% (1)	41%			
Demo School 3		9	2346 ± 21	Percent Count: 67% (6), 22% (2), 11% (1)	11%			
Demo School 4		2	2304 ± 26	Percent Count: 100% (2)	0%			
Demo School 5		20	2346 ± 17	Percent Count: 60% (12), 25% (5), 15% (3)	15%			

Exhibit 4. Overall Performance Summary Results for Grade 3 ELA/L by Gender: District Level

Breakdown of **Grade 3 ELA - Summative** (Spring 2022 (Smarter Summative)), by Gender: Demo District, 2021-2022
Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

Breakdown		Total	Total				Listening	Reading	Writing and Research/Inquiry
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent At and Above Proficient			
View Details	All		1307	2349 ± 2	Percent Count: 63% (820), 21% (269), 11% (142), 6% (76)	17%			
View Details	Male		672	2344 ± 3	Percent Count: 64% (430), 21% (143), 9% (63), 5% (36)	15%			
View Details	Female		635	2355 ± 3	Percent Count: 61% (390), 20% (126), 12% (79), 6% (40)	19%			

7.1.3 Aggregate Score Reports: Claim and Target Performance

Detailed summaries on aggregated claim and target results are also available on the same report page when a claim on the right side of the page is selected. For the claim result, (1) the average scale score and standard error of the average scale score and (2) performance distribution are presented. For the target result, the

strength or weakness indicators on each target within a claim are presented. These strength or weakness indicators are presented in two ways. The “Proficient?” measure indicates whether the group’s performance on each target is better than (check mark), less than (x mark), or not different from (half-filled circle) the proficiency standard for the selected test. The “Weak or Strong?” measure presents whether the group’s performance on each target is lower than (minus sign), higher than (plus sign), or not different from (equal sign) the group’s overall performance. If there is insufficient information in the “Proficient?” measure or “Weak or Strong?” measure, this is indicated with a star sign (*).

Like the overall performance summary results, the summary report presents results for the selected aggregate unit and the aggregate unit both above and below the selected aggregate unit. Also, the summaries on claim- and target- level performance can be presented for overall students and by subgroup.

Exhibit 5 present an example of claim- and target-level results for grade 5 mathematics at the district level.

Exhibit 5. Claim and Target Level Results for Grade 5 Mathematics: District Level

School		Average Claim Scale Score	Performance Distribution	Target A		Target B		Target C	
				Proficient?	Weak or Strong?	Proficient?	Weak or Strong?	Proficient?	Weak or Strong?
District		2414 ± 3	 Percent: 78% 15% 7% Count: 667 124 63	✗	=	✗	=	✗	=
Demo School A		2552 ± 16	 Percent: 24% 26% 50% Count: 10 11 21	✗	-	⊖	=	✓	+
Demo School B		2479 ± 13	 Percent: 48% 35% 17% Count: 22 16 8	✗	=	⊖	=	⊖	=
Demo School C		2454 ± 15	 Percent: 61% 20% 18% Count: 27 9 8	✗	=	✗	=	✗	=
Demo School D		2445 ± 24	 Percent: 50% 50% Count: 3 3	*	*	✗	-	✗	=

7.1.4 Roster Performance Report

Class, teacher, and school performance rosters provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report includes (1) the student’s overall subject scale scores with standard error of measurement, (2) reported Lexile for ELA/L or Quantile measure for mathematics tests, (2) the performance level and (3) performance category for each claim.

Exhibit 6 shows a sample roster performance report for grade 3 ELA/L.

Exhibit 6. Roster Performance Report for Grade 3 ELA/L

Score, Performance and Points Earned on **Grade 3 ELA - Summative** (Spring 2022 (Smarter Summative)) of Demo Roster, by Student and Reporting Category: 2021-2022
 Filtered By **School:** All Schools | **Test Reasons:** Spring 2022 (Smarter Summative) |

Breakdown By Download Student Results Print

Student	Student ID	Total	Scale Score	Reported Lexile® Measure	Performance	Listening	Reading	Writing and Research/Inquiry
District		2349 ± 2	n/a		 Percent: 63% 21% 11% 6% Count: 820 269 142 76			
School		2315 ± 8	n/a		 Percent: 76% 18% 4% 3% Count: 81 14 3 2			
My Students		2315 ± 8	n/a		 Percent: 76% 18% 4% 3% Count: 81 14 3 2			
Demo_Student A	123456789	2282 ± 28	290L		Level 1			
Demo_Student B	234567890	2301 ± 27	335L		Level 1			
Demo_Student C	345678901	2485 ± 26	770L		Level 3			
Demo_Student D	456789012	2262 ± 32	240L		Level 1			

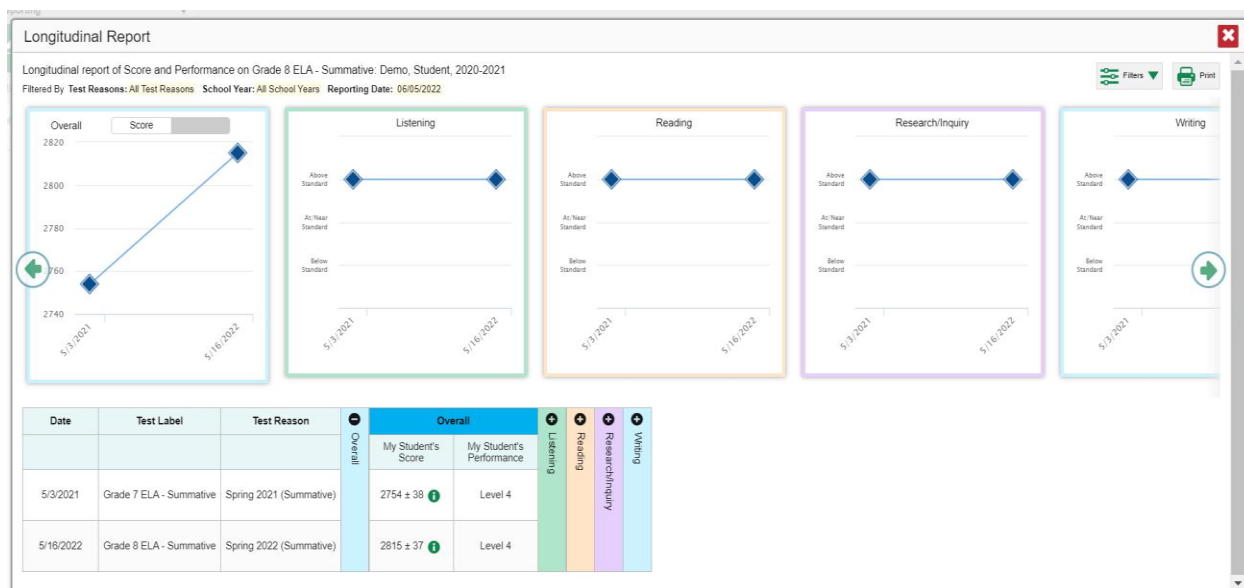
Rows per page: 80 Items: of 20

7.1.5 Trend Report

The trend (i.e., longitudinal) page provides the trend of student performance for an aggregate unit over time. The trend report can be set to plot either average scale scores or percentage of students in each achievement level on the graph for the selected aggregate unit. The trend report is also available at the individual student level.

Exhibit 7 presents an example trend report page for ELA/L at the individual student level.

Exhibit 7. Trend Report for ELA/L: Student Level



7.1.6 Individual Student Report


An individual student report can be generated and exported as a PDF file. The individual student report shows the student’s overall performance on the test with detailed information on multiple pages. In each subject area, the individual student report provides (1) the scale score and SEM; (2) achievement level for overall test; (3) reported Lexile measure for ELA/L or reported Quantile measure for mathematics; (4) performance category in each claim; and (5) average scale scores for the student’s district and school.

On the first page of the individual student report, the student’s name, scale score with the SEM, and achievement level, and reported Lexile measure for ELA/L or reported Quantile measure for mathematics are shown at the top of the page. In the middle section, the student’s performance is described in detail using a barrel chart. In the barrel chart, the student’s scale score is presented with the SEM using a “±” sign. SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered multiple times. Furthermore, in the barrel chart, achievement-level descriptors with cut scores at each achievement level are provided. This defines the content area knowledge, skills, and processes that test takers at the achievement level are expected to possess.

On the right side of the barrel chart, average scale scores and standard errors of the average scale scores for the student’s district and school are displayed so the student’s achievement can be compared with the above aggregate levels. It should be noted that the “±” next to the student’s scale score is the SEM of the scale score, whereas the “±” next to the average scale scores for aggregate levels represents the standard error of the average scale scores. On the bottom of the page, the student’s performance on each claim is displayed alongside a description of his or her performance on each claim.

Exhibits 8 presents an example of individual student reports for grade 5 ELA/L.

Exhibit 8. Individual Student Report for Grade 5 ELA/L



CONNECTICUT STATE DEPARTMENT OF EDUCATION
COMPREHENSIVE ASSESSMENT PROGRAM

Reporting

Individual Student Report

Demo, Student

Student ID: 999999999 | Student DOB: 10/12/2011 | Enrolled Grade: 5
Date Taken: 5/11/2022

Grade 5 ELA - Summative 2021-2022

Demo District
Demo School

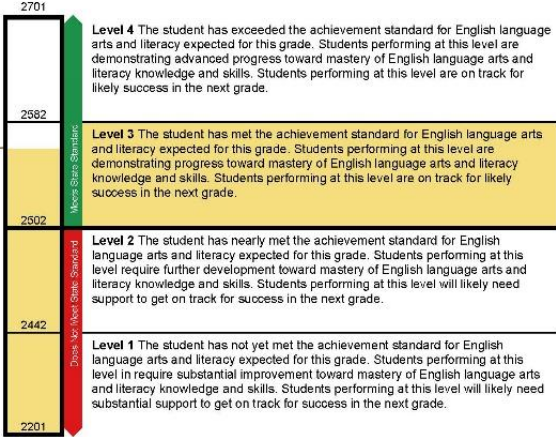
Scale Score: 2561±27

Reported Lexile® Measure: 1025L

Performance: Level 3

How Did Your Child Do on the Test?

Score
2561 ±27



Level 4 The student has exceeded the achievement standard for English language arts and literacy expected for this grade. Students performing at this level are demonstrating advanced progress toward mastery of English language arts and literacy knowledge and skills. Students performing at this level are on track for likely success in the next grade.

Level 3 The student has met the achievement standard for English language arts and literacy expected for this grade. Students performing at this level are demonstrating progress toward mastery of English language arts and literacy knowledge and skills. Students performing at this level are on track for likely success in the next grade.

Level 2 The student has nearly met the achievement standard for English language arts and literacy expected for this grade. Students performing at this level require further development toward mastery of English language arts and literacy knowledge and skills. Students performing at this level will likely need support to get on track for success in the next grade.

Level 1 The student has not yet met the achievement standard for English language arts and literacy expected for this grade. Students performing at this level in require substantial improvement toward mastery of English language arts and literacy knowledge and skills. Students performing at this level will likely need substantial support to get on track for success in the next grade.

How Does Your Child's Score Compare?

Name	Average Scale Score
Demo District	2427±3
Demo School	2377±9

Information on Standard Error of Measurement

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (±10) indicates a score range between 2290 and 2310.




Lexile® Information

The Lexile® Framework for Reading is a scientific approach to reading and text measurement. There are two Lexile measures: the Lexile reader measure and the Lexile text measure. A Lexile reader measure represents a person's reading ability on the Lexile scale. A Lexile text measure represents a text's difficulty level on the Lexile scale. When used together, they can help a reader choose a book or other reading material that is at an appropriate difficulty level.

How Did Your Child Perform on Different Areas of the Test?

The table and the graph below indicate student performance on individual reporting categories. The black dot indicates the student's score on each reporting category. The lines to the left and right of the dot show the range of likely scores your student would receive if he or she took the test multiple times.

▲ Below Standard
 ■ Approaching Standard
 ✔ Above Standard

Category	Performance	Performance	Performance level Description
Listening		■	Student may be able to employ effective listening skills for a range of purposes and audiences.
Reading		✔	Student can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Writing and Research/Inquiry		■	Student may be able to produce effective and well-grounded writing for a range of purposes and audiences. Student may be able to engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.

Generated on 5/25/2022


Page 1 of 1

Copyright © 2022 Cambium Assessment, Inc. All rights reserved.

7.1.7 Paper Family Score Reports

After the testing window is closed, parents whose children participated in a test receive a full-color paper score report (hereinafter referred to as a family report) including their child’s performance on ELA/L and mathematics. The family report includes information on student performance that is similar to the student detail page from the CRS with additional guidance on how to interpret student achievement results in the family report. An example of a family report is shown in Exhibit 9.

Exhibit 9. Sample Paper Family Score Report



CONNECTICUT STATE
DEPARTMENT OF EDUCATION
CSDE

Student Name: **Jane Doe**
 Grade: **3**
 Date of Birth: **05/20/2013**
 SASID: **1234567890**

School: **Demo Elementary School**
 District: **Demo District**
 Test Year: **2022**

Jane's ELA/Literacy Score for 2022

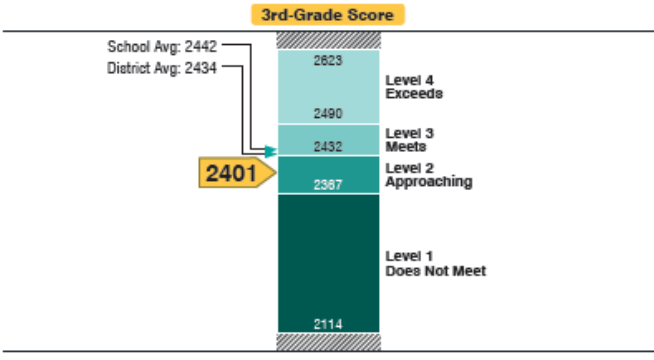
2401

Level 2
Approaching

Jane has **nearly met the achievement standard** for English language arts and literacy expected for this grade. Students performing at this standard **require further development toward mastery** of English language arts and literacy knowledge and skills. Students performing at this standard will likely need support to get on track for success in the next grade.

Areas of Knowledge and Skill	Performance
Reading	⚠ Below Standard
Listening	= Approaching Standard
Writing and Research/Inquiry	⚠ Below Standard

3rd-Grade Score



⚠ This area is outside the score range for that grade.

A student's test scores can vary if tests are taken several times. If Jane were tested again on ELA/literacy, the new scale score would probably fall between 2391 and 2411.

Jane's Mathematics Score for 2022

2410

Level 2
Approaching

Jane has **nearly met the achievement standard** for mathematics expected for this grade. Students performing at this standard **require further development toward mastery** of mathematics knowledge and skills. Students performing at this standard will likely need support to get on track for success in the next grade.

Areas of Knowledge and Skill	Performance
Concepts and Procedures	⚠ Below Standard
Problem Solving and Modeling & Data Analysis	⚠ Below Standard
Communicating Reasoning	= Approaching Standard

3rd-Grade Score



⚠ This area is outside the score range for that grade.

A student's test scores can vary if tests are taken several times. If Jane were tested again on mathematics, the new scale score would probably fall between 2400 and 2420.

7.2 INTERPRETATION OF REPORTED SCORES

A student’s performance on a test is reported in a scale score, an achievement level for the overall test, and an achievement level for each claim. Students’ scores and achievement levels are also summarized at the aggregate levels. The next section provides a description about how to interpret these scores.

7.2.1 Scale Score

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of the student’s knowledge and skills measured. The scale score is the transformed score from a theta score, which is estimated based on mathematical models. Low scale scores can be interpreted to mean that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted to mean that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. Interpretation of scale scores is more meaningful when the scale scores are used along with achievement levels and achievement-level descriptors.

7.2.2 Conditional Standard Error of Measurement

A scale score (the observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score will vary across administrations, sometimes being a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple times. When interpreting scale scores, it is recommended to consider the range of scale scores incorporating the SEM of the scale score.

The “±” next to the student’s scale score provides information about the certainty, or confidence, of the score’s interpretation. The boundaries of the score band are one SEM above and below the student’s observed scale score, representing a range of score values that is likely to contain the true score. For example, 2680 ± 10 indicates that if a student were tested again, it is likely that he or she would receive a score between 2670 and 2690. The SEM can be different for the same scale score, depending on how closely the administered items match the student’s ability.

7.2.3 Achievement Level

Achievement levels are proficiency categories on a test that students fall into based on their scale scores. For the Smarter Balanced assessments, scale scores are mapped into four achievement levels (Level 1, Level 2, Level 3, and Level 4) using three achievement standards (i.e., cut scores). Achievement-level descriptors are a description of content area knowledge and skills that test takers at each achievement level are expected to possess. Thus, achievement levels can be interpreted based on achievement-level descriptors. For the achievement level in ELA/L, for instance, achievement-level descriptors are described for grade 6 Level 3 as “The student has met the achievement standard and demonstrates progress toward mastery of the knowledge and skills in ELA/L needed for likely success in entry-level credit-bearing college coursework after high school.” Generally, students performing at Levels 3 and 4 on Smarter Balanced assessments are considered to be on track to demonstrating progress toward mastery of the knowledge and skills necessary for college and career readiness.

7.2.4 Performance Category for Claims

Students' performance on each claim is reported in three categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Unlike the achievement level for the overall test, student performance on each claim is evaluated with respect to the "Meets Standard" achievement standard. For students performing at "Below Standard" or "Above Standard," this can be interpreted to mean that their performance is clearly below or above the "Meets Standard" cut score for a specific claim. For students performing at "At/Near Standard," this can be interpreted to mean that their performance does not provide enough information to tell whether they reached the "Meets Standard" mark for the specific claim.

7.2.5 Performance Category for Targets

Teachers and educators sometimes need more detailed reports on student performance for instructional needs. The target report provides information on student performance about relative strength and weakness scores for each target within a claim. The strengths and weaknesses reports are generated for aggregate units of classroom, school, and district and provide information about how a group of students in a class, school, or district performed on each target, either relative to the proficiency standard (i.e., "Proficient?" target measure) or relative to their overall performance on the test ("Weak or Strong?" target measure). Target-level reports are produced for the aggregate units only, not for individual students, because each student is administered too few items in a target to produce a reliable score for each target.

For the "Proficient?" target measure, students' observed performance on items within the reporting element is compared to the expected performance on those items of someone who has an ability equal to the proficiency cut (i.e., the Achievement Level 3 cut). At the aggregate level, when observed performance within a target is greater than the proficiency cut, the reporting unit shows a relative strength in that target compared to the proficiency standard. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

For the "Weak or Strong?" target measure, students' observed performance on items within the reporting element is compared with the expected performance based on the overall ability estimate. At the aggregate level, when observed performance within a target is greater than the expected performance, then the reporting unit (e.g., roster, teacher, school, or district) shows a relative strength in that target. Conversely, when observed performance within a target is below the level expected based on overall achievement, then the reporting unit shows a relative weakness in that target.

Although performance categories for targets provide some evidence to help address students' strengths and weaknesses, they should not be over-interpreted because student performance on some targets may be based on relatively few items, especially for a small group.

7.2.6 Aggregated Scale Score

Students' scale scores are aggregated at roster, teacher, school, and district levels to represent how a group of students perform on a test. When students' scale scores are aggregated, the average scale scores can be interpreted as an estimate of the knowledge and skills that a group of students possess. Given that student scale scores are estimates, the average scale scores are also estimates and are subject to measures of uncertainty. In addition to the average scale scores, the percentage of students in each achievement level overall and by claim are reported at the aggregate level to represent how well a group of students performs.

7.3 APPROPRIATE USES OF TEST RESULTS

Assessment results can provide information about individual students' achievements on the test. Overall, assessment results show what students know and are able to do in certain subject areas and provide further information on whether students are on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can be used to identify students' relative strengths and weaknesses in certain content areas. For example, performance categories for targets can be used to identify a group's relative strengths and weaknesses among targets within a claim.

Assessment results on student achievement on the test can be used to help teachers or schools make decisions on how best to support students' learning. Aggregate score reports at the teacher and school level provide information regarding the strengths and weaknesses of their students and can be used to improve teaching and student learning. For example, a group of students may perform very well overall on the test but potentially not perform as well in several targets compared to their overall performance. In this case, teachers and schools would be able to identify the strengths and weaknesses of their students through the group performance by claim and target. They could then promote instruction in the specific claim or target areas in which their students perform relatively lower. Furthermore, by narrowing down the student performance results by subgroup, teachers and schools can determine which strategies may be best suited to improving student learning, particularly for students from disadvantaged subgroups. For example, teachers can examine student assessment results by LEP status and may observe that LEP students need help particularly in a certain specific area, such as reading literary responses and analysis. Teachers can then provide additional focused instruction for these students to enhance their achievement in any specific target or claim in which they are struggling.

In addition, assessment results can be used to compare performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in their school and district for overall scores and by claim. Although all students are administered different sets of items in each computer-adaptive test, scale scores are comparable across students. Furthermore, scale scores can be used to measure the growth of individual students over time when data are available. In the Smarter Balanced assessments, the scale scores across grades are on the same scale because the scores are vertically linked across grades. Therefore, scale scores from one grade can be compared with the next grade. i.e., measuring the growth.

While assessment results provide valuable information to understand students' performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and hence do not represent the precise measure for student performance. A student's scale score is associated with measurement error and thus users need to consider measurement error when using student scores to make decisions about student achievement. Moreover, although student scores may be used to help make important decisions about students' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student achievement, such as classroom assessment and teacher evaluation should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to consider the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

8. QUALITY CONTROL PROCEDURE

Quality assurance (QA) procedures are enforced through all stages of the Smarter Balanced assessment development, administration, and scoring and reporting of results. Cambium Assessment, Inc. (CAI) uses a series of quality control steps to ensure the error-free production of score reports in both online and paper-pencil formats. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

For the computer-adaptive test (CAT) component, a test configuration file is the key file that contains all specifications for the item selection algorithm and the scoring algorithm, such as the test blueprint, cut scores, the item information (i.e., answer keys, item attributes, item parameters, and passage information), and slopes and intercepts for theta-to-scale score transformation. The accuracy of the information in the configuration file is independently checked and confirmed before the testing window opens.

With the test configuration file, CAI uses simulated test administrations to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability. First, the simulator generates a sample of students with an ability distribution that matches that of the population in previous year's data. The ability of each simulated student is used to generate a sequence of item response scores while matching the blueprint and minimizing measurement error. These simulations provide a rigorous test of the adaptive algorithm. The results of these simulations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments.

After the adaptive testing simulations, another set of simulations for the combined tests (CAT and performance task [PT] components) are performed for scoring engine verification. The simulated data are generated such that verification of the scoring engine is based on a wide range of student response patterns. CAI rigorously checks whether the scoring rule specified in scoring specifications was applied accurately. The scores in the simulated data file are checked independently.

8.1.1 Platform Review

CAI's TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems such as Windows, Linux, and iOS to ensure that the item looks consistent in all of them. Some of the layouts have the stimulus and item response options/response area displayed side by side. In each of these layouts, both stimulus and response options have independent scroll bars.

Platform review is a process during which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in the Item Tracking System (ITS), and team members, each using a different platform, look at the same item to confirm that it renders as expected.

8.1.2 User Acceptance Testing and Final Review

Before deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the TDS serves as both a software evaluation and a content approval role. The UAT period provides the Department with an opportunity to interact with the exact test that the students will use.

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

The Smarter Balanced summative assessments are administered primarily online; however, a few students took paper-pencil assessments. When test documents were scanned, a quality control sample of documents consisting of 10 test cases per document type (normally between 500 and 600 documents) was created so that all possible responses and all demographic grids were verified, including various typical errors that required editing via Measurement Incorporated’s (MI) Data Inspection, Correction, and Entry (DICE) application program. This structured testing method provided exact test parameters and a methodical way of determining that the output received from the scanner(s) was correct. MI staff carefully compared the documents and the data file created from them to further ensure that the results from the scanner, the editing process (validation and data correction), and the transfer to the CAI database were correct.

8.3 QUALITY ASSURANCE IN DATA PREPARATION

CAI’s TDS has a real-time, built-in quality-monitoring component. After a test is administered to a student, the TDS passes the resulting data to CAI’s QA system. The QA system conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, and the total number of field-test items and operational items. The QA system ensures that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitor (QM) System to the Database of Record (DOR), which serves as the repository for all test information, and from which all test information for reporting is retrieved. The Data Extract Generator (DEG) is the tool that is used to retrieve data from the DOR for delivery to the Department. CAI staff ensures that data in the extract files match the DOR before delivering it to the Department.

8.4 QUALITY ASSURANCE IN ONLINE TEST DELIVERY SYSTEM

To monitor the performance of the TDS during the test administration window, CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic, state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI contracts for service in excess of this amount. Once deployed, the servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts CAI’s engineers at the first signs that trouble may be ahead. The applications log not only errors and exceptions, but also item response time information for critical database calls. This information enables CAI to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem. In addition, item response time data—such as data about how long it takes to load, view, or respond to an item—are captured for each assessed student. All of this information is logged as well, enabling CAI to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

A series of QA reports can also be generated at any time during the online assessment window, such as blueprint match rate, item exposure rate, and item statistics, for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved. In addition to these statistics, a cheating analysis report is produced to flag any unlikely patterns of behavior in a testing session, as discussed in Section 2.8, Data Forensic Program.

For example, an item statistics analysis report allows psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational testing window. The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators including item p -value and item discrimination index and item response theory item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside of a specified range are flagged for reporting or to generate reports based on all items in the pool.

For the CAT, other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to the blueprint and that items are performing as anticipated.

Table 50 presents an overview of the QA reports.

Table 50. Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology-enhanced items)
Blueprint Match Rates	To monitor unexpectedly low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

8.4.1 Score Report Quality Check

For the Smarter Balanced summative assessments, two types of score reports are produced: online reports and printed reports (family reports only).

8.4.1.1 Online Report Quality Assurance

Scores on the online assessments are assigned automatically by the systems in real time. Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QA checks and are uploaded to the

DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all the QA system’s validation checks. All of the previously mentioned processes take milliseconds to complete so that within less than one second after CAI receives handscores and they pass QA validation checks, the composite score will be available in the CRS.

8.4.1.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in CAI’s reporting specifications document. Upon approval of the specifications, analytic rules are programmed, and each program is extensively tested on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implement the agreed-on procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. The scripts are released for production when the output from both teams matches exactly.

Much of the statistical processing is repeated, and CAI has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. Small programs (called *macros*) are written to take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in CAI’s library for score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the director of score reporting, the director of psychometrics, and the project directors for affected projects.

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is mostly made up of calls to various macros, including macros that verify the data and conversion tables and the macros that perform the many complicated calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allows virtually infinite control of the visual appearance of the reports. After designers at CAI create backgrounds, CAI’s VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. CAI’s data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and are run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables CAI to test the entire system.

Programmed output goes through multiple stages of review and revision by graphics editors and the CAI Score Reporting team to ensure that design elements are accurately reproduced, and data are correctly displayed. Once CAI receives the final data and VIPP programs, the CAI Score Reporting team reviews proofs that contain actual data based on CAI’s standard quality assurance documentation. Several CAI staff

members review a large sample of the reports to ensure that all data are correctly placed on reports. This rigorous review is conducted over several days and takes place in a secure location in the CAI building. All reports containing actual data are stored in a locked storage area. Before the reports are printed, CAI provides a live data file and individual student reports with sample districts for Department staff review. CAI will work closely with the Department to resolve questions and correct any problems. The reports will not be delivered unless the Department approves the sample reports and data file.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York, NY: John Wiley & Sons, Inc.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.
- Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, *11*(6).
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253–264.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.
- Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, *16*(4), 247–260.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342.
- Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Philippine Statistician*, *52*(1–4), 81–92.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265–276.
- U.S. Department of Education. (2015). *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States*. Washington, D.C. Retrieved from <https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf>

APPENDICES

Appendix A: Summary of the 2021–2022 Interim Assessments

The Interim Comprehensive Assessments (ICAs) were fixed-form tests for each grade and subject. Most students took ICAs once, but some students took them multiple times. Table A-1 presents the number of students who took ICAs by the number of attempts. Total number of tests indicate the total ICA tests taken by the total number of students, counting multiple attempts as multiple tests. For example, if a student took ICAs twice, the number of tests for this student is counted as two. Table A-2 summarizes student performance on ICAs for all tests taken, including the average and the standard deviation of scale scores, the percentage of tests in each achievement level, and the percentage of proficient tests.

Table A-1. Number of Students Who Took ICAs

Grade	Number of Students by Number of Attempts					Total Number of Students	Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times		
ELA/L							
3	125	0	0	0	0	125	125
4	184	0	0	0	0	184	184
5	115	0	0	0	0	115	115
6	164	0	0	0	0	164	164
7	47	0	0	0	0	47	47
8	60	0	0	0	0	60	60
11	160	0	0	0	0	160	160
Mathematics							
3	202	2	0	0	0	204	206
4	119	0	0	0	0	119	119
5	98	0	0	0	0	98	98
6	200	0	0	0	0	200	200
7	112	0	0	0	0	112	112
8	143	0	0	0	0	143	143
11	165	0	0	0	0	165	165

Table A-2. ICA ELA/L and Mathematics Percentage of Tests in Achievement Levels

Subject	Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/L	3	125	2385	103	47	18	17	18	35
	4	184	2459	101	33	23	20	24	44
	5	115	2500	86	26	23	35	17	51
	6	164	2501	95	32	28	29	10	40
	7	47	2553	98	23	23	36	17	53
	8	60	2544	115	35	22	30	13	43
	11	160	2611	98	11	28	38	24	62
Math	3	206	2450	67	12	32	33	24	56
	4	119	2526	66	2	20	44	34	78
	5	98	2548	80	11	32	23	34	57
	6	200	2522	81	24	40	24	13	36
	7	112	2635	86	5	17	27	51	78
	8	143	2574	114	30	30	13	27	40
	11	165	2555	88	44	39	13	4	17

Note: The percentage of each achievement level may not add up to 100% or Percent Proficient due to rounding.

For the Interim Assessment Blocks (IABs), there were 14 to 15 IABs for English language arts/literacy (ELA/L) and 10 to 15 IABs for mathematics. Students were allowed to take as many IABs as they wanted, and to take the same IAB multiple times. Table A-3 shows the total number of students who took at least one IAB and the number of students by the number of distinct IABs taken. For example, in grade 3 ELA/L, a total of 23,667 students took at least one IAB. Among 23,667 students, 6,001 students took one IAB, 6,300 students took two distinct IABs, and so on. Tables A-4 to A-11 disaggregate the number of students in Table A-3 by each individual block. For example, . Among 6,001 students in grade 3 ELA/L who took only one IAB, 205 students took the Brief Writes IAB, 481 students took the Editing IAB, and so on.

Tables A-12 to A-17 summarize student performance on each IAB for all tests taken, including the percentage of tests in each performance category. The total number of tests indicates the total number of IAB tests taken by all students, counting multiple attempts as multiple tests. For example, if a student took the same IAB twice, the number of tests for this student is counted as two.

Table A-3. Number of Students Who Took Distinct IABs (Grades 3–8, 11)

Grade	Total Students with At Least One IAB	Number of IABs Taken														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>ELA/L</i>																
3	23,667	6,001	6,300	4,534	2,578	1,797	1,280	629	303	126	52	36	24	3	4	
4	24,184	6,869	6,899	4,581	2,791	1,210	904	399	302	127	44	37	8	12	1	
5	24,188	6,837	7,008	4,757	2,143	1,444	849	535	334	122	81	29	23	19	6	1
6	22,259	5,354	6,248	4,928	2,776	1,124	958	542	234	70	25					
7	21,775	6,206	6,468	4,251	2,269	1,425	654	173	169	122	38					
8	21,703	6,668	7,082	3,748	2,538	1,014	179	246	113	59	35	21				
11	1,080	846	232	2												
<i>Mathematics</i>																
3	26,446	7,858	8,218	5,387	2,187	1,580	736	280	111	46	30	13				
4	27,024	8,456	8,408	5,771	2,213	1,026	550	273	90	68	100	22	33	14		
5	25,887	8,169	8,591	5,298	2,174	780	312	166	149	120	87	37	4			
6	22,523	6,310	8,078	4,602	1,735	997	364	193	147	63	15	19				
7	22,625	7,441	8,243	3,839	1,826	834	356	77	7	2						
8	22,296	8,678	6,428	4,588	1,769	637	100	67	17	11	1					
11	650	374	226	50												

Table A-4: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 3–4)

Grade	Block	Number of Distinct IABs Taken														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	Brief Writes	205	1,591	526	622	645	501	263	191	83	28	33	23			
	Editing	481	785	1,240	1,397	1,112	975	533	285	105	43	31	24	3	4	
	Language and Vocabulary Use	496	1,098	1,170	955	921	676	535	257	86	41	35	24	3	4	
	Listen/Interpret	655	1,165	1,978	1,387	1,297	1031	569	293	102	45	24	24	3	4	
	Read Informational Texts	1,956	3,236	3,382	1,787	1,379	1041	444	273	110	46	24	14	3	4	
	Read Literary Texts	1,439	3,633	3,070	1,762	1,352	1088	493	286	123	51	35	24	3	4	
	Research	105	138	812	623	576	698	234	123	80	26	27	24	2	4	
	Research: Analyze Information	14	125	97	228	167	135	128	113	64	35	26	24	3	4	
	Research: Interpret and Integrate	351	122	197	114	109	219	298	116	50	30	25	23	3	4	
	Research: Use Evidence	8	137	155	106	132	143	82	29	75	26	24	13	3	4	
	Revision	155	175	460	719	621	524	242	127	70	22	27	11	3	4	
	Write & Revise Informational Texts	7	61	95	54	92	93	75	84	77	28	15	14	3	4	
	Write & Revise Narratives	53	93	206	246	273	242	205	122	44	35	29	24	2	4	
	Write & Revise Opinion Texts	44	79	136	195	196	224	228	105	53	40	27	11	2	4	
Performance Task	32	162	78	117	113	90	74	20	12	24	14	11	3	4		
4	Brief Writes	270	1,488	618	475	465	303	172	201	79	26	30	8	12	1	
	Editing	478	674	1,072	1,191	798	703	315	266	109	41	37	8	12	1	
	Language and Vocabulary Use	620	732	1,288	1,155	602	588	333	256	120	28	36	8	12	1	
	Listen/Interpret	493	1,033	1,846	1,660	945	814	350	266	118	37	30	8	12	1	
	Read Informational Texts	1,993	3,599	3,442	2,027	965	755	288	275	73	40	37	5	1	1	
	Read Literary Texts	1,809	4,755	3,091	1,998	836	695	268	199	101	42	37	8	12	1	
	Research	193	207	1,072	855	310	369	140	97	26	28	36	8	12	1	
	Research: Analyze Information	93	54	108	183	168	209	163	78	62	26	21	7	12	1	
	Research: Interpret and Integrate	393	96	137	171	115	158	200	155	88	25	33	8	12	1	
	Research: Use Evidence	40	39	81	120	79	88	79	72	34	15	22	8	11	1	
	Revision	223	513	365	663	387	369	144	160	75	23	9	7	12	1	
	Write & Revise Informational Texts	26	96	98	112	38	58	61	88	59	28	23	7	12	1	
	Write & Revise Narratives	111	212	253	356	183	190	161	163	102	22	17	3	12	1	
	Write & Revise Opinion Texts	69	197	188	157	134	89	105	125	89	30	16	3	12	1	
Performance Task	58	103	84	41	25	36	14	15	8	29	23					

Table A-5: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 5–6)

Grade	Block	Number of Distinct IABs Taken														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	Brief Writes	305	1,557	517	240	453	301	237	144	37	50	19	22	19	6	1
	Editing	786	1,133	819	779	950	674	427	252	110	77	25	23	19	6	1
	Language and Vocabulary Use	573	773	1,185	1,209	860	535	447	291	92	67	26	22	19	6	1
	Listen/Interpret	452	1,047	1,624	1,146	1,127	588	420	324	118	79	24	23	19	6	1
	Read Informational Texts	2,090	3,308	3,358	1,657	1,179	733	431	307	106	75	27	22	19	6	1
	Read Literary Texts	1,746	4,410	3,476	1,604	1,063	679	445	324	113	75	28	21	19	6	1
	Research	111	500	1,813	824	470	462	239	186	55	46	28	22	19	6	1
	Research: Analyze Information	27	171	331	135	133	134	95	108	77	38	11	16	11	1	1
	Research: Interpret and Integrate	338	175	223	252	157	172	174	168	79	36	26	22	17	6	1
	Research: Use Evidence	23	41	78	210	190	134	144	123	92	47	27	22	19	6	1
	Revision	171	561	537	223	416	282	294	133	84	68	27	22	19	6	1
	Write & Revise Informational Texts	2	21	16	21	10	24	18	5	9	3	4	2	3	5	1
	Write & Revise Narratives	156	96	190	91	77	171	173	161	54	76	17	15	19	6	1
	Write & Revise Opinion Texts	12	164	89	120	71	160	192	129	65	67	25	15	8	6	1
Performance Task	45	59	15	61	64	45	9	17	7	6	5	7	18	6	1	
6	Brief Writes	141	1,392	560	475	275	343	148	172	22	25					
	Editing	771	724	1,740	1,525	811	657	472	81	69	25					
	Language and Vocabulary Use	442	269	917	1,073	365	529	259	228	68	25					
	Listen/Interpret	521	677	1,652	1,223	562	429	424	120	70	25					
	Read Informational Texts	1,706	2,940	3,118	2,065	946	887	536	234	70	25					
	Read Literary Texts	583	3,819	3,635	2,001	845	527	457	116	70	25					
	Research	307	389	1,089	1,083	607	701	372	141	68	8					
	Research: Analyze & Integrate Info	476	821	277	185	209	365	106	189	11	25					
	Research: Evaluate Info & Sources	167	862	115	289	101	302	188	23	12	25					
	Research: Use Evidence	19	189	251	107	52	158	285	156	59	17					
	Revision	45	221	389	655	475	410	502	212	60	17					
	Write & Revise Explanatory Texts	25	40	123	81	83	53	13	37	42						
	Write & Revise Narratives	123	110	867	302	247	376	19	161							
	Performance Task	28	43	51	40	42	11	13	2	9	8					

Table A-6: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grades 7–8)

Grade	Block	Number of Distinct IABs Taken														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
7	Brief Writes	206	350	1,168	303	317	206	58	98	122	38					
	Editing	848	881	916	1,200	699	589	162	163	118	38					
	Language and Vocabulary Use	246	542	725	936	728	252	103	13							
	Listen/Interpret	584	993	646	419	518	108	129	162	120	38					
	Read Informational Texts	1,253	3,105	2,806	1,620	907	471	140	123	83	38					
	Read Literary Texts	1,373	3,289	3,449	1,767	1,228	462	123	165	121	38					
	Research	513	688	1,054	817	350	489	60	132	103	36					
	Research: Analyze & Integrate Info	368	925	177	510	357	496	97	66							
	Research: Evaluate Info & Sources	222	882	246	437	216	355	105	106	61	38					
	Research: Use Evidence	196	213	203	312	309	13	31	23	58	2					
	Revision	45	562	516	416	826	445	81	90	119	38					
	Write & Revise Argumentative Texts	123	307	738	145	410	2	35	82	121	38					
	Write & Revise Explanatory Texts	4	87	2	22	24	18	33	64	72	38					
Write & Revise Narratives	212	109	104	151	236	18	54	65								
Performance Task	13	3	3	21												
8	Brief Writes	233	133	79	366	321	69	35	14	17	8	4				
	Edit/Revise	1,487	1,045	1,661	1,875	926	119	141	101	43	23	21				
	Editing	226	225	241	145	173	99	232	101	41	28	21				
	Language and Vocabulary Use	70	361	560	653	193	55	165	20	34	34	21				
	Listen/Interpret	845	1,423	877	679	167	133	179	103	48	34	21				
	Read Informational Texts	1,334	3,785	1,984	1,776	700	132	170	113	55	35	21				
	Read Literary Texts	1,148	4,328	2,822	1,941	917	108	134	103	49	34	21				
	Research	396	571	1,362	1,009	523	58	82	99	42	8	4				
	Research: Analyze & Integrate Info	375	891	372	401	189	84	141	14	27	22	21				
	Research: Evaluate Info & Sources	301	696	404	537	323	38	64	87	42	27	17				
	Research: Use Evidence	47	335	251	529	410	103	93	95	43	27	17				
	Write & Revise Explanatory Texts	107	352	565	6	12	34	137	16	33	35	21				
	Write & Revise Narratives	45	18	65	207	214	42	148	35	57	35	21				
Performance Task	54	1	1	28	2		1	3								

Table A-7: ELA/L Number of Students Who Took Distinct IABs by Block Labels (Grade 11)

Grade	Block	Number of Distinct IABs Taken														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
11	Brief Writes															
	Editing															
	Language and Vocabulary Use			1												
	Listen/Interpret															
	Read Informational Texts	230	175	1												
	Read Literary Texts	6	130													
	Research	104	52	2												
	Research: Analyze & Integrate Info	48	57	1												
	Research: Evaluate Info & Sources		50	1												
	Research: Use Evidence															
	Revision	311														
	Write & Revise Argumentative Texts															
	Write & Revise Narratives															
	Performance Task	147														

Table A-8: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 3–4)

Grade	Block	Number of Distinct IABs Taken												
		1	2	3	4	5	6	7	8	9	10	11	12	13
3	Four Operations	730	718	666	469	446	317	166	39	20	30	13		
	Geometry	229	745	835	833	717	481	191	94	41	30	13		
	Linear and Area Measurement	158	431	231	318	341	236	184	84	34	30	13		
	Measurement and Data	402	910	916	804	848	355	189	79	39	2	13		
	Multiplication & Division	625	2,874	1,669	807	784	356	180	86	19	29	13		
	Multiply & Divide within 100	592	2,178	2,325	869	850	463	189	81	46	30	13		
	Number and Operations–Fractions	1,402	3,471	3,604	1,613	1,270	625	239	85	46	30	13		
	Number and Operations in Base 10	1,960	2,722	3,454	1,483	1,001	526	207	106	46	30	13		
	Operational and Algebraic Thinking	1,622	1,559	1,706	799	795	483	179	74	36	30	13		
	Properties of Multiplication & Division	79	412	400	398	398	308	74	62	44	29	13		
	Time, Volume, and Mass	16	170	233	297	367	216	160	97	35	30	13		
Performance Task	43	246	122	58	83	50	2	1	8					
4	Build Fractions from Unit Fractions	668	1,064	1,679	857	597	344	197	46	52	85	21	33	14
	Factors and Multiples	358	1,012	1,533	767	335	349	174	80	65	99	20	33	14
	Four Operations	443	762	586	285	163	235	117	28	59	65	22	32	14
	Fraction Equivalence and Ordering	486	2,072	1,305	1,068	583	337	238	67	41	74	18	32	14
	Fractions and Decimal Notation	330	643	983	748	571	346	155	76	53	84	20	31	14
	Generate and Analyze Patterns	14	31	98	58	73	115	99	34	47	78	17	32	14
	Geometry	358	997	1,113	516	384	255	144	47	61	95	17	32	14
	Measurement and Data	250	352	438	366	270	163	121	23	45	94	20	33	14
	Multidigit Arithmetic	476	1,591	574	582	482	219	126	65	58	84	21	32	14
	Number and Operations–Fractions	1,444	1,696	2,349	998	360	153	115	64	35	40	16	23	14
	Number and Operations in Base 10	2,767	4,383	3,431	1,232	452	273	147	75	38	60	22	33	14
	Operational and Algebraic Thinking	586	1,505	2,469	839	440	221	144	65	38	79	17	32	14
	Place Value & Multidigit Whole Numbers	227	512	731	501	414	270	118	42	13	46	11	18	14
Performance Task	49	196	24	35	6	20	16	8	7	17				

Table A-9: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 5–6)

Grade	Block	Number of Distinct IABs Taken											
		1	2	3	4	5	6	7	8	9	10	11	12
5	Add & Subtract with Equivalent Fractions	1,231	3,231	1,997	1,321	555	245	127	122	115	87	37	4
	Convert Measurements	364	139	91	140	232	183	89	70	43	62	37	4
	Geometry	354	423	738	596	417	144	106	116	113	87	37	4
	Measurement and Data	261	309	515	427	180	186	71	111	83	84	36	4
	Number and Operations–Fractions	1,319	3,583	3,223	1,097	390	176	107	119	104	80	37	4
	Number and Operations in Base 10	2,341	4,693	3,337	1,465	515	200	114	133	104	65	37	4
	Numerical Expressions	76	289	711	762	301	112	77	89	112	74	34	4
	Operations and Algebraic Thinking	697	848	341	247	128	107	91	124	105	83	37	4
	Operations with Whole Numbers & Decimals	669	792	2,049	729	268	76	113	127	106	87	37	4
	Place Value System	420	1,195	1,422	842	379	195	130	95	106	87	37	4
	Volume Concepts	380	1,496	1,431	1,001	475	233	134	85	88	73	37	4
	Performance Task	57	184	39	69	60	15	3	1	1	1	4	4
6	Algebraic Expressions	608	886	818	345	477	129	110	144	62	15	19	
	Dependent & Independent Variables	53	183	395	239	162	198	119	99	42	15	19	
	Divide Fractions by Fractions	1,133	2,451	1,334	788	693	216	159	86	56	15	19	
	Expressions and Equations	487	1,693	1,127	765	763	238	120	82	44	15	19	
	Geometry	278	1,019	1,658	974	329	222	99	46	48	7	16	
	Multidigit Numbers, Factors, & Multiples	271	2,171	495	418	227	158	105	136	63	15	19	
	One-Variable Expressions and Equations	169	929	1,823	731	402	246	112	79	35	14	19	
	Rational Number System II	128	619	373	532	172	137	84	123	59	13	18	
	Ratios and Proportional Relationships	2,112	3,770	3,520	908	771	256	166	125	61	12	19	
	Statistics and Probability	52	36	177	124	239	171	121	118	51	14	19	
	The Number System	969	2,319	1,964	846	653	172	141	130	45	15	19	
	Performance Task	50	80	122	270	97	41	15	8	1		4	

Table A-10: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grades 7–8)

Grade	Block	Number of Distinct IABs Taken												
		1	2	3	4	5	6	7	8	9	10	11	12	13
7	Algebraic Expressions and Equations	403	864	726	550	337	230	76	7	2				
	Angles, Areas, & Volume	139	329	697	428	411	166	63	7	2				
	Equivalent Expressions	275	835	628	768	415	260	29	7	2				
	Expressions and Equations	1,169	2,630	2,326	1,050	389	210	57	6	2				
	Geometric Figures	76	646	475	410	288	155	71	2	2				
	Geometry	289	352	359	265	311	172	67	7	2				
	Ratios and Proportional Relationships	2,968	5,288	2,860	1,633	726	302	66	7	2				
	Statistics and Probability	267	306	545	370	484	295	34	6	2				
	The Number System	1,840	5,109	2,855	1,625	700	330	76	7	2				
	Performance Task	15	127	46	205	109	16							
8	Analyze and Solve Linear Equations	673	2,136	2,145	737	408	67	59	15	11	1			
	Congruence and Similarity	766	1,504	1,335	770	368	87	65	17	11	1			
	Expressions and Equations I	869	1,232	2,127	1,104	308	39	62	17	11	1			
	Expressions and Equations II	465	489	1,247	745	162	62	57	10	10	1			
	Functions	1,337	1,798	2,131	811	380	95	63	13	11	1			
	Geometry	802	1,370	1,508	451	399	42	62	15	11	1			
	Proportional Relationships, Lines, & Linear Equations	1,194	2,435	2,210	1,217	571	69	23	17	11	1			
	The Number System	2,315	856	594	517	154	73	58	17	11	1			
	Volumes of Cylinders, Cones, & Spheres	204	913	367	502	340	37	20	15	11	1			
	Performance Task	53	123	100	222	95	29			1	1			

Table A-11: Mathematics Number of Students Who Took Distinct IABs by Block Labels (Grade 11)

Grade	Block	Number of Distinct IABs Taken												
		1	2	3	4	5	6	7	8	9	10	11	12	13
11	Algebraic Functions I	151	91	50										
	Algebraic Functions II	10	8	30										
	Create Equations: Linear & Exponential	12	80	33										
	Create Equations: Quadratic													
	Equations and Reasoning		3	17										
	Geometry & Right Angle Trigonometry	5	76											
	Geometry Congruence	6	76											
	Geometry Measurement & Modeling													
	Interpreting Functions		2	17										
	Number and Quantity	6												
	Seeing Structure in Expressions/Polynomial Expressions	22	58											
	Solve Equations & Inequalities: Linear & Exponential	134		3										
	Solve Equations & Inequalities: Quadratic	8	58											
	Statistics and Probability	20												
	Performance Task													

Table A-12: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels (Grades 3–5)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
3	Brief Writes	5,306	40	48	12
	Editing	7,306	30	44	26
	Language and Vocabulary Use	6,621	23	50	27
	Listen/Interpret	9,058	22	53	25
	Read Informational Texts	14,336	23	52	25
	Read Literary Texts	13,971	30	42	29
	Research	3,569	16	47	37
	Research: Analyze Information	1,195	16	47	37
	Research: Interpret and Integrate	2,013	25	43	32
	Research: Use Evidence	956	10	61	30
	Revision	3,305	24	52	24
	Write & Revise Informational Texts	706	20	58	21
	Write & Revise Narratives	1,667	28	59	13
	Write & Revise Opinion Texts	1,378	19	56	25
Performance Task	833	46	47	7	
4	Brief Writes	4,386	43	50	7
	Editing	6,160	27	51	23
	Language and Vocabulary Use	6,105	20	49	32
	Listen/Interpret	8,143	20	54	26
	Read Informational Texts	14,230	15	54	31
	Read Literary Texts	14,743	30	50	20
	Research	3,365	18	48	33
	Research: Analyze Information	1,267	23	46	31
	Research: Interpret and Integrate	1,948	26	42	32
	Research: Use Evidence	721	20	51	29
	Revision	3,041	24	56	20
	Write & Revise Informational Texts	714	25	62	13
	Write & Revise Narratives	1,814	25	59	16
	Write & Revise Opinion Texts	1,228	21	64	14
Performance Task	495	55	37	7	
5	Brief Writes	4,657	36	53	10
	Editing	6,694	20	46	34
	Language and Vocabulary Use	6,724	19	47	33
	Listen/Interpret	7,511	20	50	30
	Read Informational Texts	14,427	11	59	30
	Read Literary Texts	14,818	19	48	33
	Research	5,130	18	45	37
	Research: Analyze Information	1,361	24	50	26
	Research: Interpret and Integrate	2,123	29	40	31
	Research: Use Evidence	1,200	18	47	35
	Revision	2,903	22	51	27
	Write & Revise Informational Texts	144	19	68	13
	Write & Revise Narratives	1,348	16	48	37
	Write & Revise Opinion Texts	1,218	28	54	18
Performance Task	365	40	47	13	

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-13: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels (Grades 6–8)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
6	Brief Writes	4,133	34	56	10
	Editing	7,750	22	59	19
	Language and Vocabulary Use	4,337	24	50	26
	Listen/Interpret	5,887	15	53	32
	Read Informational Texts	13,729	20	54	27
	Read Literary Texts	12,868	23	54	22
	Research	5,625	21	51	28
	Research: Analyze & Integrate Info	3,378	16	64	20
	Research: Evaluate Info & Sources	2,106	23	55	23
	Research: Use Evidence	1,544	17	60	23
	Revision	3,509	35	53	12
	Write & Revise Explanatory Texts	499	26	61	14
	Write & Revise Narratives	2,226	17	58	25
	Performance Task	288	40	51	9
7	Brief Writes	3,041	38	52	10
	Editing	6,004	14	70	16
	Language and Vocabulary Use	3,927	25	49	26
	Listen/Interpret	3,918	18	59	24
	Read Informational Texts	11,769	26	46	28
	Read Literary Texts	13,282	25	49	26
	Research	4,408	13	60	27
	Research: Analyze & Integrate Info	3,283	21	60	18
	Research: Evaluate Info & Sources	2,688	29	45	26
	Research: Use Evidence	1,368	19	55	26
	Revision	3,645	27	55	18
	Write & Revise Argumentative Texts	2,070	13	62	25
	Write & Revise Explanatory Texts	364	13	58	29
	Write & Revise Narratives	963	19	62	19
Performance Task	40	63	30	8	
8	Brief Writes	1,307	28	63	8
	Edit/Revise	8,407	20	54	26
	Editing	1,551	22	51	27
	Language and Vocabulary Use	2,706	19	58	24
	Listen/Interpret	4,758	18	58	24
	Read Informational Texts	11,431	18	49	33
	Read Literary Texts	12,382	28	48	24
	Research	4,326	22	52	27
	Research: Analyze & Integrate Info	2,834	33	50	17
	Research: Evaluate Info & Sources	2,585	33	41	26
	Research: Use Evidence	1,981	24	54	22
	Write & Revise Explanatory Texts	1,323	17	54	29
	Write & Revise Narratives	891	17	58	25
	Performance Task	91	56	42	2

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-14: ELA/L Percentage of Tests in Performance Categories by IAB Block Labels
(Grade 11)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
11	Brief Writes				
	Editing				
	Language and Vocabulary Use	1*			
	Listen/Interpret				
	Read Informational Texts	644	46	40	14
	Read Literary Texts	137	45	43	12
	Research	158	18	54	28
	Research: Analyze & Integrate Info	106	8	40	52
	Research: Evaluate Info & Sources	51	14	65	22
	Research: Use Evidence				
	Revision	311	22	50	28
	Write & Revise Argumentative Texts				
	Write & Revise Narratives				
Performance Task	148	39	53	7	

Note: The percentage of each performance category may not add up to 100% due to rounding.

* Data suppressed due to small sample size, $n < 10$.

Table A-15: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels
(Grades 3–5)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
3	Four Operations	3,793	31	43	25
	Geometry	4,277	17	48	35
	Linear and Area Measurement	2,110	13	40	47
	Measurement and Data	4,846	24	40	36
	Multiplication & Division	7,747	27	43	30
	Multiply & Divide within 100	8,146	44	28	28
	Number and Operations–Fractions	13,155	14	42	44
	Number and Operations in Base 10	12,285	35	37	28
	Operational and Algebraic Thinking	8,243	40	43	17
	Properties of Multiplication & Division	2,391	22	44	34
	Time, Volume, and Mass	1,659	15	36	49
	Performance Task	633	21	66	13
4	Build Fractions from Unit Fractions	6,161	13	36	50
	Factors and Multiples	5,082	26	45	29
	Four Operations	2,896	36	35	29
	Fraction Equivalence and Ordering	6,658	28	31	42
	Fractions and Decimal Notation	4,188	9	33	58
	Generate and Analyze Patterns	730	14	47	39
	Geometry	4,353	9	64	28
	Measurement and Data	2,291	14	47	39
	Multidigit Arithmetic	4,503	33	47	20
	Number and Operations–Fractions	7,892	32	37	31
	Number and Operations in Base 10	14,544	40	41	19
	Operational and Algebraic Thinking	6,773	35	45	20
Place Value & Multidigit Whole Numbers	2,978	18	44	37	
Performance Task	392	22	66	12	
5	Add & Subtract with Equivalent Fractions	9,812	31	29	40
	Convert Measurements	1,558	22	34	44
	Geometry	3,274	20	51	29
	Measurement and Data	2,490	27	38	34
	Number and Operations–Fractions	11,351	42	37	20
	Number and Operations in Base 10	14,333	39	41	20
	Numerical Expressions	2,899	23	37	40
	Operations and Algebraic Thinking	2,850	23	44	33
	Operations with Whole Numbers & Decimals	5,545	36	39	25
	Place Value System	5,365	24	34	42
	Volume Concepts	5,793	12	42	46
Performance Task	438	31	57	12	

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-16: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels
(Grades 6–8)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
6	Algebraic Expressions	3,951	14	39	47
	Dependent & Independent Variables	1,642	28	43	29
	Divide Fractions by Fractions	7,404	18	31	52
	Expressions and Equations	5,824	31	39	30
	Geometry	4,804	23	49	28
	Multidigit Numbers, Factors, & Multiples	4,462	31	45	24
	One-Variable Expressions and Equations	4,591	25	39	36
	Rational Number System II	2,344	7	52	41
	Ratios and Proportional Relationships	13,190	39	33	28
	Statistics and Probability	1,152	12	38	50
	The Number System	7,977	41	36	22
	Performance Task	691	25	71	4
7	Algebraic Expressions and Equations	3,389	30	47	23
	Angles, Areas, & Volume	2,281	15	43	42
	Equivalent Expressions	3,443	14	34	52
	Expressions and Equations	8,111	26	43	30
	Geometric Figures	2,477	29	39	32
	Geometry	1,874	12	55	32
	Ratios and Proportional Relationships	15,516	26	51	23
	Statistics and Probability	2,525	23	53	24
	The Number System	13,501	31	46	22
	Performance Task	547	37	56	7
8	Analyze and Solve Linear Equations	6,595	22	46	32
	Congruence and Similarity	5,072	30	49	21
	Expressions and Equations I	6,057	32	48	19
	Expressions and Equations II	3,371	25	44	31
	Functions	7,115	33	41	26
	Geometry	4,919	25	45	30
	Proportional Relationships, Lines, & Linear Equations	7,909	21	47	32
	The Number System	5,198	32	36	32
	Volumes of Cylinders, Cones, & Spheres	2,579	9	19	71
	Performance Task	625	29	64	7

Note: The percentage of each performance category may not add up to 100% due to rounding.

Table A-17: Mathematics Percentage of Tests in Performance Categories by IAB Block Labels
(Grade 11)

Grade	Block	Total Number of Tests Taken	% Below	% At/Near	% Above
11	Algebraic Functions I	365	61	32	7
	Algebraic Functions II	48	35	56	8
	Create Equations: Linear & Exponential	191	80	18	2
	Create Equations: Quadratic				
	Equations and Reasoning	20	15	65	20
	Geometry & Right Angle Trigonometry	81	4	30	67
	Geometry Congruence	82	9	71	21
	Geometry Measurement & Modeling				
	Interpreting Functions	19	11	74	16
	Number and Quantity	6*			
	Seeing Structure in Expressions/Polynomial Expressions	137	66	30	4
	Solve Equations & Inequalities: Linear & Exponential	137	14	59	27
	Solve Equations & Inequalities: Quadratic	66	11	64	26
	Statistics and Probability	20		70	30
	Performance Task				

Note: The percentage of each performance category may not add up to 100% due to rounding.

* Data suppressed due to small sample size, $n < 10$.

Appendix B: Student Performance Across Four Years for All Students and by Subgroups

Table B-1. ELA/L Student Performance Across Four Years (Grades 3 and 4)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	37,525	53	2435	90	36,516	54	2437	91	34,389	45	2416	95	35,315	47	2419	96
Female	18,417	57	2443	88	17,890	58	2445	89	16,893	48	2424	95	17,296	49	2426	95
Male	19,108	49	2427	91	18,626	51	2429	92	17,494	42	2409	96	18,017	44	2413	96
Black or African American	4,764	33	2395	84	4,603	34	2395	86	4,123	22	2366	84	4,217	27	2375	86
AmerIndian/Alaskan	110	50	2422	87	101	48	2416	83	70	36	2399	88	114	36	2390	105
Asian	2,022	73	2479	85	1,945	73	2481	87	1,959	68	2467	89	1,911	67	2466	93
Hispanic or Latino	10,287	32	2392	84	10,122	35	2397	87	9,734	24	2368	86	10,855	27	2374	89
Pacific Islander	46	46	2438	85	29	45	2413	70	42	52	2415	99	29	38	2405	86
White	18,889	67	2464	80	18,236	67	2464	82	16,845	60	2449	86	16,485	62	2454	86
Multi-Racial	1,407	58	2445	90	1,480	58	2447	93	1,616	52	2435	92	1,704	54	2436	94
LEP	4,153	18	2360	76	4,287	22	2369	79	4,279	16	2347	79	4,710	15	2347	80
IDEA	4,871	16	2355	78	5,018	18	2358	80	4,922	16	2348	82	5,153	16	2347	82
Grade 4																
All Students	38,376	55	2479	97	37,727	55	2478	99	34,883	47	2458	101	35,940	49	2463	102
Female	18,646	59	2488	95	18,486	58	2487	96	17,159	49	2465	99	17,683	51	2470	100
Male	19,730	52	2470	99	19,239	51	2470	101	17,721	45	2451	102	18,253	47	2457	103
Black or African American	4,854	34	2431	90	4,820	34	2432	91	4,271	24	2407	90	4,392	27	2415	93
AmerIndian/Alaskan	105	41	2451	85	104	49	2463	87	79	28	2415	95	82	38	2445	91
Asian	2,010	75	2525	89	2,015	76	2530	91	1,859	71	2515	93	1,975	72	2522	94
Hispanic or Latino	10,195	35	2432	93	10,477	35	2432	93	9,887	27	2410	94	10,663	29	2415	95
Pacific Islander	37	65	2502	93	42	55	2476	78	34	38	2429	99	47	45	2458	98
White	19,781	68	2509	87	18,857	68	2510	89	17,174	61	2492	90	17,071	64	2498	90
Multi-Racial	1,394	59	2490	100	1,412	58	2489	97	1,579	49	2465	99	1,710	56	2481	100
LEP	3,776	18	2392	83	3,999	18	2391	84	3,927	13	2375	83	4,590	17	2383	87
IDEA	5,174	17	2388	86	5,443	18	2389	87	5,259	15	2378	86	5,642	18	2384	90

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table B-2. ELA/L Student Performance Across Four Years (Grades 5 and 6)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	39,594	58	2517	98	38,605	58	2516	100	35,310	50	2499	105	36,300	52	2502	106
Female	19,454	63	2528	95	18,733	63	2528	97	17,330	53	2507	103	17,859	55	2510	104
Male	20,140	54	2506	100	19,871	54	2506	102	17,977	47	2491	106	18,432	49	2494	108
Black or African American	5,034	36	2467	90	4,955	36	2466	94	4,368	26	2441	93	4,561	29	2449	96
AmerIndian/Alaskan	82	55	2489	100	111	43	2479	96	100	35	2468	89	75	43	2467	98
Asian	2,109	79	2571	90	2,003	78	2568	90	1,866	73	2558	97	1,884	75	2562	99
Hispanic or Latino	10,458	38	2470	94	10,371	38	2470	95	9,905	30	2449	98	10,822	32	2453	101
Pacific Islander	49	43	2495	101	36	67	2526	101	33	42	2470	90	35	49	2481	106
White	20,476	72	2547	87	19,683	72	2547	89	17,521	65	2534	93	17,262	67	2539	94
Multi-Racial	1,386	63	2529	95	1,446	61	2526	103	1,517	56	2513	102	1,661	56	2512	105
LEP	3,186	13	2410	79	3,387	14	2415	80	3,266	9	2394	76	3,926	12	2401	83
IDEA	5,520	18	2423	86	5,647	18	2420	88	5,521	16	2409	91	5,795	17	2411	92
Grade 6																
All Students	39,019	54	2534	101	39,588	55	2538	99	35,794	47	2520	101	36,627	48	2521	102
Female	19,152	59	2546	97	19,412	60	2550	96	17,510	51	2529	99	18,057	51	2529	100
Male	19,866	50	2522	103	20,175	51	2526	101	18,270	44	2511	103	18,558	45	2512	103
Black or African American	5,034	32	2484	92	5,069	34	2493	91	4,530	26	2472	92	4,676	28	2476	91
AmerIndian/Alaskan	119	36	2498	99	80	41	2510	86	97	38	2499	90	99	33	2495	83
Asian	1,931	77	2591	93	2,059	79	2597	88	1,900	72	2582	94	1,872	74	2584	98
Hispanic or Latino	9,938	32	2482	95	10,575	35	2490	95	9,994	27	2471	95	10,838	29	2474	96
Pacific Islander	32	56	2533	91	45	40	2507	97	40	35	2518	101	31	29	2494	83
White	20,706	68	2565	89	20,320	68	2567	89	17,840	61	2552	90	17,548	62	2554	91
Multi-Racial	1,259	58	2542	99	1,440	58	2547	97	1,393	50	2528	101	1,563	53	2532	102
LEP	2,502	6	2406	73	2,710	7	2415	75	2,746	4	2401	70	3,259	5	2406	71
IDEA	5,839	15	2436	89	5,759	15	2442	86	5,564	13	2431	86	5,835	13	2431	84

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table B-3. ELA/L Student Performance Across Four Years (Grades 7 and 8)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																
All Students	39,391	55	2556	104	39,165	56	2559	105	35,995	51	2544	107	37,794	50	2541	109
Female	19,421	61	2572	100	19,200	61	2574	100	17,463	56	2557	102	18,435	54	2553	105
Male	19,970	49	2541	107	19,961	51	2546	107	18,526	46	2532	109	19,321	46	2529	112
Black or African American	4,895	31	2501	97	5,068	33	2507	98	4,491	29	2493	101	4,917	30	2496	102
AmerIndian/Alaskan	95	52	2544	107	117	38	2521	98	106	33	2506	97	98	38	2517	105
Asian	1,942	76	2612	94	1,922	77	2619	97	1,853	75	2607	94	1,926	76	2611	97
Hispanic or Latino	9,757	33	2502	101	10,134	36	2510	101	9,574	32	2494	105	11,184	30	2489	106
Pacific Islander	46	59	2560	122	29	59	2573	103	30	57	2542	135	42	48	2542	103
White	21,546	68	2588	92	20,584	70	2591	93	18,525	63	2575	93	18,115	64	2576	96
Multi-Racial	1,110	57	2564	104	1,311	59	2567	105	1,416	53	2552	109	1,512	54	2552	107
LEP	2,410	5	2421	79	2,429	6	2425	78	2,390	4	2411	80	3,126	5	2408	82
IDEA	5,632	15	2454	92	6,086	17	2460	93	5,479	14	2445	95	5,997	14	2443	96
Grade 8																
All Students	39,427	56	2575	103	39,372	56	2574	104	37,035	51	2562	108	38,522	49	2558	109
Female	19,178	62	2591	99	19,362	62	2591	101	18,100	57	2578	104	18,684	54	2572	106
Male	20,245	50	2560	104	20,006	50	2558	105	18,925	45	2548	109	19,797	45	2544	109
Black or African American	4,932	33	2522	95	4,917	34	2522	96	4,640	31	2511	100	5,007	30	2512	101
AmerIndian/Alaskan	98	38	2546	96	100	50	2563	102	82	37	2537	96	113	32	2523	92
Asian	1,975	76	2629	95	1,917	78	2635	92	1,948	78	2633	96	1,941	75	2626	101
Hispanic or Latino	9,258	34	2522	98	9,883	34	2521	100	9,796	32	2513	103	11,037	31	2510	104
Pacific Islander	37	62	2595	109	48	58	2574	121	37	54	2562	102	31	52	2561	126
White	22,056	69	2605	92	21,345	69	2605	94	19,133	63	2592	97	18,847	62	2590	97
Multi-Racial	1,071	56	2581	102	1,162	57	2581	103	1,399	55	2575	108	1,546	54	2568	111
LEP	2,112	5	2437	72	2,225	3	2432	69	2,091	2	2420	69	2,764	4	2422	74
IDEA	5,557	16	2476	89	5,790	16	2474	88	5,529	14	2466	90	6,063	14	2463	91

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table B-4. Mathematics Student Performance Across Four Years (Grades 3 and 4)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	37,472	54	2440	84	36,460	55	2443	86	34,100	44	2417	92	35,220	47	2426	92
Female	18,393	53	2439	81	17,877	54	2441	83	16,756	42	2413	89	17,252	45	2421	89
Male	19,079	55	2442	87	18,583	56	2445	88	17,342	46	2420	95	17,966	50	2430	95
Black or African American	4,751	30	2395	79	4,597	31	2397	79	4,053	18	2360	82	4,195	23	2374	83
AmerIndian/Alaskan	110	45	2427	77	101	47	2429	76	70	31	2401	85	114	40	2406	101
Asian	2,024	79	2496	77	1,944	79	2497	81	1,943	72	2480	87	1,911	74	2486	91
Hispanic or Latino	10,270	33	2400	78	10,107	35	2404	81	9,610	22	2369	84	10,820	27	2382	84
Pacific Islander	46	50	2441	72	29	59	2432	71	42	29	2400	84	29	45	2424	93
White	18,866	68	2467	74	18,202	69	2470	76	16,784	59	2449	80	16,448	64	2459	80
Multi-Racial	1,405	56	2448	84	1,480	57	2448	87	1,598	50	2431	90	1,703	53	2437	91
LEP	4,158	24	2380	77	4,286	28	2388	79	4,249	18	2357	83	4,710	20	2367	81
IDEA	4,865	19	2361	83	5,028	19	2364	82	4,870	16	2345	89	5,146	18	2351	89
Grade 4																
All Students	38,307	51	2484	85	37,675	52	2486	87	34,527	40	2458	92	35,860	45	2469	94
Female	18,618	50	2482	80	18,467	51	2484	83	16,979	38	2454	87	17,640	43	2465	90
Male	19,689	52	2485	90	19,206	54	2489	91	17,545	43	2462	96	18,216	47	2473	98
Black or African American	4,839	26	2434	79	4,805	28	2437	80	4,166	15	2401	79	4,375	20	2413	83
AmerIndian/Alaskan	104	42	2462	80	104	49	2475	72	79	20	2417	78	82	28	2442	88
Asian	2,007	78	2541	78	2,013	80	2547	78	1,853	71	2526	87	1,975	75	2537	86
Hispanic or Latino	10,178	30	2443	79	10,454	31	2445	81	9,723	19	2411	82	10,636	24	2423	86
Pacific Islander	37	49	2491	88	42	48	2480	69	34	24	2438	86	47	38	2466	83
White	19,747	65	2511	75	18,848	67	2515	76	17,110	55	2490	80	17,039	61	2503	81
Multi-Racial	1,395	53	2491	87	1,409	56	2495	87	1,562	43	2465	95	1,706	51	2482	95
LEP	3,773	19	2418	76	3,992	21	2420	79	3,869	13	2392	79	4,582	17	2405	82
IDEA	5,169	16	2402	82	5,448	17	2404	83	5,191	13	2383	85	5,630	15	2390	88

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table B-5. Mathematics Student Performance Across Four Years (Grades 5 and 6)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	39,540	45	2510	92	38,514	47	2513	94	34,876	35	2485	97	36,200	39	2493	98
Female	19,439	44	2510	89	18,690	45	2512	90	17,129	33	2482	94	17,808	36	2489	94
Male	20,101	46	2510	96	19,823	48	2513	98	17,744	37	2488	101	18,383	41	2497	102
Black or African American	5,031	19	2453	82	4,940	21	2456	85	4,257	11	2423	80	4,537	14	2433	83
AmerIndian/Alaskan	82	29	2488	78	110	28	2477	84	96	23	2462	73	75	19	2456	80
Asian	2,107	74	2577	85	1,997	75	2578	86	1,861	66	2558	95	1,878	68	2566	93
Hispanic or Latino	10,442	24	2466	85	10,344	27	2469	88	9,709	16	2438	86	10,780	20	2446	88
Pacific Islander	49	33	2475	99	36	47	2509	97	33	27	2475	79	35	23	2459	92
White	20,449	59	2539	82	19,644	61	2543	83	17,423	48	2518	87	17,238	54	2529	86
Multi-Racial	1,380	48	2520	90	1,443	49	2519	98	1,497	38	2492	99	1,657	41	2501	99
LEP	3,188	9	2425	77	3,375	13	2434	79	3,200	6	2404	72	3,912	9	2415	76
IDEA	5,511	12	2422	82	5,632	12	2422	85	5,406	9	2404	84	5,781	11	2411	86
Grade 6																
All Students	38,946	44	2527	107	39,488	45	2530	109	35,115	34	2501	111	36,426	37	2506	115
Female	19,115	45	2531	102	19,374	47	2534	104	17,155	33	2500	107	17,966	36	2504	112
Male	19,830	43	2523	112	20,113	44	2527	113	17,948	35	2502	115	18,448	38	2508	119
Black or African American	5,020	19	2464	100	5,051	22	2471	101	4,384	12	2435	101	4,636	14	2440	102
AmerIndian/Alaskan	118	31	2495	107	81	31	2497	94	94	21	2484	101	98	24	2486	88
Asian	1,929	73	2608	100	2,055	78	2616	95	1,889	66	2588	106	1,868	69	2594	112
Hispanic or Latino	9,918	22	2472	101	10,537	24	2476	103	9,640	14	2446	100	10,760	18	2450	107
Pacific Islander	32	47	2532	93	44	34	2511	93	39	28	2493	103	31	19	2481	94
White	20,674	58	2561	92	20,286	59	2564	94	17,696	47	2538	96	17,482	51	2548	98
Multi-Racial	1,255	46	2536	108	1,434	46	2537	108	1,373	35	2509	112	1,551	41	2515	118
LEP	2,495	5	2407	88	2,697	6	2409	92	2,653	3	2388	85	3,244	4	2391	88
IDEA	5,832	9	2415	100	5,725	10	2419	102	5,422	8	2399	102	5,777	8	2401	104

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Table B-6. Mathematics Student Performance Across Four Years (Grades 7 and 8)

Group	2017–2018				2018–2019				2020–2021				2021–2022			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																
All Students	39,265	44	2542	113	39,002	46	2547	115	35,119	37	2526	111	37,528	38	2524	116
Female	19,382	45	2546	110	19,125	46	2550	111	17,004	36	2526	107	18,315	36	2522	112
Male	19,883	44	2539	117	19,873	46	2544	119	18,109	38	2526	114	19,176	39	2526	120
Black or African American	4,873	18	2473	100	5,038	20	2478	104	4,309	14	2462	97	4,858	15	2459	100
AmerIndian/Alaskan	95	38	2521	112	116	31	2513	109	102	20	2488	89	97	22	2497	107
Asian	1,939	73	2628	106	1,917	76	2636	110	1,832	69	2613	110	1,923	71	2621	111
Hispanic or Latino	9,719	21	2481	104	10,072	24	2487	106	9,153	19	2474	100	11,090	18	2466	103
Pacific Islander	46	39	2550	141	29	55	2567	113	28	32	2520	128	42	36	2524	96
White	21,486	58	2578	99	20,525	61	2584	100	18,307	49	2558	99	18,022	53	2566	102
Multi-Racial	1,107	44	2550	113	1,305	50	2560	117	1,388	40	2533	117	1,496	42	2536	117
LEP	2,405	5	2417	91	2,406	5	2419	88	2,277	4	2410	81	3,109	3	2404	81
IDEA	5,607	9	2427	99	6,042	11	2435	101	5,287	8	2423	95	5,911	8	2421	98
Grade 8																
All Students	39,294	43	2558	120	39,216	44	2558	123	35,943	35	2536	121	38,238	34	2532	124
Female	19,100	44	2564	115	19,290	45	2565	118	17,561	36	2541	115	18,527	34	2534	119
Male	20,190	42	2553	125	19,922	42	2552	128	18,373	34	2532	125	19,669	35	2531	128
Black or African American	4,909	18	2483	105	4,890	19	2483	106	4,388	13	2466	102	4,965	13	2463	106
AmerIndian/Alaskan	98	23	2518	107	98	37	2532	120	81	22	2497	104	112	16	2493	104
Asian	1,975	72	2646	114	1,914	74	2653	116	1,916	70	2641	118	1,935	67	2632	123
Hispanic or Latino	9,209	20	2493	108	9,811	21	2493	109	9,289	16	2477	106	10,918	15	2472	107
Pacific Islander	37	57	2589	127	47	38	2575	130	36	25	2507	107	31	45	2545	142
White	21,997	56	2595	107	21,295	57	2597	110	18,867	45	2571	109	18,739	47	2574	111
Multi-Racial	1,069	43	2563	118	1,161	43	2564	127	1,366	37	2544	123	1,538	37	2542	130
LEP	2,101	4	2426	89	2,202	3	2418	83	1,992	2	2403	78	2,743	2	2404	80
IDEA	5,527	8	2438	100	5,712	8	2438	101	5,258	7	2429	96	5,971	6	2426	97

2019–2020 is not included because the summative assessments were canceled due to the COVID-19 pandemic.

Appendix C: Classification Accuracy and Consistency Index by Subgroups

Table C-1. ELA/L Classification Accuracy and Consistency by Subgroup (Grades 3–4)

Group	N	%Accuracy						Proficiency Cut	%Consistency					
		All	L1	L2	L3	L4	All		L1	L2	L3	L4	Proficiency Cut	
Grade 3														
All Students	35,315	80	91	69	65	87	93	72	86	58	54	82	90	
Female	17,296	79	90	69	65	88	93	72	85	58	54	82	90	
Male	18,017	80	91	69	65	87	93	73	87	58	53	81	90	
Black or African American	4,217	81	91	69	65	83	93	74	88	58	53	74	90	
AmerIndian/Alaskan	114	83	93	70	62	88	95	77	90	59	53	80	94	
Asian	1,911	80	88	68	65	90	93	73	81	55	54	87	91	
Hispanic or Latino	10,855	82	92	69	65	85	94	75	89	58	53	75	91	
Pacific Islander	29	80	90	70*	65*	90*	93	72	84	56*	57*	84*	90	
White	16,485	78	88	69	65	88	92	70	80	58	54	83	89	
Multi-Racial	1,704	79	90	70	64	89	92	72	84	58	54	83	89	
LEP	4,710	84	93	68	64	82	95	78	91	58	52	70	93	
IDEA	5,153	85	93	69	64	83	95	79	91	57	51	73	93	
Grade 4														
All Students	35,940	79	90	60	63	88	92	71	86	48	52	81	89	
Female	17,683	78	90	60	63	88	92	71	85	48	52	82	89	
Male	18,253	79	91	60	63	87	93	72	87	47	53	81	90	
Black or African American	4,392	81	92	60	63	85	93	74	89	48	51	76	90	
AmerIndian/Alaskan	82	78	89	61	64	84	93	70	86	48	51	77	90	
Asian	1,975	80	88	60	63	91	93	73	79	48	52	87	90	
Hispanic or Latino	10,663	80	92	60	63	85	93	74	89	48	52	74	90	
Pacific Islander	47	78	91	60*	64*	84	92	70	86	47*	51*	80	89	
White	17,071	77	87	60	63	88	92	69	80	47	53	82	88	
Multi-Racial	1,710	78	90	60	63	88	92	70	84	48	52	83	89	
LEP	4,590	83	93	60	63	81	94	78	91	48	52	66	92	
IDEA	5,642	85	94	60	63	84	95	79	92	46	52	72	93	

*The classification index is based on $n < 10$.

Table C-2. ELA/L Classification Accuracy and Consistency by Subgroup (Grades 5–6)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 5													
All Students	36,300	80	91	64	72	86	93	72	86	52	62	80	90
Female	17,859	79	91	64	72	86	93	72	85	53	62	81	90
Male	18,432	80	91	64	72	86	93	72	87	52	62	79	90
Black or African American	4,561	81	92	64	72	81	93	74	88	53	61	71	90
AmerIndian/Alaskan	75	80	92	67	72	76*	92	71	89	51	65	62*	89
Asian	1,884	82	91	64	72	89	94	74	82	52	61	86	92
Hispanic or Latino	10,822	81	92	64	72	83	93	74	89	53	61	74	90
Pacific Islander	35	81	95	61*	74	87*	91	73	91	49*	67	72*	88
White	17,262	78	88	64	71	87	92	70	80	52	62	81	89
Multi-Racial	1,661	80	90	64	72	87	93	72	85	51	63	81	90
LEP	3,926	85	94	64	71	72	94	79	91	52	59	49	92
IDEA	5,795	85	94	64	71	80	95	80	92	52	60	69	92
Grade 6													
All Students	36,627	79	90	69	72	84	92	70	84	58	63	75	88
Female	18,057	78	89	68	72	85	91	70	82	58	63	76	88
Male	18,558	79	91	69	72	84	92	71	85	58	63	75	89
Black or African American	4,676	80	91	68	72	80	92	72	86	59	62	66	89
AmerIndian/Alaskan	99	77	88	70	72	93*	89	69	82	61	64	62*	85
Asian	1,872	80	90	68	72	88	94	73	81	56	63	83	91
Hispanic or Latino	10,838	81	92	68	72	81	93	73	87	59	62	68	90
Pacific Islander	31	73	80	68	69*	67*	92	64	76	58	55*	55*	89
White	17,548	77	87	69	72	85	91	68	77	58	64	77	87
Multi-Racial	1,563	79	89	69	73	85	92	70	83	58	64	77	89
LEP	3,259	88	94	68	70	72	97	83	92	58	52	49	95
IDEA	5,835	85	93	68	72	79	95	79	91	57	60	62	93

*The classification index is based on $n < 10$.

Table C-3. ELA/L Classification Accuracy and Consistency by Subgroup (Grades 7–8)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 7													
All Students	37,794	79	90	67	75	83	92	71	85	56	67	74	88
Female	18,435	78	89	67	75	84	91	70	83	56	67	75	88
Male	19,321	80	91	67	75	83	92	72	86	56	67	73	89
Black or African American	4,917	80	91	67	74	79	92	72	87	57	65	66	88
AmerIndian/Alaskan	98	79	91	65	77	80	93	71	85	57	68	66	90
Asian	1,926	80	89	68	75	87	93	72	80	55	66	82	90
Hispanic or Latino	11,184	81	92	67	75	81	92	74	88	57	66	67	89
Pacific Islander	42	78	93	65	73	87*	90	69	82	56	66	76*	86
White	18,115	77	87	67	75	83	91	69	79	55	68	74	87
Multi-Racial	1,512	79	90	67	75	85	91	70	83	56	67	75	88
LEP	3,126	89	95	67	71	78*	97	85	93	55	56	48*	95
IDEA	5,997	85	93	67	74	78	94	80	91	56	62	61	92
Grade 8													
All Students	38,522	79	89	69	76	84	92	71	83	58	68	75	89
Female	18,684	79	88	69	76	84	92	70	81	58	68	76	89
Male	19,797	80	90	69	76	83	92	72	85	58	68	73	89
Black or African American	5,007	80	89	69	76	81	93	73	85	58	67	70	90
AmerIndian/Alaskan	113	79	89	69	79	80*	92	71	82	60	69	66*	88
Asian	1,941	81	88	70	76	87	94	73	79	58	68	82	91
Hispanic or Latino	11,037	80	90	68	76	81	92	73	86	58	67	68	89
Pacific Islander	31	80	95*	65*	72	87*	96	73	87*	57*	68	70*	94
White	18,847	78	86	69	76	84	91	69	77	58	69	76	88
Multi-Racial	1,546	80	88	69	76	84	92	72	84	57	68	78	89
LEP	2,764	89	94	68	75	51*	98	85	93	54	57	28*	96
IDEA	6,063	84	92	69	75	78	95	78	89	57	62	65	93

*The classification index is based on $n < 10$.

Table C-4. Mathematics Classification Accuracy and Consistency by Subgroup (Grades 3–4)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 3													
All Students	35,220	83	89	73	78	90	94	76	85	62	70	85	92
Female	17,252	83	89	73	78	89	94	76	85	61	70	84	91
Male	17,966	83	88	73	78	90	94	76	85	62	70	86	92
Black or African American	4,195	83	89	73	78	88	95	77	87	60	69	79	92
AmerIndian/Alaskan	114	85	90	74	77	97	94	80	90	58	73	89	92
Asian	1,911	85	86	73	78	93	95	80	82	61	70	91	93
Hispanic or Latino	10,820	83	90	73	78	86	95	77	87	61	69	79	93
Pacific Islander	29	85	91	76*	75*	90*	94	78	88	63*	62*	90*	91
White	16,448	82	87	73	78	90	93	75	80	63	71	86	91
Multi-Racial	1,703	83	89	72	79	90	94	76	84	62	71	86	91
LEP	4,710	84	90	73	78	87	95	78	88	61	68	79	93
IDEA	5,146	84	88	72	79	87	96	78	88	57	69	80	95
Grade 4													
All Students	35,860	85	91	80	79	90	95	79	87	72	71	85	92
Female	17,640	85	91	80	79	89	94	78	87	72	71	84	92
Male	18,216	85	92	80	79	90	95	79	88	72	70	86	93
Black or African American	4,375	86	92	80	77	87	96	81	89	72	68	79	94
AmerIndian/Alaskan	82	85	91	82	76	87	96	79	86	76	65	83	94
Asian	1,975	87	89	80	79	94	96	82	79	72	71	92	94
Hispanic or Latino	10,636	86	92	80	78	88	95	80	89	72	69	81	93
Pacific Islander	47	84	87	78	83	92*	95	77	84	74	72	81*	93
White	17,039	84	89	80	79	90	94	77	81	72	71	85	91
Multi-Racial	1,706	85	91	81	79	91	95	79	87	73	71	86	92
LEP	4,582	87	93	79	77	85	96	81	90	71	68	78	94
IDEA	5,630	89	94	80	78	87	97	84	92	71	69	80	96

*The classification index is based on $n < 10$.

Table C-5. Mathematics Classification Accuracy and Consistency by Subgroup (Grades 5–6)

Group	N	%Accuracy						%Consistency					
		All	L1	L2	L3	L4	Proficiency Cut	All	L1	L2	L3	L4	Proficiency Cut
Grade 5													
All Students	36,200	84	92	77	71	90	95	77	88	67	61	85	92
Female	17,808	84	91	77	71	90	94	77	88	67	61	84	92
Male	18,383	84	92	77	71	90	95	78	88	67	60	86	93
Black or African American	4,537	87	93	76	71	86	96	81	91	66	59	76	95
AmerIndian/Alaskan	75	86	93	81	70	86*	93	79	91	72	56	73*	91
Asian	1,878	86	89	76	71	94	95	80	82	67	60	91	93
Hispanic or Latino	10,780	86	93	77	71	87	96	80	90	66	60	79	94
Pacific Islander	35	82	92	73	64*	89*	90	75	89	63	52*	87*	87
White	17,238	82	89	77	71	90	93	75	82	68	61	85	91
Multi-Racial	1,657	84	91	76	71	91	94	77	87	67	61	87	92
LEP	3,912	89	94	75	70	86	97	84	92	63	59	73	96
IDEA	5,781	90	95	76	71	86	97	85	93	64	60	77	96
Grade 6													
All Students	36,426	84	93	77	71	89	94	77	89	69	60	84	91
Female	17,966	84	92	77	71	89	94	77	88	69	60	83	91
Male	18,448	84	93	77	71	90	94	78	89	68	60	84	91
Black or African American	4,636	87	94	76	70	85	95	82	92	68	58	75	93
AmerIndian/Alaskan	98	83	90	78	71	84*	94	75	85	69	61	76*	92
Asian	1,868	86	91	76	71	93	95	80	87	67	59	91	93
Hispanic or Latino	10,760	87	94	76	70	87	95	81	91	68	60	77	93
Pacific Islander	31	81	88	77	67*	100*	95	73	84	67	58*	62*	93
White	17,482	81	89	77	71	89	92	74	82	69	61	84	89
Multi-Racial	1,551	85	93	78	70	92	94	78	88	69	61	86	92
LEP	3,244	92	96	75	69	85	97	89	95	65	55	70	95
IDEA	5,777	91	96	76	71	87	96	88	95	67	58	78	94

*The classification index is based on $n < 10$.

Table C-6. Mathematics Classification Accuracy and Consistency by Subgroup (Grades 7–8)

Group	N	%Accuracy					Proficiency Cut	%Consistency					Proficiency Cut
		All	L1	L2	L3	L4		All	L1	L2	L3	L4	
Grade 7													
All Students	37,528	84	92	76	74	90	94	78	88	66	65	85	91
Female	18,315	84	92	76	74	89	94	77	88	67	65	84	91
Male	19,176	85	92	75	74	91	94	78	89	66	65	86	91
Black or African American	4,858	87	94	76	74	87	95	82	91	66	63	78	93
AmerIndian/Alaskan	97	85	91	75	85*	88	96	78	86	70	63*	84	93
Asian	1,923	87	89	76	75	94	95	81	82	66	64	92	93
Hispanic or Latino	11,090	86	93	75	74	87	95	81	91	65	63	78	92
Pacific Islander	42	82	89	74	76*	85*	91	74	86	61	63*	85*	88
White	18,022	82	89	76	74	90	93	75	82	67	65	85	90
Multi-Racial	1,496	84	91	75	76	91	94	78	87	65	67	87	92
LEP	3,109	92	96	74	74	85	97	89	95	62	57	72	95
IDEA	5,911	91	95	75	74	87	96	87	94	63	62	78	94
Grade 8													
All Students	38,238	83	91	72	71	90	94	77	88	61	59	85	92
Female	18,527	83	91	72	71	90	94	76	87	61	60	84	92
Male	19,669	84	92	72	71	90	94	77	88	61	59	86	92
Black or African American	4,965	87	93	72	72	86	96	81	91	59	58	78	95
AmerIndian/Alaskan	112	85	93	74	69	93*	95	78	88	67	55	82*	93
Asian	1,935	85	90	72	71	94	94	79	83	61	61	92	92
Hispanic or Latino	10,918	86	93	72	70	87	96	80	90	60	58	79	94
Pacific Islander	31	84	91	65*	68*	94*	95	78	90	51*	59*	86*	93
White	18,739	81	88	72	71	90	93	73	82	63	60	85	90
Multi-Racial	1,538	84	92	72	71	91	95	78	88	62	59	87	93
LEP	2,743	93	95	71	69	86	99	89	95	49	54	71	98
IDEA	5,971	90	95	71	70	86	98	86	94	56	55	78	97

*The classification index is based on $n < 10$.