

**Connecticut Smarter Balanced  
Summative Assessments  
2016–2017 Technical Report  
Addendum to the Smarter Balanced  
Technical Report**



**Submitted to  
Connecticut State Department of Education  
by American Institutes for Research**

## TABLE OF CONTENTS

1. OVERVIEW.....	1
2. TESTING ADMINISTRATION.....	3
2.1 Testing Windows.....	3
2.2 Test Options and Administrative Roles.....	3
2.2.1 Administrative Roles.....	4
2.2.2 Online Administration.....	6
2.2.3 Paper-and-Pencil Test Administration.....	7
2.2.4 Braille Test Administration.....	7
2.3 Training and Information for Test Coordinators and Administrators.....	8
2.3.1 Online Training.....	8
2.3.2 District Training Workshops.....	11
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 Data Forensics Program.....	25
2.7.1 Data Forensics Report.....	25
2.7.2 Changes in Student Performance.....	25
2.7.3 Item Response Time.....	26
2.7.4 Inconsistent Item Response Pattern (Person Fit).....	27
2.8 Prevention and Recovery of Disruptions in Test Delivery System.....	27

2.8.1 High-Level System Architecture.....	28
2.8.2 Automated Backup and Recovery.....	29
2.8.3 Other Disruption Prevention and Recovery.....	30
3. SUMMARY OF 2016–2017 OPERATIONAL TEST ADMINISTRATION .....	31
3.1 Student Population.....	31
3.2 Summary of Student Performance.....	31
3.3 Test Taking Time .....	38
3.4 Student Ability–Item Difficulty Distribution for the 2016–2017 Operational Item Pool .....	39
4. VALIDITY .....	42
4.1 Evidence on Test Content.....	42
4.2 Evidence on Internal Structure .....	46
5. RELIABILITY .....	49
5.1 Marginal Reliability.....	49
5.2 Standard Error Curves .....	50
5.3 Reliability of Achievement Classification.....	54
5.4 Reliability for Subgroups .....	58
5.5 Reliability for Claim Scores .....	59
6. SCORING .....	61
6.1 Estimating Student Ability Using Maximum Likelihood Estimation .....	61
6.2 Rules for Transforming Theta to Vertical Scale Scores .....	62
6.3 Lowest/Highest Obtainable Scores (LOSS/HOSS).....	63
6.4 Scoring All Correct and All Incorrect Cases .....	63
6.5 Rules for Calculating Strengths and Weaknesses for Reporting Categories (Claim Scores).....	64
6.6 Target Scores.....	64
6.6.1 Target Scores Relative to Student’s Overall Estimated Ability.....	64
6.6.2 Target Scores Relative to Proficiency Standard (Level 3 Cut) .....	65
6.7 Handscoring.....	66

6.7.1 Reader Selection .....	67
6.7.2 Reader Training .....	67
6.7.3 Reader Statistics.....	69
6.7.4 Reader Monitoring and Retraining.....	70
6.7.5 Reader Validity Checks.....	70
6.7.6 Reader Dismissal .....	71
6.7.7 Reader Agreement.....	71
7. REPORTING AND INTERPRETING SCORES .....	73
7.1 Online Reporting System for Students and Educators .....	73
7.1.1 Types of Online Score Reports.....	73
7.1.2 The Online Reporting System.....	75
7.2 Paper Family Score Reports .....	85
7.3 Interpretation of Reported Scores.....	87
7.3.1 Scale Score.....	87
7.3.2 Standard Error of Measurement .....	87
7.3.3 Achievement Level.....	87
7.3.4 Performance Category for Claims.....	88
7.3.5 Performance Category for Targets.....	88
7.3.6 Aggregated Score.....	88
7.4 Appropriate Uses for Scores and Reports.....	89
8. QUALITY CONTROL PROCEDURE.....	90
8.1 Adaptive Test Configuration .....	90
8.1.1 Platform Review.....	90
8.1.2 User Acceptance Testing and Final Review.....	91
8.2 Quality Assurance in Document Processing.....	91
8.3 Quality Assurance in Data Preparation .....	91
8.4 Quality Assurance in Handscoring .....	91

8.4.1 Double Scoring Rates, Agreement Rates, Validity Sets, and Ongoing Read-Behinds ..... 91

8.4.2 Handscoring QA Monitoring Reports..... 92

8.4.3 Monitoring by Connecticut State Department of Education ..... 92

8.4.4 Identifying, Evaluating, and Informing the State on Alert Responses..... 92

8.5 Quality Assurance in Test Scoring ..... 93

    8.5.1 Score Report Quality Check..... 94

REFERENCES ..... 96

APPENDICES ..... 97

## **LIST OF TABLES**

Table 1. 2016–2017 Testing Windows.....	3
Table 2. Summary of Tests and Testing Options in 2016–2017 .....	3
Table 3. SY 2016–2017 Universal Tools, Designated Supports, and Accommodations .....	21
Table 4. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations.....	22
Table 5. ELA/L Total Students with Allowed Embedded Designated Supports.....	22
Table 6. ELA/L Total Students with Allowed Non-Embedded Designated Supports .....	23
Table 7. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations .....	23
Table 8. Mathematics Total Students with Allowed Embedded Designated Supports .....	24
Table 9. Mathematics Total Students with Allowed Non-Embedded Designated Supports .....	24
Table 10. Number of Students in Summative ELA/L Assessment .....	31
Table 11. Number of Students in Summative Mathematics Assessment .....	31
Table 12. ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 3–5).....	32
Table 13. ELA/L Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 6–8).....	33
Table 14. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 3–5).....	34
Table 15. Mathematics Percentage of Students in Achievement Levels for Overall and by Subgroups (Grades 6–8).....	35
Table 16. ELA/L Percentage of Students in Performance Categories for Reporting Categories .....	37
Table 17. Mathematics Percentage of Students in Performance Categories for Reporting Categories.	38
Table 18. ELA/L Test Taking Time .....	39
Table 19. Mathematics Test Taking Time.....	39
Table 20. Percentage of ELA/L Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered.....	43
Table 21. ELA/L Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of- Knowledge and Item Type .....	43
Table 22. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: Grades 3–5 Mathematics .....	44
Table 23. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: Grades 6–8 Mathematics .....	45

Table 24. Average and the Range of the Number of Unique Targets Assessed within Each Claim Across all Delivered Tests .....	46
Table 25. Correlations among Reporting Categories for ELA/L.....	47
Table 26. Correlations among Reporting Categories for Mathematics .....	48
Table 27. Marginal Reliability for ELA/L and Mathematics .....	50
Table 28. Average Conditional Standard Error of Measurement by Achievement Levels .....	53
Table 29. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the SEMs between Two Cuts .....	53
Table 30. Classification Accuracy and Consistency by Achievement Levels.....	57
Table 31. Marginal Reliability Coefficients for Overall and by Subgroup for ELA/L .....	58
Table 32. Marginal Reliability Coefficients for Overall and by Subgroup for Mathematics .....	58
Table 33. Marginal Reliability Coefficients for Claim Scores in ELA/L.....	59
Table 34. Marginal Reliability Coefficients for Claim Scores in Mathematics .....	60
Table 35. Vertical Scaling Constants on the Reporting Metric .....	62
Table 36. Cut Scores in Scale Scores .....	63
Table 37. Lowest and Highest Obtainable Scores .....	63
Table 38. ELA/L Reader Agreements for Short-Answer Items .....	71
Table 39. Mathematics Reader Agreements.....	72
Table 40. Types of Online Score Reports by Level of Aggregation .....	74
Table 41. Types of Subgroups.....	74
Table 42. Overview of Quality Assurance Reports.....	94

## **LIST OF FIGURES**

Figure 1. ELA/L % Proficient Across Years .....	36
Figure 2. Mathematics ELA/L % Proficient Across Years .....	36
Figure 3. Student Ability–Item Difficulty Distribution for ELA/L .....	40
Figure 4. Student Ability–Item Difficulty Distribution for Mathematics .....	41
Figure 5. Conditional Standard Error of Measurement for ELA/L .....	51
Figure 6. Conditional Standard Error of Measurement for Mathematics .....	52

## **LIST OF EXHIBITS**

Exhibit 1. Home Page: District Level .....	75
Exhibit 2. Subject Detail Page for ELA/L by Gender: District Level .....	76
Exhibit 3. Claim Detail Page for ELA/L by LEP Status: District Level .....	77
Exhibit 4. Target Detail Page for ELA/L: School Level .....	78
Exhibit 5. Target Detail Page for ELA/L: Class Level .....	79
Exhibit 6. Target Detail Page for Mathematics: School Level .....	80
Exhibit 7. Target Detail Page for Mathematics: Teacher Level .....	81
Exhibit 8. Student Detail Page for ELA/L .....	83
Exhibit 9. Student Detail Page for Mathematics .....	84
Exhibit 10. Participation Rate Report at District Level .....	85
Exhibit 11. Sample Paper Family Score Report .....	86

## **LIST OF APPENDICES**

Appendix A	Number of Students for Interim Assessments
Appendix B	Percentage of Proficient Students in 2014–2015, 2015–2016, and 2016–2017 for All Students and by Subgroups
Appendix C	Classification Accuracy and Consistency Index by Subgroups



## 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, and 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, an AA-AAAS, are taught to the same academic content 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 SAT to replace the Smarter Balanced grade 11 assessments for high school students. American Institutes for Research (AIR) delivered and scored the Smarter Balanced assessments and produced score reports. Measurement Incorporated (MI) scored the handscored items.

The Smarter Balanced assessments consist 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:** An online adaptive test that provides an individualized assessment for each student.
- **Performance Task:** A task that challenges students to apply their knowledge and skills to respond to real-world problems. Performance tasks can best be described as collections of questions 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 that cannot be adequately assessed with selected-response or constructed-response items. Some performance task items can be scored by the computer, but most are handscored.

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, Writing and Language, and mathematics tests.

- Removed the summative field test items and off-grade items from the ELA/L and mathematics CAT item pool.
- Removed performance tasks (PT) in ELA/L while keeping PTs in mathematics assessment. For the paper tests, the test booklet will include both non-PT and PT components, but only the non-PT component will be scored for ELA/L.
- Combined claim 2 (writing) and 4 (research/inquiry) in ELA/L reporting categories.

Optional interim assessments allow teachers to check student progress throughout the year, giving them information they can use to improve their instruction and learning. These tools are used at the discretion of schools and district, and teachers can employ them to check students' progress at mastering specific concepts at strategic points during the school year. The interim assessments are available as fixed-form tests and consist of the following features:

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

This report provides a technical summary of the 2016–2017 summative assessments in ELA/L and mathematics administered in grades 3–8 under the Connecticut Smarter Balanced assessments. The report includes eight chapters covering an overview, test administration, the 2016–2017 operational administration, validity, reliability, scoring, reporting and interpreting scores, and the quality control process. 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 is 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 Smarter Balanced technical report. The 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 is included in the Smarter Balanced technical report.

Smarter Balanced produces a technical report for the Smarter Balanced assessments, including all aspects of the technical qualities for the Smarter Balanced assessments described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) and the requirements of the U.S. Department of Education peer review of State Assessment Systems Non-Regulatory Guidance for States. The Smarter Balanced technical report includes information using the data at the consortium level, combining data from the consortium states.

## 2. TESTING ADMINISTRATION

### 2.1 TESTING WINDOWS

The 2016–2017 Smarter Balanced assessments testing window spanned approximately two 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 two-month online summative window. Table 1 shows the testing windows for both online and paper-pencil assessments.

Table 1. 2016–2017 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	03/27/2017	05/26/2017	Online Adaptive Test
	3–8	03/27/2017	05/26/2017	Paper Fixed-Forms
Interim Comprehensive Assessments	3–8, 11	10/18/2016	06/09/2017	Online Fixed-Forms
Interim Assessment Blocks	3–8, 11	10/18/2016	06/09/2017	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, a number of assessment options were available for the 2016–2017 administration to accommodate students’ needs. Table 2 lists the testing options that were offered in 2016–2017. A testing option is selected by content area. Once a testing option is selected, it would apply to all tests in the content area.

Table 2. Summary of Tests and Testing Options in 2016–2017

Assessments	Test Options	Test Mode
Summative Assessments	English	Online
	Braille	Online
	Braille Fixed-Form (mathematics only)	Online
	Spanish (mathematics only)	Online
	Paper-Pencil Large-Print Fixed-Form*	Paper
	Paper-Pencil Braille Fixed-Form*	Paper
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 into 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 the day(s) of testing. 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). The main responsibilities of these key personnel are described below. More detailed descriptions can be found in the TAM provided online at this URL: <http://ct.portal.airast.org/resources/>.

#### **District Administrator (DA)**

The DA may add users with District Test Coordinator (DTC) roles in 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 (DTC)**

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

DTCs are responsible for the following:

- Reviewing all Smarter Balanced policies and test administration documents
- Reviewing scheduling and test requirements with STCs, TEs, and TAs
- Working with STCs and Technology Coordinators (TCs) to ensure that all systems, including the secure browser, are properly installed and functional
- Importing users (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 properly administer the Smarter Balanced assessments
- 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 (STC)**

The STC is primarily responsible for coordinating the administration of the Smarter Balanced assessment 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 CSDE.

STCs are responsible for the following:

- Based on test administration 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 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 (TE)**

A TE responsible for administering the Smarter Balanced assessments must have the same qualifications as a TA. They also have the same test administration responsibilities as a TA. 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 (TA)**

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 Administration

Within the state’s testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long test 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 an online TA Certification Course. Staff who complete this course receive a certificate of completion and appear in the online testing system.

To start a test session, the TEs or TAs 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), first name, and 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 feature(s) (see Section 2.6 for a list of accommodations). Students can begin testing only when the TA or TE confirms the setting. The TA or TE then reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the student(s) and walks them through the login process.

Once an assessment is started, the student must answer all test questions presented on a page before proceeding to the next page. Skipping questions 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. No pause rule is implemented for the performance tasks. The same rules that apply to the CAT for reviews and changes to responses also applies to performance tasks.

For the summative test, an assessment can be started in one component and completed in another component. For the CAT, the assessment must be completed within 45 calendar days of the start date, otherwise, the assessment opportunity will expire. For the performance tasks, 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 break for a student or group of students. It is up to the TEs or TAs to determine an appropriate stopping point; however, for ELA/L and mathematics CAT, to ensure the integrity of the test scores and testing, the assessments cannot be paused for more than 20

minutes. If that happens, the student must restart a new test session, which starts from where the student left off. Editing previous responses is no longer an option.

The TAs or TEs must remain in the room at all times 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, and collect and send for secure shredding any handouts or scratch paper that students used during the assessment.

### **2.2.3 Paper-and-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, or students who are visually impaired. For Connecticut, paper-pencil tests were offered only in braille and large print.

The DA must submit a request for accommodated test materials on behalf of the students who need to take the paper-pencil test. If the request is approved, 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 performance task components are combined into one test booklet, including two sessions for CAT and one session for performance tasks in both content areas. The TEs and TAs are asked not to administer the ELA performance task 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.

### **2.2.4 Braille Test Administration**

In SY 2016–2017, the online braille test was also available. The interface is described below in several formats:

- 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 Nemeth braille through the CAT or the performance task via a braille embosser.
- Mathematics items are presented to students in Nemeth braille through a fixed-form CAT test. TAs may decide whether to administer the online fixed-form braille test or the online braille CAT test.
- Students taking the summative ELA/L assessment can emboss both reading passages and items as they progress through the assessment. If a student has a Refreshable Braille Display (RBD), a 40-cell RBD is recommended. The summative ELA/L assessment 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 a 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, while TEs and TAs administer the online assessments. The online 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*

All school personnel who serve as TEs and TAs are encouraged to complete an online TA Certification Course to administer assessments. 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 actually 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.

#### *Webinars*

The following three webinars were offered to the field:

*Technology Requirements for Online Testing:* The webinar provides an overview of the technology requirements needed on all computers and devices used for online testing, information on secure browser installation, and voice packs for text-to-speech accommodations.

*TIDE:* This webinar provides an overview of how to navigate the Test Information Distribution Engine (TIDE).

*AIR Ways Reporting System:* This webinar provides an overview of how to navigate the new AIR Ways Reporting system that provides student performance reports for interim assessments. This system offers more reporting features and greater access to item level data that is not available through the Online Reporting System (ORS). AIR Ways provides access to item level data in addition to individual and group data.

The length of each of these webinars is about one hour. The interactive nature of these training webinars allows the participants to ask questions during and after the presentation. The audio portion of the webinar is recorded. The PowerPoint slides and audio files of the interactive webinars are made available on the portal after the live webinars at <http://ct.portal.airast.org/resources/?section=training-materials>.

#### *Practice and Training Test Site*

In January 2015, separate training sites were opened for TEs/TAs and students, and were refreshed before the 2016–2017 school year. TEs/TAs can practice administering assessments and starting and ending test



sessions on the TA Training Site, and 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 question types and difficulty levels (approximately 30 items each in ELA/L and mathematics), as well as a performance task.

The training tests are designed to provide students and teachers with opportunities to quickly familiarize themselves with the software and navigational tools that 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 are organized by grade bands (grades 3–5, grades 6–8, and grade 11), with each test containing 5–10 questions.

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/TA in the TA Training Site. Items in the student training test include all item types that are included 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 question types.

#### *Manuals and User Guides*

The following manuals and user guides are available on the CT portal, <http://ct.portal.airast.org/>.

The *Test Coordinator Manual* provides information for DCs and STCs regarding policies and procedures for the 2017 Smarter Balanced assessments in ELA/L and mathematics.

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

The *Braille Requirements and Configuration Manual* includes information about supported operating systems and required hardware and software for braille testing. It provides information on how to configure JAWS, navigate an online test with JAWS, and administer a test to a student requiring braille.

The *System Requirements for Online Testing Manual* outlines the basic technology requirements for administering an online assessment, including operating system requirements and supported web browsers.

The *Secure Browser Installation Manual* provides instructions for downloading and installing the secure browser on supported operating systems used for online assessments.

The *Technical Specifications Manual for Online Testing* provides technology staff with the technical specifications for online testing, including information on Internet and network requirements, general hardware and software requirements, and the text-to-speech function.

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

The *Online Reporting System User Guide* provides information about the ORS, including instructions for viewing score reports, accessing test management resources, creating and editing rosters, and searching for students.

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 Assessment Viewing Application (AVA). AVA allows teachers to view items on the Smarter Balanced interim assessments.

The *Teacher Hand Scoring System User Guide* provides information on the Teacher Hand Scoring System (THSS) for scorers and score managers responsible for handscored item responses on the Smarter Balanced interim assessments.

The *AIR Ways User Guide* provides instructions and support for users viewing student interim assessment performance reports in AIR Ways.

All manuals and user guides pertaining to the 2016–2017 online testing were available on the portal, and DAs, DTCs, and STCs can use 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 CT portal, <http://ct.portal.airast.org/>.

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

*How to Access the Data Entry Interface (DEI):* This brochure describes how to access the Data Entry Interface (DEI) to submit the Smarter Balanced Paper Tests. The submission of the LCI is required to confirm student eligibility and register the student for participation in alternate assessments prior to administration.

*How to Activate a Test Session for the Interim Assessments:* This document provides a quick 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.

*Technology Coordinator Brochure:* This brochure provides a quick overview of the basic system and software requirements needed to administer the online tests.

*Test Information Distribution Engine Brochure:* This brochure provides a brief overview of the steps for logging into the Test Information Distribution Engine (TIDE), activating your TIDE account, and managing user accounts in TIDE.

*TIDE Test Settings Brochure:* This brochure provides a brief overview on how to manage student test settings in TIDE. Embedded accommodations and designated supports must be set in TIDE prior to test administration for these settings to be reflected in the TDS.

*User Role Permissions for Online Systems Brochure:* This brochure outlines the user roles and permissions for each secure online testing system, including TIDE, ORS, TDS, THSS, and AVA.

### *Training Modules*

The following training modules were created to help users in the field understand the overall Smarter Balanced assessments, as well as how each system works. All modules were provided in Microsoft PowerPoint (PPT) format; two modules were also narrated.

*Accessibility and Accommodations Training Module:* This course covers the accessibility options, including designated supports and accommodations for students taking the Smarter Balanced assessments. It focuses on students with disabilities, students with a Section 504 Plan, students identified as English Learners, as well as general education students.

*Assessment Viewing Application Module:* This module explains how to navigate AVA. AVA allows authorized users to view the interim comprehensive assessments (ICAs) and IABs for administrative and instructional purposes.

*Embedded Universal Tools and Online Features Module:* The module acquaints students and teachers with the online universal tools (e.g., types of calculators, expandable text) available in the Smarter Balanced assessments.

*Online Reporting System Module:* This module explains how to navigate the ORS, including participation reports and score reports.

*Student Interface for Online Testing Module:* This module explains how to navigate the Student Interface, including how students log in to the testing system, select a test, navigate through the layout of the test, and use the functionality of the test tools.

*Teacher Hand Scoring System Module:* This module provides an overview of THSS. Teachers can use this handscoring system to score items on the interim assessments.

*Technology Requirements for Online Testing Module:* This module provides current information about technology requirements, site readiness, supported devices, and secure browser installation.

*Test Administration Overview Module:* This module gives a general overview of the necessary steps that staff must know in order to prepare for online test administration.

*Test Administrator Interface for Online Testing Module:* This module presents an overview on how to navigate the TA Interface.

*Test Information Distribution Engine Module:* This module provides an overview of the TIDE. It includes information on logging in to TIDE and managing user accounts, student information, rosters, and appeals.

*What Is A CAT? Module:* This module describes a computer adaptive test and how it works when taking ELA/L and mathematics online assessments.

### **2.3.2 District Training Workshops**

District Test Coordinator (DTC) Workshops were held on January 18–20, 2017, at the Institute of Technology and Business Development (ITBD) in New Britain, CT. 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 secured 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 ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. Our systems use role-based security models that ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

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

1. *Test eligibility* refers to the assignment of a test for 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 e-mail message. If information must be sent via e-mail 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 secured 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 DAs, DTCs, STCs, or TEs 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, DAs, DTCs, STCs, TAs, and TEs have defined roles and access to the testing system. When the TIDE testing window opens, 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 DTC’s and STC’s 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 e-mail addresses assigned by the school.

**Secure browser:** A key role of the Technology Coordinator (TC) is to ensure that the secure browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the secure browser prevents students from accessing other computers or Internet applications and from copying test information. The 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 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 interruptions are important factors to be considered 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 without disrupting others, and to instruct students where they are expected to report once they leave. 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 answers provided 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 questions 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 out of sight in the testing room. 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 spacing 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 questions as other students; however, students should be discouraged from communicating through appropriate seating arrangements. For the performance tasks, different forms are spiraled within a classroom so that students receive different forms of the performance tasks.

### **After the Test**

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 at the end of a test session. These materials should be securely shredded or stored in a locked area immediately. The printed reading passages and questions for any content area assessment provided for a student who is 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 are provided in the *Paper and Pencil Test Administration Manual* on how to package and secure the test booklets to be returned to the testing contractor’s office.

## **2.4.4 Test Security Violations**

Everyone who administers or proctors the assessments is responsible for understanding the security procedures for administering the assessments. 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. (Example: Student[s] 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. (Example: 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. (Example: 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

The following students are exempt from participating in the Smarter Balanced assessments:

- A student who has a significant medical emergency
- A student who is classified as Limited English Proficiency (LEP) who has moved to the country within the year (ELA/L exemption only)

## 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 *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 *Guidelines* are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The Smarter Balanced 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 *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 preselected 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 Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* for complete information <http://www.smarterbalanced.org/wp-content/uploads/2015/09/Usability-Accessibility-Accommodations-Guidelines.pdf>.

### **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 preset in TIDE. In the SY 2016–2017 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: <http://ct.portal.airast.org>.

#### **Embedded Universal Tools**

*Zoom in:* Students are able to zoom in and zoom out on test questions, text, or graphics.

*Highlight:* This tool is used to highlight passages or sections of passages and test questions.

*Pause:* 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 questions.

*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 dictionary:* An English dictionary is available for the full write portion of an ELA/L performance task.

*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:* Each passage or stimulus can be expanded so that it takes up a larger portion of the screen.

*Global notes:* Global notes is a notepad that is available for ELA/L performance tasks in which students complete a full write. The student clicks the notepad icon for the notepad to appear. During the ELA/L performance tasks, the notes are retained from segment to segment so that the student may go back to the notes even though he or she may not return to specific items in the previous segment.

*Cross-out response options:* This function allows students to use the strikethrough function.

*Mark a question for review:* Students can mark a question 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 questions.

*Take as much time as needed to complete a Smarter Balanced assessment:* Testing may be split across multiple sessions so that the testing does not interfere with class schedules. The CAT must be completed within 45 calendar days of its starting date. The performance tasks must be completed within 20 calendar days of the starting date.

### **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-based 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.

*English dictionary:* An English dictionary can be provided for the full write portion of an ELA/L performance task. A full write is the second part of a performance task. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

*Scratch paper:* 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.

*Thesaurus:* A thesaurus provides synonyms of terms while a student interacts with text included in the assessment, and is available for the full write. A full write is the second part of a performance task. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

### **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 are able to 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.

*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.

*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 before beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically a 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, Cantonese, Spanish, Korean, Mandarin, Punjabi, Russian, Filipino, 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**

*Bilingual dictionary:* A bilingual/dual language word-to-word dictionary is a language support that can be provided for the full write portion of an ELA/L performance task.

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

*Color overlays:* Color transparencies may be placed over a paper-pencil-based assessment.

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

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

*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*: 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.

*Scribe* (for ELA/L non-writing items): Students dictate their responses to a human who records verbatim what they dictate. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

*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*: This is 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, and illustrations) is presented in a raised format (paper or thermoform). Contracted and non-contracted braille is available, and Nemeth code is available for mathematics.

*Closed captioning for ELA/L listening stim items*: This is printed text that appears on the computer screen as audio materials are presented.

*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 (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 (grade 4 and above mathematics tests)*: A paper-based list of all the digits from 1 to 100 in table format will be available from Smarter Balanced for reference.

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

*Alternate response option:* Alternate response options include but are not limited to and adapted keyboard, large keyboard, StickyKey, MouseKey, FilterKey, adapted mouse, touch screen, head wand, and switches.

*Calculator (for grades 6–8 and grade 11 mathematics tests):* A non-embedded calculator may be provided for students needing a special calculator, such as a braille calculator or a talking calculator that is currently unavailable within the assessment platform.

*Multiplication table (grade 4 and above mathematics tests):* A paper-pencil-based single digit (1–9) multiplication table is available from Smarter Balanced for reference.

*Print-on-demand:* Paper copies of passages, stimuli, and/or items are printed for students. For those students needing 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 (for ELA/L writing items):* Students dictate their responses to a human who records verbatim what they dictate. The scribe must be trained and qualified, and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

*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, and saving work). Voice recognition software generally can recognize speech up to 160 words per minute. Students may use their own assistive technology devices.

Table 3 presents a list of universal tools, designated supports, and accommodations that were offered in the 2016–2017 administration. Tables 4–9 provide the number of students who were offered the accommodations and designated supports.

Table 3. SY 2016–2017 Universal Tools, Designated Supports, and Accommodations

	<b>Universal Tools</b>	<b>Designated Supports</b>	<b>Accommodations</b>
Embedded	Breaks Calculator <sup>1</sup> Digital Notepad English Dictionary <sup>2</sup> English Glossary Expandable Passages Global Notes Highlighter Keyboard Navigation Mark for Review Mathematics Tools <sup>3</sup> Spell Check Strikethrough Writing Tools <sup>4</sup> Zoom	Color Contrast Masking Text-to-Speech <sup>5</sup> Translated Test Directions <sup>6</sup> Translations (Glossary) <sup>7</sup> Translations (Stacked) <sup>8</sup> Turn off Any Universal Tools	American Sign Language <sup>9</sup> Braille Closed Captioning <sup>10</sup> Streamline Text-to-Speech <sup>11</sup>
Non-embedded	Breaks English Dictionary <sup>12</sup> Scratch Paper Thesaurus <sup>13</sup>	Bilingual Dictionary <sup>14</sup> Color Contrast Color Overlay Magnification Read Aloud <sup>15</sup> Noise Buffers Scribe <sup>16</sup> Separate Setting Simplified Test Directions Translated Test Directions Translations (Glossary) <sup>17</sup>	Abacus Alternate Response Options <sup>18</sup> Calculator <sup>19</sup> Multiplication Table <sup>20</sup> Print on Demand Read Aloud <sup>21</sup> Scribe Speech-to-Text 100s Number Table <sup>20</sup>

\*Items shown are available for ELA/L and mathematics unless otherwise noted.

<sup>1</sup> For calculator-allowed items only in grades 6–8 and 11

<sup>2</sup> For ELA/L performance task full-writes

<sup>3</sup> Includes embedded ruler, embedded protractor

<sup>4</sup> Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo

<sup>5</sup> 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.

<sup>6</sup> For mathematics items

<sup>7</sup> For mathematics items

<sup>8</sup> For mathematics test

<sup>9</sup> For ELA/L listening items and mathematics items

<sup>10</sup> For ELA/L listening items

<sup>11</sup> For ELA/L reading passages. Must be set in TIDE by state-level user.

<sup>12</sup> For ELA/L performance task full writes

<sup>13</sup> For ELA/L performance task full writes

<sup>14</sup> For ELA/L performance task full writes

<sup>15</sup> For ELA/L items (not ELA/L reading passages) and mathematics items

<sup>16</sup> For ELA/L non-writing items and mathematics items

<sup>17</sup> For mathematics items on the paper-pencil test

<sup>18</sup> Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches

<sup>19</sup> For calculator-allowed items only in grades 6–8 and 11

<sup>20</sup> For mathematics items beginning in grade 4

<sup>21</sup> For ELA/L reading passages, all grades

Table 4. ELA/L Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
<b>Embedded Accommodations</b>						
American Sign Language	5	10	5	3	7	9
Closed Captioning	28	25	21	27	22	34
Streamlined Mode	101	94	89	82	60	46
Text-to-Speech: Passage and Items	839	760	809	741	749	775
<b>Non-Embedded Accommodations</b>						
Alternate Response Options	6	9	3	9	2	1
Read Aloud Stimuli	43	37	35	21	32	27
Scribe Items (Writing)	9	5	5	3	3	
Speech-to-Text	80	88	91	71	50	27

Table 5. ELA/L Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	18	15	22	24	21	33
	LEP	5	5				1
	IDEA Eligible	6	7	10	17	14	11
Masking	Overall	196	179	185	116	98	107
	LEP	62	52	42	22	27	27
	IDEA Eligible	111	117	124	86	78	84
Text-to-Speech: Items	Overall	4,762	4,631	4,549	3,368	2,734	2,386
	LEP	2,281	1,998	1,805	1,073	907	785
	IDEA Eligible	1,850	2,243	2,497	2,167	1,726	1,445

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

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	4	2	2	4	4	2
	LEP					1	
	IDEA Eligible	3	2	1	1	2	1
Color Overlay	Overall	7	6	5	3	5	2
	LEP					1	
	IDEA Eligible	5	5	4	2	4	1
Magnification	Overall	8	3	5	4	8	13
	LEP	1	2			1	1
	IDEA Eligible	3	2	4	2	4	7
Noise Buffers	Overall	53	36	17	5	4	5
	LEP	6	3	2			1
	IDEA Eligible	17	17	11	1	2	1
Read Aloud Items	Overall	147	141	79	52	35	58
	LEP	64	66	31	16	17	26
	IDEA Eligible	86	87	57	35	30	42
Scribe Items (Non-Writing)	Overall	5	3	4	1	2	
	LEP			1			
	IDEA Eligible	3	2	4	1	2	
Separate Setting	Overall	2,970	3,038	3,300	2,699	2,365	2,417
	LEP	700	710	640	441	403	333
	IDEA Eligible	1,995	2,141	2,440	2,110	1,818	1,850
Simplified Test Directions	Overall	777	434	397	267	232	190
	LEP	364	268	204	136	132	96
	IDEA Eligible	243	223	218	175	146	121
Translated Test Directions	Overall	211	197	210	272	221	202
	LEP	204	186	200	258	204	197
	IDEA Eligible	23	32	20	23	35	29

Table 7. Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade					
	3	4	5	6	7	8
<b>Embedded Accommodations</b>						
American Sign Language	5	8	4	3	7	9
Streamlined Mode	93	88	86	84	58	41
<b>Non-Embedded Accommodations</b>						
100s Number Table	335	216	147	43	51	31
Abacus	3		5	1	4	
Alternate Response Options	6	10	3	7	2	2
Calculator	11	17	26	137	214	230
Multiplication Table		1,688	2,281	2,042	1,694	1,453
Speech-to-Text	74	76	81	64	47	23

Table 8. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	18	15	23	23	22	34
	LEP	5	5				1
	IDEA Eligible	6	7	11	16	15	11
Masking	Overall	237	215	243	161	128	153
	LEP	72	57	59	31	32	35
	IDEA Eligible	149	155	178	132	108	128
Translation (Glossary): Spanish	Overall	583	558	510	585	547	535
	LEP	570	536	495	561	525	517
	IDEA Eligible	42	62	58	79	103	67
Translation (Glossary): Other Languages	Overall	96	69	78	60	52	36
	LEP	94	67	74	57	51	35
	IDEA Eligible	1			2	1	
Text-to-Speech: Stimuli and Items	Overall	6,177	5,878	5,677	4,397	3,619	3,263
	LEP	2,624	2,262	1,962	1,221	989	862
	IDEA Eligible	2,802	3,117	3,394	2,957	2,506	2,228

Table 9. Mathematics Total Students with Allowed Non-Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	5	2	2	4	3	4
	LEP	1					
	IDEA Eligible	3	2	1	1	1	3
Color Overlay	Overall	6	3	5	3	4	2
	LEP					1	
	IDEA Eligible	4	2	4	2	3	1
Translation (Glossary): Spanish	Overall	89	96	113	112	115	119
	LEP	85	88	113	107	103	113
	IDEA Eligible	12	14	9	7	16	10
Translation (Glossary): Other Languages	Overall	25	11	25	15	17	11
	LEP	24	11	25	14	16	8
	IDEA Eligible			1		1	2
Magnification	Overall	8	3	5	5	7	13
	LEP	1	1			1	1
	IDEA Eligible	3	2	4	2	3	7
Noise Buffers	Overall	53	36	16	5	4	5
	LEP	6	3	2			1
	IDEA Eligible	17	17	10	1	2	1
Read Aloud Stimuli and Items	Overall	233	175	101	87	91	90
	LEP	107	82	39	34	62	61
	IDEA Eligible	93	100	61	45	49	39
Read Aloud Stimuli and Items (Spanish)	Overall	97	90	67	56	27	29
	LEP	96	86	67	53	26	27
	IDEA Eligible	8	17	12	8	10	10



Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Scribe Items (Non-Writing)	Overall	9	4	5	2	1	1
	LEP	2	1	1			
	IDEA Eligible	4	2	5	1	1	
Separate Setting	Overall	2,984	3,030	3,308	2,719	2,361	2,404
	LEP	711	720	644	446	394	326
	IDEA Eligible	1,995	2,133	2,438	2,124	1,838	1,842
Simplified Test Directions	Overall	798	457	416	293	242	197
	LEP	362	276	211	146	136	100
	IDEA Eligible	264	242	230	198	156	127
Translated Test Directions	Overall	236	232	217	282	251	239
	LEP	226	216	206	268	234	229
	IDEA Eligible	26	37	18	30	35	26

## 2.7 DATA FORENSICS PROGRAM

### 2.7.1 Data Forensics Report

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, which include clear test administration policies, effective TA training, and tools to identify possible irregularities in test administrations.

Online test administration allows collection of information that was impossible in paper-pencil testing, such as item response changes, item response time, the number of visits for an item or an item group, test starting and ending times, and scores in both the current year and the previous year. AIR’s Test Delivery System (TDS) captures all of this information.

For online administration, a set of quality assurance (QA) reports are generated during and after the testing window. One of the QA reports focuses on flagging possible testing anomalies. Testing anomalies are analyzed for changes in test scores between administrations, testing time, and item response patterns using a person-fit index. Flagging criteria used for these analyses are configurable and can be changed by an authorized user. Analyses are performed at student level and summarized for each aggregate unit, including testing session, TA, and school. The QA reports are provided to state clients to monitor testing anomalies throughout the testing window.

### 2.7.2 Changes in Student Performance

Score changes between years are examined using a regression model. For between-year comparisons, the scores between past and current years are compared, with the current-year score regressed on the test score from the previous year and the number of days between test end days between two years to control the instruction time between the two test scores. Between-year comparisons are performed between the current year (e.g., 2017) and the year before the current year (e.g., 2016).

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. To detect unusual residuals, we compute the

studentized  $t$  residuals. An unusual increase or decrease in student scores between opportunities is flagged when studentized  $t$  residuals are greater than  $|3|$ .

The number of students with a large score gain or loss is 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 studentized  $t$  residuals in an aggregate unit (e.g., testing session, TA, and school). For each aggregate unit, a critical  $t$  value is computed and flagged when  $t$  was greater than  $|3|$ ,

$$t = \frac{\text{Average residuals}}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^n \text{var}(\hat{e}_i)}{n^2}}},$$

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

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

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

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

The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit. If the aggregate unit size is 1–5 students, the aggregate unit is flagged if the percentage of flagged students is greater than 50%. The aggregate unit size for the score change is based on the number of students included in the between-year regression analyses in the aggregate unit.

### **2.7.3 Item Response Time**

The online environment also allows item response time to be captured as the item page time (the length of time that each item page is presented) in milliseconds. Discrete items appear on the screen one item at a time. However, for stimulus-based items selected as part of an item group, all items associated with the stimulus are selected and loaded as a group. 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.

The expectation is that the item response time will be shorter than the average time if students have a prior knowledge of items. An example of unusual item response time is a test record for an individual who scores very well on the test even though the average time spent for each item was far less than that required of students statewide. If students already know the answers to the questions, the response time will be much shorter than the response time for those items where the student has 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 average and standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units are flagged if the test-taking time is greater than  $|3|$  standard deviations of the state average. The state average and standard deviation is computed based on all students

when the analysis was performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

#### 2.7.4 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify examinees 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 test), 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. We note, however, that 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 of a test. An unlikely response to a single test question 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), Sotaridona, Pornell, and Vallejo (2003), aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of  $I_z$  is asymptotically normal (i.e., with an increasing number of administered items,  $i$ ). 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  $I_z$  for systematic flagging of aberrant response patterns. Students with  $I_z$  values greater than  $|3|$  are flagged. Aggregate units are flagged with  $t$  greater than  $|3|$ ,

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

where  $s$  = standard deviation of  $I_z$  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 with the number of flagged students in the aggregate unit (e.g., test session, TA, and school).

## 2.8 PREVENTION AND RECOVERY OF DISRUPTIONS IN TEST DELIVERY SYSTEM

AIR is continuously improving our ability to protect our systems from interruptions. AIR’s test delivery system is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described below, is designed to recover from failure of any component with little interruption. Each system is redundant, and critical student response data is transferred to a different data center each night.

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

AIR has also implemented an escalation procedure that enables us to alert clients within minutes of any disruption. Our emergency alert system notifies our executive and technical staff by text message, who immediately join a call to understand the problem.

The section below describes AIR system architecture and how it recovers from device failures, internet interruptions, and other problems.

### **2.8.1 High-Level System Architecture**

Our architecture provides redundancy, robustness, and reliability required by a large-scale, high stakes testing program. Our general approach, which has been adopted by Smarter Balanced as standard policy, is pragmatic and well supported by our architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. Our system is designed to ensure that the testing results and experience are able to respond robustly to such inevitable failures. Thus, AIR’s test delivery system (TDS) is designed to protect data integrity and prevent student data loss at every point in the process.

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

#### **Student Machine**

Student responses are conveyed to our 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 our 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 for every four satellites serves as a backup hub. 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 failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described below), 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 above. This real-time backup copy remains on the hub until the hub receives 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 quality assurance (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 immediate notification goes out to our 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.

### **2.8.2 Automated Backup and Recovery**

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure 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.

### **2.8.3 Other Disruption Prevention and Recovery**

We have designed our system to be extremely fault-tolerant. The system can withstand failure of any component with little or no interruption of service. One way that we achieve this robustness is through redundancy. Key redundant systems are as follows:

- Our 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.
- Our hosting provider has multiple redundancies in the flow of information to and from our 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, we have redundant firewalls and load balancers throughout the environment.
- We use redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, AIR 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 will need to rerun the backup.

AIR's test delivery system 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 itself is redundant at every component, and the unique design ensures that data is always stored in at least two locations in the event of failure. The engineering that led to this system protects the student responses from loss.

### 3. SUMMARY OF 2016–2017 OPERATIONAL TEST ADMINISTRATION

#### 3.1 STUDENT POPULATION

All students enrolled in grades 3–8 in all public elementary and secondary schools are required to participate in the Smarter Balanced ELA/L and mathematics assessments. Tables 10–11 present the demographic composition of Connecticut students who meet attemptedness requirements for scoring and reporting of the Smarter Balanced summative assessments.

Table 10. Number of Students in Summative ELA/L Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	38,097	39,228	38,748	39,180	39,212	40,139
Female	18,506	19,281	19,028	19,355	19,056	19,440
Male	19,591	19,947	19,720	19,825	20,156	20,699
American Indian/Alaska Native	97	86	104	105	100	108
Asian	2,049	2,109	1,992	1,980	1,982	1,973
African American	4,841	4,939	5,019	4,889	4,933	4,978
Hispanic/Latino	9,847	10,078	9,580	9,438	8,956	9,068
Native Hawaiian/Pacific Islander	33	42	29	44	34	41
White	19,903	20,623	20,830	21,699	22,182	22,921
Multiple Ethnicities	1,327	1,351	1,194	1,025	1,025	1,050
LEP	4,011	3,372	2,779	2,315	2,110	1,857
IDEA	4,490	5,006	5,464	5,415	5,368	5,358

Table 11. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	38,016	39,162	38,656	39,031	39,033	39,955
Female	18,464	19,254	18,990	19,287	18,969	19,350
Male	19,552	19,908	19,666	19,744	20,064	20,605
American Indian/Alaska Native	96	86	101	103	100	109
Asian	2,042	2,106	1,987	1,976	1,983	1,970
African American	4,826	4,927	4,994	4,864	4,906	4,950
Hispanic/Latino	9,817	10,055	9,545	9,397	8,883	9,008
Native Hawaiian/Pacific Islander	33	41	29	44	33	41
White	19,881	20,598	20,805	21,627	22,106	22,831
Multiple Ethnicities	1,321	1,349	1,195	1,020	1,022	1,046
LEP	4,005	3,370	2,770	2,307	2,091	1,845
IDEA	4,484	4,998	5,453	5,391	5,334	5,297

#### 3.2 SUMMARY OF STUDENT PERFORMANCE

Tables 12–15 present a summary of overall student performance in the 2016–2017 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 1–2 compare the percentage of proficient students in 2014–2015, 2015–2016, and 2016–2017 for all students (cohort comparisons). The percentages of proficient students by subgroups across three years are provided in Appendix B. In ELA/L, student performance is compared for 2015–2016 and 2016–2017 only because

ELA/L scores in 2014–2015 were based on both CAT and PT components while ELA scores in 2015–2016 and 2016–2017 were based on CAT component only.

Table 12. ELA/L Percentage of Students in Achievement Levels  
for Overall and by Subgroups (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
<b>Grade 3</b>								
All Students	38,097	2432	91	25	23	23	29	52
Female	18,506	2442	89	21	23	23	32	56
Male	19,591	2423	91	29	23	23	25	48
American Indian/Alaska Native	97	2399	83	38	25	24	13	37
Asian	2,049	2472	84	12	17	26	45	71
African American	4,841	2388	83	43	27	18	12	30
Hispanic/Latino	9,847	2390	85	42	27	18	13	31
Native Hawaiian/Pacific Islander	33	2444	84	18	21	30	30	61
White	19,903	2459	83	14	21	27	39	65
Multiple Ethnicities	1,327	2443	91	22	22	21	34	55
LEP	4,011	2361	76	55	27	12	6	18
IDEA Eligible	4,490	2349	78	63	21	10	6	16
<b>Grade 4</b>								
All Students	39,228	2477	96	27	19	24	31	54
Female	19,281	2487	93	23	19	24	34	58
Male	19,947	2468	97	31	19	23	27	50
American Indian/Alaska Native	86	2465	84	26	28	21	26	47
Asian	2,109	2530	88	12	13	23	53	76
African American	4,939	2428	88	46	22	19	13	32
Hispanic/Latino	10,078	2430	90	45	22	19	14	33
Native Hawaiian/Pacific Islander	42	2457	92	38	19	19	24	43
White	20,623	2506	86	15	17	27	41	67
Multiple Ethnicities	1,351	2489	92	23	19	24	34	58
LEP	3,372	2386	80	64	21	11	4	15
IDEA Eligible	5,006	2389	85	65	18	11	6	17
<b>Grade 5</b>								
All Students	38,748	2512	100	25	18	30	26	56
Female	19,028	2524	97	21	18	32	29	61
Male	19,720	2501	102	29	19	29	23	52
American Indian/Alaska Native	104	2480	97	39	22	22	16	38
Asian	1,992	2564	96	12	13	27	48	75
African American	5,019	2454	91	46	24	23	8	31
Hispanic/Latino	9,580	2461	93	43	23	24	10	34
Native Hawaiian/Pacific Islander	29	2526	74	14	17	45	24	69
White	20,830	2544	89	14	16	35	36	71
Multiple Ethnicities	1,194	2526	96	21	16	34	29	62
LEP	2,779	2400	76	71	20	8	1	9
IDEA Eligible	5,464	2416	86	65	19	12	4	16

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



Table 13. ELA/L Percentage of Students in Achievement Levels  
for Overall and by Subgroups (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
<b>Grade 6</b>								
All Students	39,180	2534	98	22	24	33	21	54
Female	19,355	2547	95	18	23	35	24	59
Male	19,825	2522	99	26	25	31	18	49
American Indian/Alaska Native	105	2521	94	25	29	31	15	47
Asian	1,980	2585	91	9	17	35	39	74
African American	4,889	2483	89	39	30	24	7	31
Hispanic/Latino	9,438	2481	94	40	28	24	7	31
Native Hawaiian/Pacific Islander	44	2523	107	30	25	23	23	45
White	21,699	2564	87	12	21	39	28	67
Multiple Ethnicities	1,025	2547	95	19	24	32	25	57
LEP	2,315	2406	73	75	20	5	1	5
IDEA Eligible	5,415	2438	84	62	24	12	3	14
<b>Grade 7</b>								
All Students	39,212	2556	102	23	22	36	19	55
Female	19,056	2568	99	19	21	38	22	60
Male	20,156	2544	104	27	23	33	17	50
American Indian/Alaska Native	100	2539	96	29	25	31	15	46
Asian	1,982	2607	95	11	15	37	37	74
African American	4,933	2499	96	42	28	24	6	30
Hispanic/Latino	8,956	2501	99	41	27	26	6	32
Native Hawaiian/Pacific Islander	34	2574	111	26	15	24	35	59
White	22,182	2586	90	12	20	42	25	68
Multiple Ethnicities	1,025	2561	99	21	23	37	19	56
LEP	2,110	2421	77	77	18	5	0	5
IDEA Eligible	5,368	2455	91	61	24	12	2	15
<b>Grade 8</b>								
All Students	40,139	2569	103	22	24	36	17	54
Female	19,440	2585	98	17	23	39	21	60
Male	20,699	2554	104	27	25	34	14	48
American Indian/Alaska Native	108	2544	92	25	31	33	10	44
Asian	1,973	2627	94	9	15	40	36	76
African American	4,978	2513	94	40	30	24	5	30
Hispanic/Latino	9,068	2516	99	39	29	26	6	32
Native Hawaiian/Pacific Islander	41	2590	100	15	24	41	20	61
White	22,921	2597	93	13	22	43	23	65
Multiple Ethnicities	1,050	2578	102	20	24	38	19	57
LEP	1,857	2428	71	80	16	3	0	3
IDEA Eligible	5,358	2470	89	60	25	12	2	14

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

Table 14. Mathematics Percentage of Students in Achievement Levels  
for Overall and by Subgroups (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
<b>Grade 3</b>								
All Students	38,016	2432	83	24	23	30	24	53
Female	18,464	2439	79	24	24	30	22	53
Male	19,552	2440	86	24	23	29	25	54
American Indian/Alaska Native	96	2417	67	30	28	31	10	42
Asian	2,042	2490	78	8	16	30	47	76
African American	4,826	2393	77	44	27	20	9	29
Hispanic/Latino	9,817	2401	77	39	28	23	10	33
Native Hawaiian/Pacific Islander	33	2441	77	24	24	30	21	52
White	19,881	2464	74	13	21	35	31	66
Multiple Ethnicities	1,321	2448	83	21	22	30	28	58
LEP	4,005	2385	75	46	29	19	6	24
IDEA Eligible	4,484	2361	81	61	21	13	5	18
<b>Grade 4</b>								
All Students	39,162	2482	85	20	29	27	23	50
Female	19,254	2480	81	20	31	28	21	49
Male	19,908	2483	89	21	28	27	24	51
American Indian/Alaska Native	86	2474	74	19	38	27	16	43
Asian	2,106	2530	78	6	17	27	50	77
African American	4,927	2432	78	40	35	18	7	25
Hispanic/Latino	10,055	2439	79	37	35	20	8	29
Native Hawaiian/Pacific Islander	41	2465	85	20	34	32	15	46
White	20,598	2508	75	10	26	33	30	64
Multiple Ethnicities	1,349	2491	82	16	31	28	25	53
LEP	3,370	2411	73	51	35	11	4	15
IDEA Eligible	4,998	2402	80	56	29	10	4	15
<b>Grade 5</b>								
All Students	38,656	2505	93	30	27	20	23	43
Female	18,990	2504	89	30	29	20	22	42
Male	19,666	2506	96	31	25	19	25	44
American Indian/Alaska Native	101	2480	89	43	26	17	15	32
Asian	1,987	2570	90	11	19	19	51	70
African American	4,994	2445	81	57	27	10	6	16
Hispanic/Latino	9,545	2458	83	50	29	13	8	21
Native Hawaiian/Pacific Islander	29	2506	83	21	31	28	21	48
White	20,805	2535	82	17	27	25	32	57
Multiple Ethnicities	1,195	2515	93	28	26	18	28	46
LEP	2,770	2417	72	71	22	5	2	7
IDEA Eligible	5,453	2418	82	70	20	6	4	10

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

Table 15. Mathematics Percentage of Students in Achievement Levels  
for Overall and by Subgroups (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
<b>Grade 6</b>								
All Students	39,031	2526	106	28	28	22	22	44
Female	19,287	2530	101	26	29	23	22	44
Male	19,744	2523	111	30	27	21	22	43
American Indian/Alaska Native	103	2511	102	34	29	20	17	37
Asian	1,976	2602	99	9	20	21	50	71
African American	4,864	2461	97	52	30	13	5	18
Hispanic/Latino	9,397	2467	100	49	31	14	7	20
Native Hawaiian/Pacific Islander	44	2524	126	36	25	14	25	39
White	21,627	2559	92	15	28	27	29	57
Multiple Ethnicities	1,020	2538	102	25	30	21	24	45
LEP	2,307	2405	88	77	17	4	1	5
IDEA Eligible	5,391	2413	97	72	20	6	2	8
<b>Grade 7</b>								
All Students	39,033	2541	111	30	28	21	21	43
Female	18,969	2542	106	29	29	22	20	42
Male	20,064	2541	115	31	26	21	22	43
American Indian/Alaska Native	100	2508	102	38	35	12	15	27
Asian	1,983	2618	106	12	18	22	48	70
African American	4,906	2469	97	56	28	11	5	16
Hispanic/Latino	8,883	2479	102	51	29	13	7	20
Native Hawaiian/Pacific Islander	33	2569	122	27	24	15	33	48
White	22,106	2575	97	17	28	27	29	56
Multiple Ethnicities	1,022	2540	109	31	29	20	20	40
LEP	2,091	2416	88	80	15	3	2	5
IDEA Eligible	5,334	2430	97	72	19	6	3	9
<b>Grade 8</b>								
All Students	39,955	2554	120	34	24	19	22	42
Female	19,350	2560	114	31	26	21	22	43
Male	20,605	2549	125	37	23	18	22	40
American Indian/Alaska Native	109	2520	98	43	28	18	10	28
Asian	1,970	2645	114	12	16	21	52	72
African American	4,950	2475	103	63	22	10	5	15
Hispanic/Latino	9,008	2489	108	57	24	12	7	19
Native Hawaiian/Pacific Islander	41	2593	116	24	17	24	34	59
White	22,831	2589	107	21	25	24	29	54
Multiple Ethnicities	1,046	2561	123	33	23	19	25	43
LEP	1,845	2418	90	85	11	3	1	4
IDEA Eligible	5,297	2438	101	76	15	5	3	8

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

Figure 1. ELA/L %Proficient Across Years

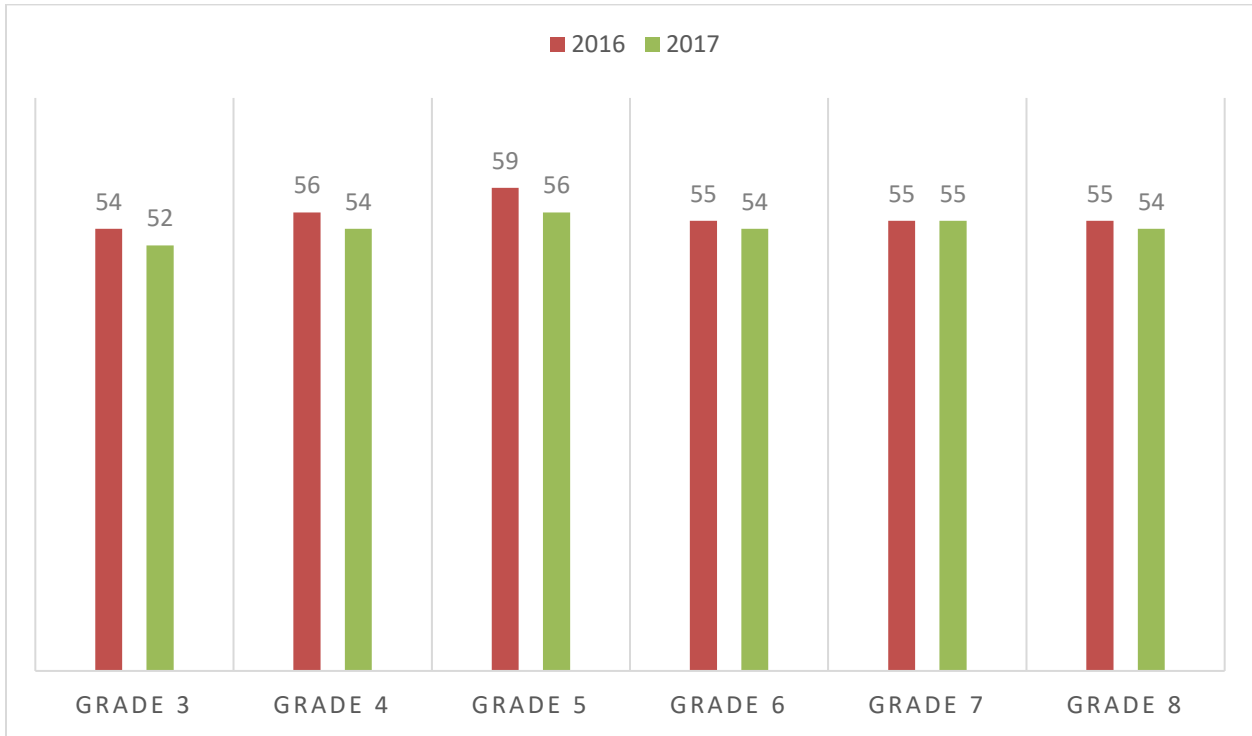
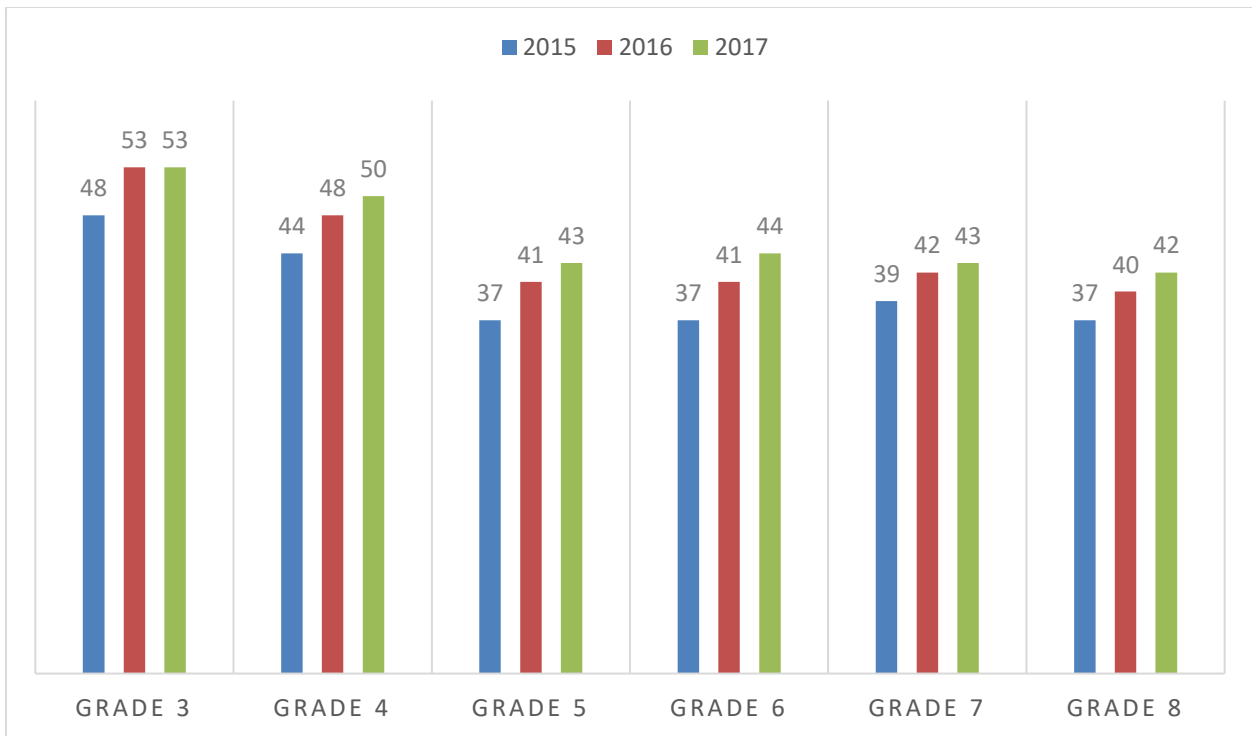


Figure 2. Mathematics %Proficient Across Years



For the reporting categories, because the precision of scores in each reporting category is not sufficient to report scores, given a small number of items, the scores on each reporting category are reported using one of the three performance categories, taking into account the SEM of the reporting category score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 16 and 17 present the distribution of performance categories for each reporting category. The reporting categories are claim 1, claims 2 and 4 combined, and claim 3 in both ELA/L and mathematics.

Table 16. ELA/L Percentage of Students in Performance Categories  
for Reporting Categories

Grade	Performance Category	Claim 1 Reading	Claim 2 & 4: Writing & Research	Claim 3 Listening
3	Below	29	29	16
	At/Near	42	41	62
	Above	29	29	23
4	Below	22	27	19
	At/Near	48	45	57
	Above	31	28	24
5	Below	23	27	17
	At/Near	45	41	59
	Above	32	32	25
6	Below	25	27	15
	At/Near	49	45	64
	Above	27	28	22
7	Below	24	25	18
	At/Near	45	47	64
	Above	31	29	18
8	Below	26	28	14
	At/Near	44	45	65
	Above	30	27	21

Table 17. Mathematics Percentage of Students in Performance Categories  
for Reporting Categories

Grade	Performance Category	Claim 1	Claim 2 & 4	Claim 3
3	Below	30	25	20
	At/Near	33	45	48
	Above	37	30	31
4	Below	33	28	27
	At/Near	33	45	44
	Above	34	27	30
5	Below	40	32	32
	At/Near	32	43	46
	Above	28	25	23
6	Below	37	33	31
	At/Near	35	44	45
	Above	28	23	24
7	Below	40	30	24
	At/Near	32	45	53
	Above	29	24	23
8	Below	40	27	27
	At/Near	33	47	50
	Above	27	26	23

Legend:

Claim 1: Concepts and Procedures;

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

Claim 3: Communicating Reasoning

### 3.3 TEST TAKING TIME

The Smarter Balanced summative assessments are not timed, and an individual student may need more or less time overall. The length of a test session is determined by or 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 questions.

In the Test Delivery System (TDS), item response time is captured as the item page time (the length of time that each item page is presented) in milliseconds. Discrete items appear on the screen one at a time. For items associated with a stimulus, the page time is the time spent on all items associated with the stimulus because all items associated with the stimulus appear on the screen together. For each student, the total time taken to finish the test is computed by adding up the page time for all items. For the items associated with a stimulus, the page time for each item is computed by dividing the page time by the number of items associated with the stimulus.

Tables 18 and 19 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 18. ELA/L Test Taking Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75 <sup>th</sup>	80 <sup>th</sup>	85 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
<b>Overall Test (CAT Component)</b>							
3	1:36	0:43	1:55	2:02	2:12	2:25	2:48
4	1:40	0:43	1:58	2:05	2:14	2:27	2:49
5	1:34	0:37	1:52	1:58	2:06	2:18	2:38
6	1:35	0:39	1:53	2:00	2:08	2:21	2:44
7	1:27	0:35	1:44	1:50	1:58	2:09	2:31
8	1:18	0:31	1:33	1:39	1:46	1:55	2:13

Table 19. Mathematics Test Taking Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75 <sup>th</sup>	80 <sup>th</sup>	85 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
<b>Overall Test</b>							
3	1:59	0:56	2:26	2:38	2:52	3:10	3:42
4	2:01	0:58	2:27	2:38	2:53	3:11	3:46
5	2:20	1:05	2:51	3:03	3:19	3:41	4:18
6	2:11	0:57	2:37	2:47	3:00	3:20	3:56
7	1:46	0:48	2:08	2:17	2:28	2:44	3:12
8	1:51	0:49	2:15	2:24	2:35	2:51	3:19
<b>CAT Component</b>							
3	1:18	0:38	1:35	1:43	1:53	2:06	2:28
4	1:24	0:43	1:42	1:50	2:00	2:15	2:41
5	1:23	0:38	1:41	1:48	1:57	2:10	2:31
6	1:25	0:36	1:41	1:48	1:56	2:08	2:30
7	1:17	0:34	1:33	1:39	1:48	1:59	2:20
8	1:18	0:35	1:35	1:41	1:49	2:00	2:20
<b>PT Component</b>							
3	0:41	0:25	0:52	0:57	1:03	1:12	1:26
4	0:37	0:22	0:47	0:52	0:57	1:04	1:16
5	0:57	0:35	1:12	1:18	1:27	1:40	2:02
6	0:46	0:29	0:57	1:02	1:09	1:19	1:36
7	0:29	0:19	0:36	0:40	0:45	0:52	1:03
8	0:33	0:20	0:42	0:46	0:51	0:58	1:09

### 3.4 STUDENT ABILITY–ITEM DIFFICULTY DISTRIBUTION FOR THE 2016–2017 OPERATIONAL ITEM POOL

Figures 3 and 4 display the empirical distribution of the Connecticut student scale scores in the 2016–2017 administration and the distribution of the summative item difficulty parameters in the operational pool. The student ability distribution is shifted to the left in all grades and subjects, 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 measure high performing students accurately but needs additional easy items to better measure low performing students. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool, and 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 3. Student Ability–Item Difficulty Distribution for ELA/L

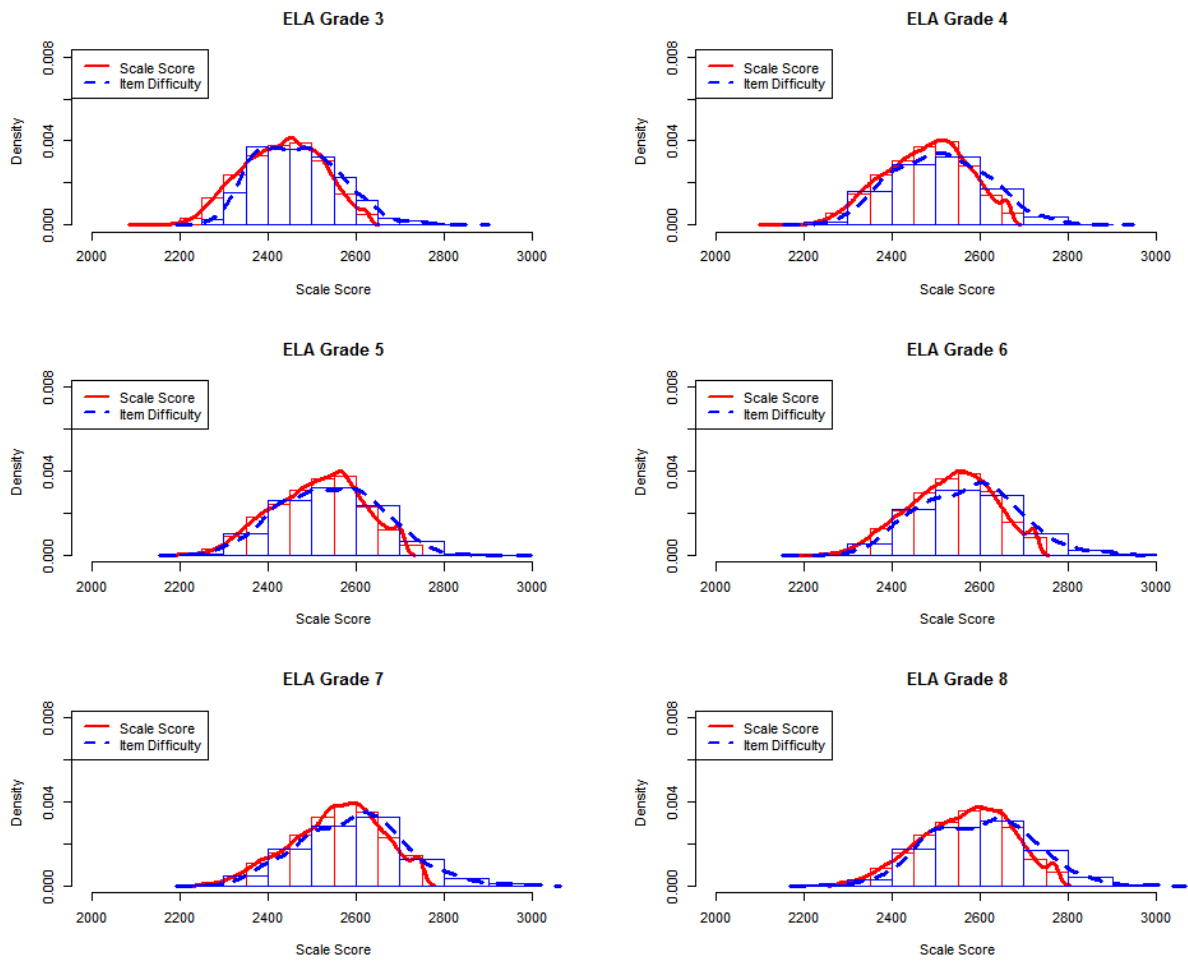
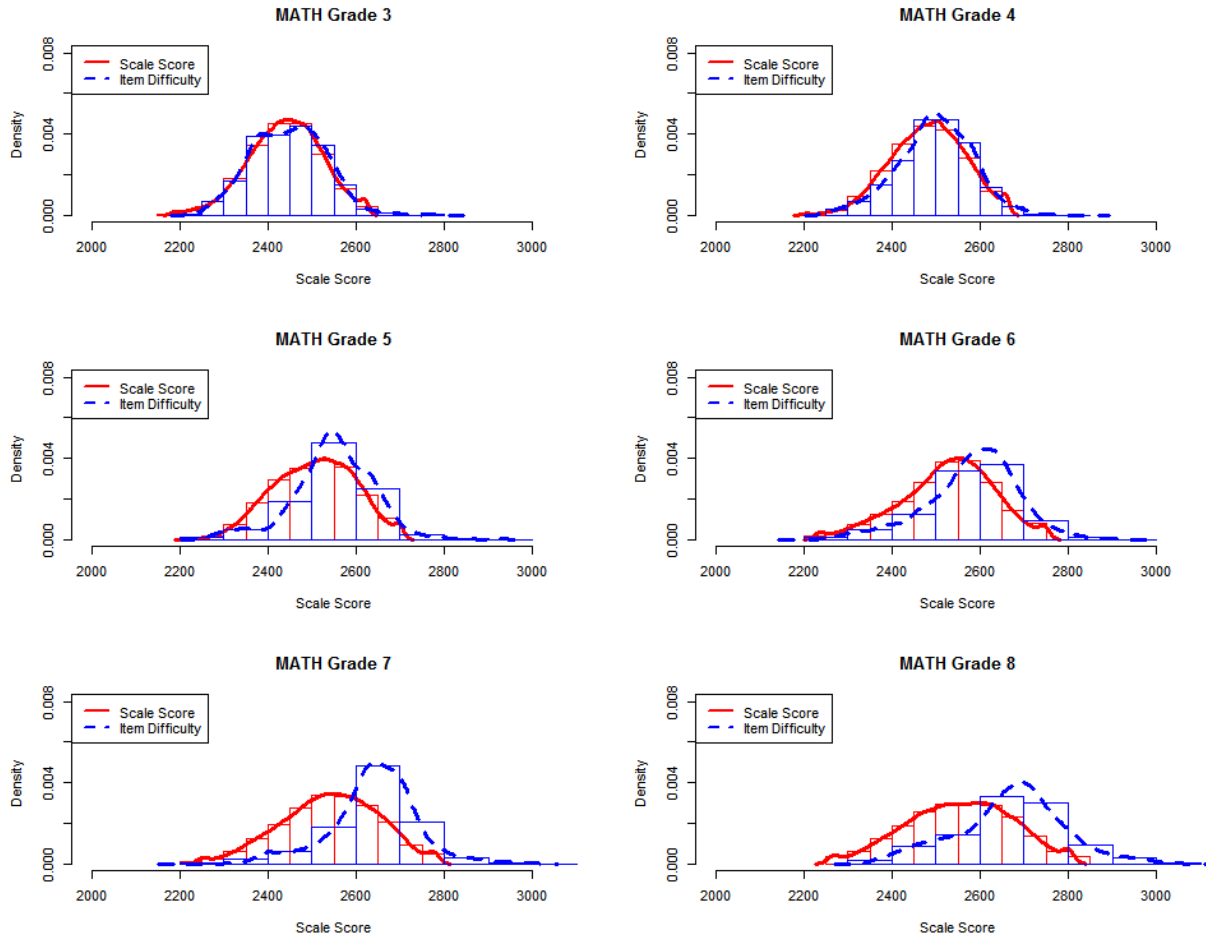




Figure 4. Student Ability–Item Difficulty Distribution for Mathematics



## 4. VALIDITY

According to the *Standards for Educational and Psychological Testing* (AERA, APA, and 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 reporting category 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 include two components: computer adaptive test (CAT) and performance task (PT). For CAT, each student receives a different set of items, adapting to his/her ability. For 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 (Smarter Balanced Assessment Consortium, 2015) specify a range of items to be administered in each claim, content domain/standards, and/or targets. Moreover, blueprints constrain the DoK and item and passage types. 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 ELA/L, the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 20–21 present the percentages of tests aligned with the test blueprint constraints for ELA/L CAT. Table 20 provides the blueprint match rates for item and passage requirements for each claim. For DoK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 21 presents the percentages of tests that satisfied the DoK and item type constraints for each claim. All tests met the requirements, except for the claim 2 DoK2 requirement in grades 3, and 6, which each administered one DoK2 item fewer than required in claim 2.

Tables 22–23 provide the percentages of tests aligned with the test blueprint constraints for mathematics CAT, the blueprint match rates for claims, DoK, and target constraints. In mathematics, all tests met the blueprint requirements except for grades 3, 6, and 8. In grade 3, the violation was in claim 1 for target sets of E, J, and K, which administered one item fewer than required. In grade 6, the violation was in claim 1 no-calculator segment for target sets of E and F and target B, which administered a few items fewer or more

than required. Another violation was in claim 3 calculator segment for target sets of A and D, which administered one or two items fewer than the item requirement. In grade 8, the violation was in claim 1 non-calculator segment for target B and C, and DoK2 or higher, each administered one item fewer or one item more than required.

Table 20. Percentage of ELA/L Delivered Tests Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

Grade	Claim	Min	Max	%BP Match for Item	%BP Match for
3	1-IT	7	8	100%	100%
	1-LT	7	8	100%	100%
	2-W	10	10	100%	
	3-L	8	8	100%	100%
	4-CR	6	6	100%	
4	1-IT	7	8	100%	100%
	1-LT	7	8	100%	100%
	2-W	10	10	100%	
	3-L	8	8	100%	100%
	4-CR	6	6	100%	
5	1-IT	7	8	100%	100%
	1-LT	7	8	100%	100%
	2-W	10	10	100%	
	3-L	8	9	100%	100%
	4-CR	6	6	100%	
6	1-IT	10	12	100%	100%
	1-LT	4	4	100%	100%
	2-W	10	10	100%	
	3-L	8	9	100%	100%
	4-CR	6	6	100%	
7	1-IT	10	12	100%	100%
	1-LT	4	4	100%	100%
	2-W	10	10	100%	
	3-L	8	9	100%	100%
	4-CR	6	6	100%	
8	1-IT	12	12	100%	100%
	1-LT	4	4	100%	100%
	2-W	10	10	100%	
	3-L	8	9	100%	100%
	4-CR	6	6	100%	

Legend: 1-IT: Reading with Information Text; 1-LT: Reading with Literary Text; 2-W: Writing; 3L: Listening; 4-CR: Research

Table 21. ELA/L Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Item Type

DoK and Item Type Constraints	Minimum Required	%Blueprint Match					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Claim 1 DoK2	7	100%	100%	100%	100%	100%	100%
Claim 1 DoK3 or higher	2	100%	100%	100%	100%	100%	100%
Claim 2 DoK2	4	91%	100%	100%	70%	100%	100%
Claim 2 DoK3 or higher	1	100%	100%	100%	100%	100%	100%
Claim 2 Brief Write	1	100%	100%	100%	100%	100%	100%
Claim 3 DoK2 or higher	3	100%	100%	100%	100%	100%	100%

Table 22. Percentage of Delivered Tests Meeting Blueprint Requirements  
for Each Claim and Target: Grades 3–5 Mathematics

Claim	Target	Grade 3		Grade 4		Grade 5	
		Required Items	%BP Match	Required Items	%BP Match	Required Items	%BP Match
Total Adaptive Test Length		34	100%	34	100%	34	100%
1	Overall	20	100%	20	100%	20	100%
	<i>Priority Cluster</i>	15	100%				
	Targets B, C, G, I	6	100%				
	Targets D, F	6	100%				
	Target A	3	100%				
	<i>Supporting Cluster</i>	5	100%				
	Targets E, J, K	4	99%				
	Target H	1	100%				
	<i>Priority Cluster</i>			15	100%		
	Target A, E, F			9	100%		
	Target G			3	100%		
	Target D			2	100%		
	Target H			1	100%		
	<i>Supporting Cluster</i>			5	100%		
	Target I, K			3	100%		
Target B, C, J			1	100%			
Target L			1	100%			
<i>Priority Cluster</i>					15	100%	
Target E, I					6	100%	
Target F					5	100%	
Target C, D					4	100%	
<i>Supporting Cluster</i>					5	100%	
Target J, K					3	100%	
Target A, B, G, H					2	100%	
	DOK 2 or higher	7	100%	7	100%	7	
2	Overall	3	100%	3	100%	3	100%
	Target A	2	100%	2	100%	2	100%
	Targets B, C, D	1	100%	1	100%	1	100%
3	Overall	8	100%	8	100%	8	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%
	DOK 3 or higher	2	100%	2	100%	2	100%
4	Overall	3	100%	3	100%	3	100%
	Targets A, D	1	100%	1	100%	1	100%
	Targets B, E	1	100%	1	100%	1	100%
	Targets C, F	1	100%	1	100%	1	100%
2&4	DOK 3 or higher	2	100%	2	100%	2	100%

Table 23. Percentage of Delivered Tests Meeting Blueprint Requirements  
for Each Claim and Target: Grades 6–8 Mathematics

Claim	Target	Grade 6		Grade 7		Grade 8	
		Required	%BP	Required	%BP	Required	%BP
Total Adaptive Test Length		33	100%	34	100%	34	
1-Calc	Overall	6	100%	10	100%	14	100%
	<i>Priority Cluster</i>	3	100%	6	100%	11	100%
	Target A	2	100%				
	Target G	1	100%				
	Targets A, D			6	100%		
	Target D					4	100%
	Targets E, G					4	100%
	Targets F, H					3	100%
	<i>Supporting Cluster</i>	3	100%	4	100%	3	100%
	Targets H, I, J	3	100%				
	Targets E, F			2	100%		
	Targets G, H, I			2	100%		
	Targets I, J					3	100%
	DOK 2 or higher	2	100%	4	100%	5	100%
1-No Calc	Overall	13	100%	10	100%	6	100%
	<i>Priority Cluster</i>	11	100%	9	100%	4	100%
	Targets E, F	6	99%				
	Target A	2	100%				
	Target B	1	99%			2	89%
	Target D	2	100%	3	100%		
	Target B, C			6	100%		
	Target C					2	89%
	<i>Supporting Cluster</i>	2	100%	1	100%	2	100%
	Target C	2	100%				
Target E			1	100%			
Target A					2	100%	
	DOK 2 or higher	5	100%	4	100%	4	95%
2	Overall	3	100%	3	100%	3	100%
	Target A	2	100%	2	100%	2	100%
	Targets B, C, D	1	100%	1	100%	1	100%
3-Calc	Overall	7	100%	8	100%	8	100%
	Targets A, D	3	99%	2	100%	2	100%
	Targets B, E	2	100%	3	100%	3	100%
	Targets C, F, G	2	100%	1	100%	1	100%
	DOK 3 or higher	1	100%	2	100%	2	100%
3-No Calc	Overall	1	100%		100%		100%
4	Overall	3	100%	3	100%	3	100%
	Targets A, D	1	100%	1	100%	1	100%
	Targets B, E	1	100%	1	100%	1	100%
	Targets C, F	1	100%	1	100%	1	100%
2&4	DOK 3 or higher	2	100%	2	100%	2	100%

Table 24 summarizes the target coverage, the number of unique targets administered in each delivered test by claim. The table includes the number of targets specified in the blueprints and the mean and range of the number of targets administered to students. Since 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 24. Average and the Range of the Number of Unique Targets Assessed within Each Claim Across all Delivered Tests

Grade	Total Targets in BP				Mean				Range (Minimum - Maximum)			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
<b>ELA/L</b>												
3	14	5	1	3	10.2	4.0	1.0	3.0	8-13	3-5	1-1	3-3
4	14	5	1	3	10.3	4.1	1.0	3.0	8-13	3-5	1-1	3-3
5	14	5	1	3	10.1	4.7	1.0	3.0	7-13	3-5	1-1	3-3
6	14	5	1	3	9.3	4.1	1.0	3.0	8-11	3-5	1-1	3-3
7	14	5	1	3	9.4	4.9	1.0	3.0	7-11	3-5	1-1	3-3
8	14	5	1	3	9.4	4.0	1.0	3.0	8-11	3-4	1-1	3-3
<b>Mathematics</b>												
3	11	4	6	6	10.8	2.0	5.5	3.0	9-11	2-2	3-6	2-3
4	12	4	6	6	10.0	2.0	5.5	3.0	9-10	2-2	3-6	3-3
5	11	4	6	6	9.0	2.0	5.3	3.0	9-9	2-2	3-6	3-4
6	10	4	7	6	10.0	2.0	4.8	3.0	8-10	1-2	3-7	2-3
7	9	3	7	6	8.0	2.0	4.8	3.0	8-8	2-2	3-6	3-4
8	10	4	7	6	10.0	2.0	5.2	3.0	10-10	2-2	3-6	3-4

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). However, scores from the test should be comparable, and each test form should measure the same content, albeit with a different set of test items, ensuring the comparability of assessments in content and scores. The blueprint match and target coverage results demonstrate that test forms conform to the same content as specified, thus providing evidence of content comparability. In other words, while each form is unique with respect to its items, all forms align with the same curricular expectations set forth in the test blueprints.

## 4.2 EVIDENCE ON INTERNAL STRUCTURE

The measurement and reporting model used in the Smarter Balanced summative assessments assumes a single underlying latent trait, with achievement reported as a total score as well as scores for each reporting category measured. The evidence on the internal structure is examined based on the correlations among reporting category scores.

The correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 25 and 26. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates. The observed correlation between two reporting category 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 reporting scores are higher than observed correlations. The disattenuated correlations are quite high. The correction for attenuation is large because the marginal reliabilities of claim 3 scores in ELA/L and the marginal reliabilities of claim 2 & 4 and claim 3 scores in mathematics are low. The low reliabilities are due to the low performance with large standard errors, due to a shortage of easy items in the item pool.

Because the reliabilities for reporting category scores are low, the performance of each reporting category scores is reported in three performance categories. The distribution of performance categories for each reporting category is provided in Tables 16–17, Section 3.2. Scale scores are not reported for reporting categories.

Table 25. Correlations among Reporting Categories for ELA/L

Grade	Reporting Categories	Observed and Disattenuated Correlation		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1: Reading		0.96	0.96
	Claim 2 & 4: Writing & Research	0.78		0.95
	Claim 3: Listening	0.65	0.66	
4	Claim 1: Reading		0.97	0.93
	Claim 2 & 4: Writing & Research	0.75		0.94
	Claim 3: Listening	0.64	0.67	
5	Claim 1: Reading		0.99	0.99
	Claim 2 & 4: Writing & Research	0.78		0.98
	Claim 3: Listening	0.68	0.71	
6	Claim 1: Reading		0.98	1
	Claim 2 & 4: Writing & Research	0.76		1
	Claim 3: Listening	0.66	0.69	
7	Claim 1: Reading		0.99	1
	Claim 2 & 4: Writing & Research	0.78		1
	Claim 3: Listening	0.66	0.66	
8	Claim 1: Reading		0.99	1
	Claim 2 & 4: Writing & Research	0.79		1
	Claim 3: Listening	0.64	0.65	

Table 26. Correlations among Reporting Categories for Mathematics

Grade	Reporting Categories	Observed and Disattenuated Correlation		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1		1.00	0.97
	Claim 2 & 4	0.79		1
	Claim 3	0.79	0.74	
4	Claim 1		0.99	0.99
	Claim 2 & 4	0.81		1
	Claim 3	0.81	0.76	
5	Claim 1		1.00	0.98
	Claim 2 & 4	0.78		1
	Claim 3	0.78	0.73	
6	Claim 1		1	0.99
	Claim 2 & 4	0.83		1
	Claim 3	0.80	0.77	
7	Claim 1		1	1
	Claim 2 & 4	0.80		1
	Claim 3	0.78	0.73	
8	Claim 1		1	1
	Claim 2 & 4	0.77		1
	Claim 3	0.78	0.70	

Legend:

Claim 1: Concepts and Procedures; Claims 2 & 4: Problem Solving & Modeling and Data Analysis; Claim 3: Communicating Reasoning



## 5. RELIABILITY

Reliability refers to the consistency of test scores. Reliability is evaluated in terms of the standard errors 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. Within the IRT framework, measurement error varies conditioning on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is the inverse of the measurement error of the test; the larger the measurement error, the less test information is being provided. In computer adaptive testing, because selected items vary across students, the measurement error can vary for the same ability depending on the selected items for each student.

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

### 5.1 MARGINAL RELIABILITY

The marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional SEM, 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 conditional SEM of the scale score for student  $i$ ; and  $\sigma^2$  is the variance of the scale score. The higher 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 CAT, items administered vary across all students, so the SEM also can vary across students, which yield conditional SEM. The average conditional SEM can be computed as

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

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

Table 27 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores.

Table 27. 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
		Min	Max				
<b>ELA/L</b>							
3	38,097	38	40	0.92	2432	91	26
4	39,228	38	40	0.91	2477	96	29
5	38,748	38	41	0.91	2512	100	29
6	39,180	38	41	0.90	2534	98	31
7	39,212	38	41	0.91	2556	102	32
8	40,139	40	41	0.91	2569	103	32
<b>Mathematics</b>							
3	38,016	39	40	0.94	2439	83	19
4	39,162	37	40	0.95	2482	85	20
5	38,656	38	40	0.94	2505	93	23
6	39,031	38	39	0.94	2526	106	26
7	39,033	38	40	0.93	2541	111	29
8	39,955	38	40	0.93	2554	120	32

## 5.2 STANDARD ERROR CURVES

Figures 5 and 6 present plots of the conditional SEM of scale scores across the range of ability. The vertical lines indicate the cut scores for Level 2, Level 3, and Level 4. The item selection algorithm matched items to each student’s ability and to the test blueprints with the same precision across the range of abilities.

Overall, the standard error curves suggest that students are measured with a high degree of precision given that the standard errors are consistently low. However, 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 5. Conditional Standard Error of Measurement for ELA/L

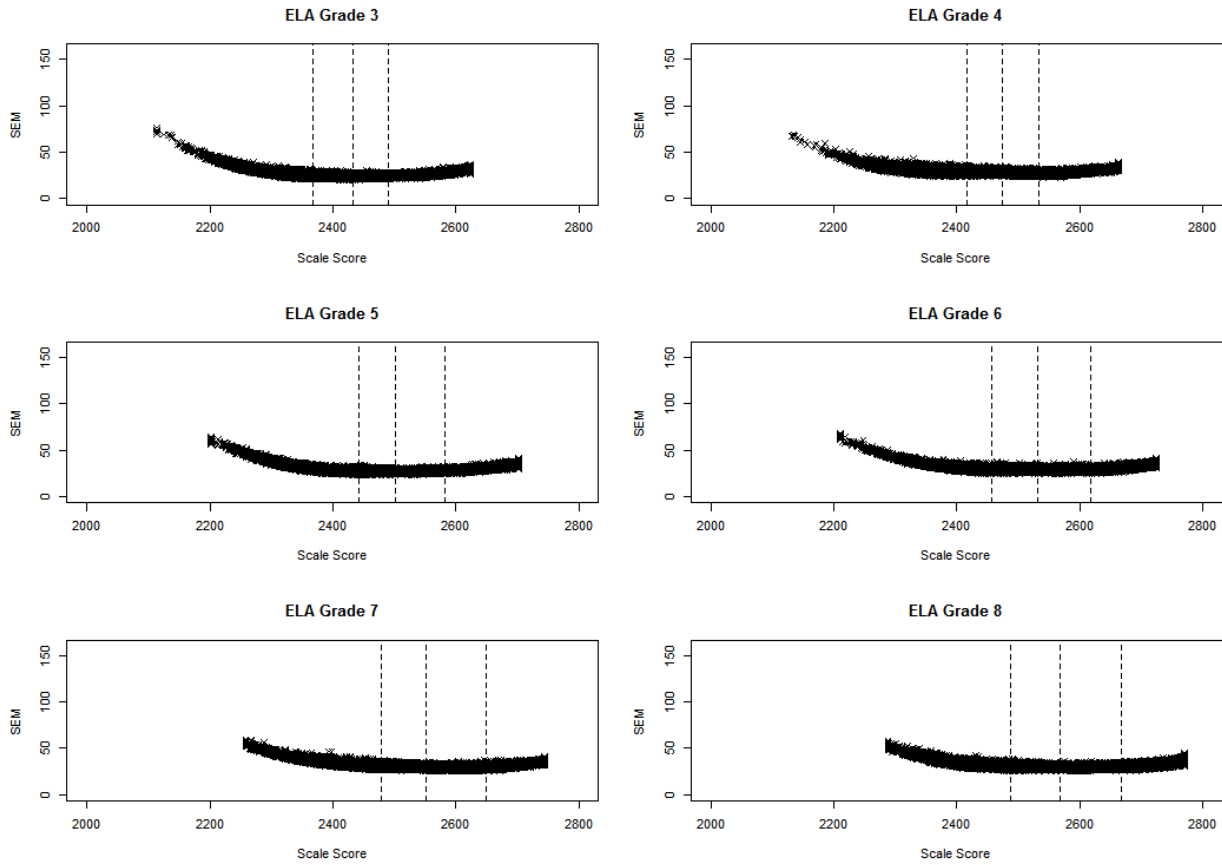
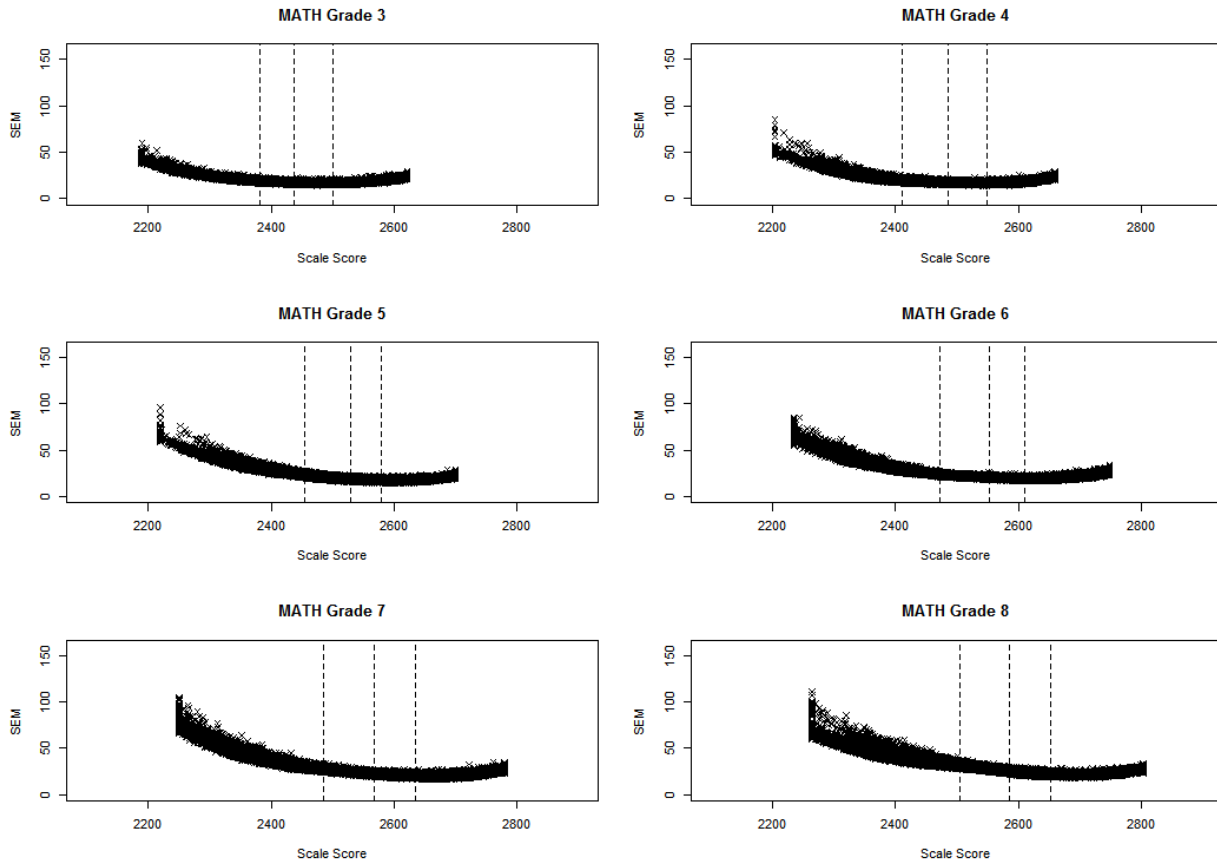


Figure 6. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures above are summarized in Tables 28 and 29. Table 28 provides the average conditional SEM for all scores and scores in each achievement level. Table 29 presents the average conditional SEMs at the each cut score and the difference in average conditional SEMs between two cut scores. As shown in Figures 5 and 6, the greatest average conditional SEM is in Level 1 in both ELA/L and mathematics. Average conditional SEMs at all cut scores are similar in ELA/L, but larger in Level 2 cut scores in mathematics.

Table 28. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
<b>ELA/L</b>					
3	30	24	24	26	26
4	32	29	28	29	29
5	32	27	27	31	29
6	33	29	29	31	31
7	36	30	29	32	32
8	35	30	30	33	32
<b>Mathematics</b>					
3	24	18	17	18	19
4	25	18	17	19	20
5	30	21	18	18	23
6	35	22	20	21	26
7	40	25	21	21	29
8	42	29	24	23	32

Table 29. Average Conditional Standard Error of Measurement at Each Achievement Level Cut and Difference of the SEMs between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	L2-L3	L3-L4	L2-L4
<b>ELA/L</b>						
3	25	24	24	1	0	1
4	29	28	27	1	1	2
5	27	27	28	0	1	1
6	29	30	29	1	1	0
7	31	30	29	1	1	2
8	31	29	30	2	1	1
<b>Mathematics</b>						
3	20	18	17	2	1	3
4	19	17	17	2	0	2
5	23	19	18	4	1	5
6	24	21	19	3	2	5
7	28	23	20	5	3	8
8	32	26	22	6	4	10

### 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, and NCME, 2014). The indexes 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 CAT, because the adaptive testing algorithm constructs a test form unique to each student, the classification indexes 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 actually taken and the classifications that would be made on the basis of 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 and 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 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  $\theta_i$  is the unknown true ability of the  $i$ th student and  $\Phi$  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  $\hat{\theta}_i \sim N(\theta_i, se^2(\hat{\theta}_i))$ , we can estimate the above probabilities 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, the probability of at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

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  $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_j)$ , using the  $J$  administered items, can be estimated as

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

$$p_{i1} = P(-\infty < \theta_i < cut_1 | \mathbf{z}, \mathbf{b}) = \frac{\int_{-\infty}^{cut_1} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}{\int_{-\infty}^{+\infty} L(\theta | \mathbf{z}, \mathbf{b}) d\theta}$$

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

where the likelihood function, based on general IRT models, is

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

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 2PL models,  $c_j = 0$ ),  $D$  is 1.7 for non-Rasch models and 1 for Rasch model.

### Classification Accuracy

Using  $p_{il}$ , we can construct a  $L \times L$  table as

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

where  $n_{alm} = \sum_{pl_i=l} p_{im}$ .  $n_{alm}$  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.

### Classification Consistency

Using  $p_{il}$ , similar to accuracy, we can construct another  $L \times L$  table by assuming the test is administered twice independently to the same student group, hence we have

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

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 equivalent test form.

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

$$CC_l = \frac{n_{c ll}}{\sum_{m=1}^L n_{c lm}},$$

and the overall classification consistency is

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

The analysis of the classification index is performed based on overall scale scores. Table 30 provides the proportion of classification accuracy and consistency for overall and by achievement level.

The overall classification index ranged from 77% to 84% for the accuracy and from 69% to 77% 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 intervals used to compute the classification probability to classify students into L1  $[-\infty, L2 \text{ cut}]$  or L4  $[L4 \text{ cut}, \infty]$  is wider than the intervals used in L2  $[L2 \text{ cut}, L3 \text{ cut}]$  and L3  $[L3 \text{ cut}, L4 \text{ cut}]$ . The misclassification probability tends to be higher for narrow intervals.

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



Table 28. Classification Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	79	71	83	76
	L1	89	83	90	84
	L2	70	59	73	64
	L3	67	56	79	71
	L4	88	83	90	84
4	Overall	77	69	84	77
	L1	89	83	90	84
	L2	61	49	80	73
	L3	63	53	79	71
	L4	87	81	90	85
5	Overall	79	71	83	76
	L1	90	84	91	86
	L2	64	52	77	68
	L3	72	63	71	61
	L4	87	80	90	85
6	Overall	78	69	83	76
	L1	88	81	92	87
	L2	68	57	77	70
	L3	73	65	72	62
	L4	85	76	90	84
7	Overall	78	70	83	76
	L1	89	82	91	85
	L2	67	56	76	68
	L3	76	68	75	65
	L4	85	76	90	85
8	Overall	79	70	82	76
	L1	88	81	91	86
	L2	70	59	72	62
	L3	77	69	72	62
	L4	83	74	91	86

## 5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores and achievement levels are also computed by subgroups. Tables 31 and 32 present the marginal reliability coefficients by the subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for Limited English Proficiency (LEP) and IDEA subgroups, a large percentage of whom received Level 1 with large SEMs.

Table 29. Marginal Reliability Coefficients for Overall and by Subgroup for ELA/L

<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>
All Students	0.92	0.91	0.91	0.90	0.91	0.91
Female	0.91	0.90	0.91	0.90	0.90	0.90
Male	0.91	0.91	0.92	0.90	0.91	0.91
American Indian/Alaska Native	0.84	0.88	0.91	0.89	0.90	0.88
Asian	0.90	0.89	0.90	0.88	0.89	0.89
African American	0.89	0.89	0.89	0.88	0.89	0.88
Hispanic/Latino	0.90	0.89	0.90	0.89	0.89	0.89
Pacific Islander	0.91	0.91	0.85	0.92	0.92	0.90
White	0.90	0.89	0.89	0.88	0.88	0.89
Multiple Ethnicities	0.92	0.90	0.91	0.90	0.90	0.90
Limited English Proficiency	0.85	0.84	0.82	0.78	0.78	0.75
IDEA	0.86	0.87	0.87	0.85	0.85	0.85

Table 30. Marginal Reliability Coefficients for Overall and by Subgroup for Mathematics

<b>Subgroup</b>	<b>Grade 3</b>	<b>Grade 4</b>	<b>Grade 5</b>	<b>Grade 6</b>	<b>Grade 7</b>	<b>Grade 8</b>
All Students	0.94	0.95	0.94	0.94	0.93	0.93
Female	0.94	0.94	0.93	0.94	0.93	0.93
Male	0.95	0.95	0.94	0.94	0.93	0.93
American Indian/Alaska Native	0.92	0.93	0.93	0.93	0.91	0.89
Asian	0.94	0.94	0.95	0.95	0.94	0.94
African American	0.93	0.93	0.89	0.90	0.88	0.86
Hispanic/Latino	0.93	0.93	0.90	0.91	0.89	0.88
Pacific Islander	0.94	0.94	0.93	0.95	0.95	0.94
White	0.94	0.94	0.93	0.94	0.93	0.93
Multiple Ethnicities	0.95	0.95	0.94	0.94	0.93	0.93
Limited English Proficiency	0.92	0.90	0.82	0.83	0.76	0.73
IDEA	0.92	0.91	0.87	0.87	0.83	0.83

## 5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability coefficients and the measurement errors are also computed for the claim scores. In mathematics, claims 2 and 4 are combined to have enough items to generate a score. Because the precision of scores in claims is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three achievement categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 33 and 34 present the marginal reliability coefficients for each reporting category score in ELA/L and mathematics, respectively.

Table 31. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Reporting Categories	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1: Reading	14	16	0.79	2430	101	46
	Claims 2 & 4: Writing & Research	16	16	0.83	2428	98	40
	Claim 3: Listening	8	8	0.59	2436	117	75
4	Claim 1: Reading	14	16	0.74	2477	106	54
	Claims 2 & 4: Writing & Research	16	16	0.80	2473	102	46
	Claim 3: Listening	8	8	0.63	2472	126	76
5	Claim 1: Reading	14	16	0.75	2512	110	55
	Claims 2 & 4: Writing & Research	16	16	0.83	2509	107	44
	Claim 3: Listening	8	9	0.62	2512	125	76
6	Claim 1: Reading	14	16	0.75	2528	110	55
	Claims 2 & 4: Writing & Research	16	16	0.80	2528	105	47
	Claim 3: Listening	8	9	0.53	2555	122	83
7	Claim 1: Reading	14	16	0.78	2557	111	52
	Claims 2 & 4: Writing & Research	16	16	0.80	2550	113	51
	Claim 3: Listening	8	9	0.55	2558	121	81
8	Claim 1: Reading	16	16	0.78	2569	111	52
	Claims 2 & 4: Writing & Research	16	16	0.80	2560	111	49
	Claim 3: Listening	8	9	0.51	2582	125	87

Table 32. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Reporting Categories	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1	20	20	0.90	2442	87	27
	Claims 2 and 4	8	11	0.69	2429	98	55
	Claim 3	9	11	0.73	2435	95	49
4	Claim 1	20	20	0.90	2483	88	28
	Claims 2 and 4	8	10	0.74	2474	100	51
	Claim 3	9	10	0.75	2477	98	49
5	Claim 1	20	20	0.89	2506	96	32
	Claims 2 and 4	8	10	0.61	2489	123	77
	Claim 3	9	10	0.71	2497	112	60
6	Claim 1	19	19	0.89	2528	113	38
	Claims 2 and 4	9	10	0.72	2515	123	65
	Claim 3	10	11	0.74	2523	117	59
7	Claim 1	20	20	0.88	2541	116	40
	Claims 2 and 4	10	10	0.65	2525	133	79
	Claim 3	8	10	0.67	2537	126	73
8	Claim 1	20	20	0.88	2554	125	43
	Claims 2 and 4	8	10	0.60	2537	149	95
	Claim 3	9	10	0.67	2545	136	79

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 item parameters that are vertically scaled by linking across 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 performance category for each reporting category. This section describes the rules used in generating scores and the handscoring procedure.

### 6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The Smarter Balanced tests 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

$$L_j(\theta_j | \mathbf{z}_j, \mathbf{a}, b_1, \dots, b_k) = \prod_{i=1}^I p_{ij}(z_{ij} | \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}),$$

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  $j$ ,  $k$  indexes 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, we have  $m_i = 1$ ,

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

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:

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

where  $m_i$  is the maximum possible score point (starting from 0) for the  $i$ th item,  $D$  is the scale factor, 1.7. The SE is calculated based only on the answered item(s) 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 35 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 33. Vertical Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA/L	3–8	85.8	2508.2
Math	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,  $SE_{\theta}$  is the standard error of the ability estimate on the  $\Theta$  scale, and  $a$  is the slope of the scaling constant that transforms  $\Theta$  to the reporting scale.

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 36 provides three achievement standards for each grade and content area.

Table 34. 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	2493	2583	2682	2543	2628	2718

### 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 37 presents the lowest obtainable score (LOT or LOSS) and the highest obtainable score (HOT or HOSS) in both theta and scale score metrics. Estimated theta's lower than LOT or higher than HOT are truncated to the LOT and HOT values, and assign LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and subscores). The standard error for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 35. 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
ELA/L	4	-4.3962	1.8014	2131	2663
ELA/L	5	-3.5763	2.2498	2201	2701
ELA/L	6	-3.4785	2.5140	2210	2724
ELA/L	7	-2.9114	2.7547	2258	2745
ELA/L	8	-2.5677	3.0430	2288	2769
Math	3	-4.1132	1.3335	2189	2621
Math	4	-3.9204	1.8191	2204	2659
Math	5	-3.7276	2.3290	2219	2700
Math	6	-3.5348	2.9455	2235	2748
Math	7	-3.3420	3.3238	2250	2778
Math	8	-3.1492	3.6254	2265	2802

### 6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

In 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 REPORTING CATEGORIES (CLAIM SCORES)

In both ELA/L and mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim, three performance categories, relative strength and weakness are produced. If the difference between the proficiency cut score and the claim score is greater (or less) than 1.5 times standard error of the claim, a plus or minus indicator appears on the student’s score report as shown in Section 7.

For summative tests, the specific rules are as follows:

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

where  $SS_{rc}$  is the student’s scale score on a reporting category;  $SS_p$  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 reporting category. For HOSS and LOSS are automatically assigned to *Above Standard* and *Below Standard*, respectively.

## 6.6 TARGET SCORES

The target-level reports are not possible to produce for a fixed-form test because the number of items included per target is too few 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 reflect the benchmark narrowly 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 level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given target. A target score is an aggregate of the differences in student overall proficiency and the differences in the difficulty of the items measuring a target in a class, school, or district area. Target scores are computed for attempted tests based on the responded items. Target scores are computed in each claim (three claims) in ELA/L and Claim 1 only in mathematics.

Target scores are computed in two ways: (1) target scores relative to a student’s overall estimated ability ( $\theta$ ), 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, we use the 2PL IRT model 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 generalized partial credit model, 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:



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

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 individual student target scores for the target, across students of different abilities receiving different items measuring the same target at different levels of difficulty,

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

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.

In the aggregate, a target performance is reported as a group of students performing better, worse, or as expected on this target. In some cases, insufficient information will be available and that will be indicated as well.

For target level strengths/weakness, we will report the following:

- 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  $p_{ij} = p(z_{ij} = 1)$ , 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 we use the 2PL IRT model 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 generalized partial credit model, 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:

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

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 individual student target scores for the target, across students of different abilities receiving different items measuring the same target at different levels of difficulty,

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

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.

We do not suggest direct reporting of the statistic  $\bar{\delta}_{Tg}$ ; instead, we recommend reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this target. In some cases, insufficient information will be available and that will be indicated as well.

For target level strengths/weakness, we will report the following:

- 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 HANDSCORING

AIR provides the automated electronic scoring and Measurement Incorporated (MI) provides all handscoring for the Smarter Balanced summative tests. In ELA/L, short-answer (SA) items and Full Write items are scored by human raters; this is also referred to as “handscored.” In mathematics, SA items and other constructed-response items are handscored. The procedure for scoring these items is provided by Smarter Balanced.

Outlined below is the scoring process MI follows. This procedure is used to score responses to all constructed-response or written composition items.

### **6.7.1 Reader Selection**

MI maintains a large pool of readers at each scoring center, as well as distributive readers who work remotely from their homes. Experienced readers are defined as those who have worked on one or more previous projects and typically comprise 50–65% of all readers. 2016–2017 was the third year that MI scored operational Smarter Balanced assessments, and it is estimated that approximately twice as many experienced readers returned in comparison to 2015–2016, particularly in the distributive reader pool. MI only needs to inform experienced readers that a project is pending and invite them to return. MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. MI employs many of these experienced readers for the Smarter Balanced project and recruits new ones as well.

MI procedures for selecting new readers are very thorough. After advertising and receiving applications, MI staff review the applications and schedule interviews for qualified applicants (i.e., those with a four-year college degree). Each qualified applicant must pass an interview by experienced MI staff, complete ELA/L and mathematics placement assessments, complete a grammar exercise, write an acceptable essay, and receive good recommendations from references. MI then reviews all the information about an applicant before offering employment.

In selecting team leaders, MI management staff and scoring directors review the files of all returning staff. They look for people who are experienced team leaders with a record of good performance on previous projects and also consider readers who have been recommended for promotion to the team leader position.

MI is an equal opportunity employer that actively recruits minority staff. Historically, MI's temporary staff on major projects averages about 51% female, 49% male, 76% Caucasian, and 24% minority. MI strongly opposes illegal discrimination against any employee or applicant for employment with respect to hiring, tenure, terms, conditions, or privileges of employment; or any matter directly or indirectly related to employment, because of race, color, religion, sex, age, handicap, national origin, or ancestry.

MI requires all handscoring project staff (scoring directors, team leaders, readers, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or 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 Reader Training**

All readers hired for Smarter Balanced assessment handscoring are 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. The only changes made to anchor sets across the years include occasional updates to annotations and removal of individual responses, as determined during annual meetings between the vendors and Smarter Balanced. Additionally, several of the Brief Writes anchor sets were revised between the 2014–2015 and 2015–2016 test administrations. Finally, based on challenges observed scoring the 2014–2015 and 2015–2016 administrations, in the summer of 2016 MI scoring managers developed additional item-level supplemental training materials for their respective content areas to use in conjunction with the Smarter Balanced-provided materials.

Once hired, readers are placed into a scoring group that corresponds to the subject/grade that they are deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). Readers are trained on a specific item type (i.e., Brief Writes, Reading, Research, Full Writes, and/or Mathematics). Within each group, readers are divided into teams consisting of one team leader and 10–15 readers. Each team leader and reader is assigned a unique number for easy identification of their scoring work throughout the scoring session. For the 2016–2017 administration, scoring directors attempted to minimize the number of items an individual reader scored so that the reader became highly experienced in scoring responses to those items.

MI’s Virtual Scoring Center (VSC) includes an online training interface which presents rubrics, scoring guides, and training/qualifying sets. Readers are trained by a scoring director (in-person) or using scripted videos (online). The same training protocol is followed for both site-based and distributive readers.

After the contracts and nondisclosure forms are signed and the scoring director completes his or her introductory remarks, training begins. Reader training and team leader training follow the same format. The scoring director presents the writing or constructed-response task and introduces the scoring guide (anchor set), then discusses each score point with the entire room. This presentation is followed by practice scoring on the training/qualifying sets. The scoring director reminds the readers to compare each training/qualifying set response to anchor responses in the scoring guide to ensure consistency in scoring the training/qualifying responses.

All scoring personnel log in to MI’s secure Scoring Resource Center (SRC). The SRC includes all online training modules, is the portal to the VSC interface, and is the data repository of all scoring reports that are used for reader monitoring.

After completing the first training set, readers are provided a rationale for the score of each response presented in the set. Training continues until all training/qualifying sets have been scored and discussed.

Like team leaders, readers must demonstrate their ability to score accurately by attaining the qualifying agreement percentage established by Smarter Balanced before they may score actual student responses. Any readers unable to meet the qualifying standards are not permitted to score that item. Readers who reach the qualifying standard on some items but not others will only score the items on which they have successfully qualified. All readers understand this stipulation when they are hired.

Training is carefully orchestrated so that readers understand how to apply the rubric in scoring the responses, reference the scoring guide, develop the flexibility needed to handle a variety of responses, and retain the consistency needed to score all responses accurately. In addition to completing all of the initial training and qualifications, significant time is allotted for demonstrations of the VSC handscoring system, explanations of how to “flag” unusual responses for review by the scoring director, and instructions about other procedures necessary for the conduct of a smooth project.

Training design varies slightly depending on Smarter Balanced item type:

- Full writes: readers train and qualify on baseline sets for each grade and writing purpose (Grade 3 Narrative, Grade 6 Argumentative, etc.), then take qualifying sets for each item in that grade and purpose.
- Brief writes, reading, and research: readers train and qualify on a baseline set within a specific grade band and target.
- Mathematics: readers train on baseline items, which qualify the readers for that item as well as

any items associated with it; for items with no associated items, training is for the specific item.

Reader training time varies by grade and content area. Training for brief writes, reading, research, and many mathematics items can be accomplished in one day, while training for full writes may take up to five days to complete. Readers generally work 6.5 hours per day, excluding breaks. Evening shift readers work 3.75 hours, excluding breaks.

Multiple strategies are used to minimize rater bias. First, readers do not have access to any student identifiers. Unless the students sign their names, write about their hometowns, or in some way provide other identifying information as part of their response, the readers have no knowledge of student characteristics. Second, all readers are trained using Smarter Balanced-provided materials, which were approved as unbiased examples of responses at the various score points. Training involves constant comparisons with the rubric and anchor papers so that readers' judgments are based solely on the scoring criteria. Finally, following training, a cycle of diagnosis and feedback is used to identify any issues. Specifically, during scoring, readers are monitored and any instances of readers making scoring decisions based on anything but the criteria are discussed. Readers are further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback they are dismissed.

### **6.7.3 Reader Statistics**

One concern regarding the scoring of any open-response assessment is the reliability and accuracy of the scoring. MI appreciates and shares this concern and continually develops new and technically sound methods of monitoring reliability. Reliable scoring starts with detailed scoring rubrics and training materials, and thorough training sessions by experienced trainers. Quality results are achieved by daily monitoring of each reader.

In addition to extensive experience in the preparation of training materials and employing management and staff with unparalleled expertise in the field of handscored educational assessment, MI constantly monitors the quality of each reader's work throughout every project. Reader status reports are used to monitor readers' scoring habits during the Smarter Balanced handscoring project.

MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the readers' scores are submitted into the VSC handscoring system, the data are uploaded into the scoring data report servers located at MI's corporate headquarters in Durham, North Carolina.

More than 20 reports are available and can be customized to meet the information needs of the client and MI's scoring department, providing the following data:

- Reader ID and team
- Number of responses scored
- Number of responses assigned each score point (1–4 or other)
- Percentage of responses scored that day in exact agreement with a second reader
- Percentage of responses scored that day within one point agreement with a second reader
- Number and percentage of responses receiving adjacent scores at each line (0/1, 1/2, 2/3, etc.)
- Number and percentage of responses receiving nonadjacent scores at each line
- Number of correctly assigned scores on the validity responses

Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available for access by the handscoring project monitors at each MI scoring center via a secure website, and the handscoring project monitors provide updated reports to the scoring directors several times per day. MI scoring directors are experienced in examining these reports and using the information to determine the need for retraining of individual readers or the group as a whole. It can easily be determined if a reader is consistently scoring high or low, and the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all retraining efforts.

#### **6.7.4 Reader Monitoring and Retraining**

Team leaders spot-check (i.e., read behind) each reader’s scoring to ensure that he or she is on target, and conduct one-on-one retraining sessions about any problems found. At the beginning of the project, team leaders read behind every reader every day; they become more selective about the frequency and number of read-behinds as readers become more proficient at scoring. The daily reader reliability reports and validity/calibration results are used to identify the readers who need more frequent monitoring.

Retraining is an ongoing process once scoring is underway. Daily analysis of the reader status reports enables management personnel to identify individual or group retraining needs. If it becomes apparent that a whole team or a whole group is having difficulty with a particular type of response, large group training sessions are conducted. Standard retraining procedures include room-wide discussions led by the scoring director, team discussions conducted by team leaders, and one-on-one discussions with individual readers. It is standard practice to conduct morning room-wide retraining at MI each day, with a more extensive retraining on Monday mornings in order to re-anchor the readers after a weekend away from scoring.

Each student response is scored holistically by a trained and qualified reader using the scoring criteria developed and approved by Smarter Balanced, with a second read conducted on 15% of responses for each item for reliability purposes. Responses are selected randomly for second reading and scored by readers who are not aware of the score assigned by the first reader or even that the response has been read before. MI’s QA/reliability procedures allow the handscoring staff to identify struggling readers very early and begin retraining at once. While retraining these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI’s monitoring is also used as a retraining method. MI shows readers responses that the readers have scored incorrectly, explains the correct scores, and has the readers change the scores. Between the 2014–2015 and 2015–2016 test administrations, MI developed dynamic “threshold” reports which, based on inputted criteria, immediately identify potential scoring performance issues. This enhancement allows scoring leadership to pinpoint areas of concern and take corrective action with greater efficiency than ever before.

During scoring, readers occasionally send responses to their leadership for review and/or scoring. These types of responses most commonly include non-scorable responses such as off-topic or foreign language responses that are difficult to score using the available rubrics and reference responses, and at-risk responses that are alerted for action by the client State.

#### **6.7.5 Reader Validity Checks**

Approved responses are loaded into the VSC system as validity responses. A small set of validity responses are 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. These responses are imbedded into live scoring on an ongoing basis to be scored by the readers. A validity report is generated that includes the response identification number, the score(s) assigned by the readers, and the “true” scores. A daily and project-to-date summary of percentages of correct scores and low/high considerations at each score point is also provided. If it is determined that a validity response and/or item is performing poorly, scoring management reviews the validity responses to ensure that the true scores have been entered correctly. If so, then retraining may be conducted with the readers using the validity data as a guide for how to focus the retraining. If the true scores have been entered incorrectly, then the database is updated to show the correct true scores. Validity results are not used in isolation but as one piece of evidence along with the second read and read-behind agreement to make decisions about retraining and dismissing readers.

### **6.7.6 Reader Dismissal**

When read-behinds or daily statistics identify a reader who cannot maintain acceptable agreement rates, the reader is retrained and monitored by scoring leadership personnel. A reader may be released from the project if retraining is unsuccessful. In these situations, all items scored by a reader during the timeframe in question can be identified, reset, and released back into the scoring pool. The aberrant reader’s scores are deleted, and the responses are redistributed to other qualified readers for rescoring.

### **6.7.7 Reader Agreement**

The inter-reader reliability is computed based on scorable responses (numeric scores) scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) which are scored by scoring leadership, not by two independent readers. The inter-reader reliability is based on the readers who scored student responses in Connecticut.

In ELA/L, the short answer items are scored in 0–2. In mathematics, the maximum score points of the hand-scored items range from 1–3.

Tables 38–39 provide a summary of the inter-reader reliability based on items with a sample size greater than 50. The inter-reader reliability is presented with %exact agreement, minimum and maximum %exact agreements, combined %exact and %adjacent agreement, and quadratic weighted Kappa (QWK).

Table 38. ELA/L Reader Agreements for Short-Answer Items

Grade	# of Items	%Exact			% (Exact+Adjacent)	QWK
		Average	Min	Max		
3	18	79	71	91	100	0.71
4	30	79	65	90	100	0.74
5	21	76	66	84	100	0.70
6	18	74	62	85	100	0.66
7	23	73	60	88	100	0.63
8	25	74	56	90	100	0.67

Table 39. Mathematics Reader Agreements

Grade	Score Points	# of Items	%Exact			% (Exact+ Adjacent)	QWK
			Average	Min	Max		
3	1	12	92	90	95	100	0.82
4	1	8	86	80	93	100	0.68
5	1	4	93	91	97	100	0.72
6	1	14	96	86	99	100	0.90
7	1	8	97	94	98	100	0.82
8	1	15	89	80	98	100	0.75
3	2	26	89	79	100	100	0.91
4	2	36	91	77	98	100	0.91
5	2	41	90	78	97	100	0.88
6	2	32	88	78	98	100	0.88
7	2	30	87	73	93	100	0.84
8	2	26	86	75	99	100	0.87
3	3	4	95	92	96	99	0.97
4	3	4	88	85	91	99	0.93
5	3	8	86	81	99	97	0.82
7	3	3	75	66	82	98	0.83



## 7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that include the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and the tests are handscored. Because the score report on students' performance are updated each time that students complete tests and these tests are handscored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to individual students' score reports, the Online Reporting System also produces aggregate score reports by class, schools, districts, and states. It should be noted that the ORS does not produce aggregate score reports for state. The timely accessibility of aggregate score reports could help users monitor students testing 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. Additionally, the ORS provides participation data that helps monitor the student participation rate. In 2016–2017, some new features are added to ORS reports.

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

### 7.1 ONLINE REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

#### 7.1.1 Types of Online Score Reports

The ORS is designed to help educators and students answer questions regarding how well students have performed on ELA/L and mathematics assessments. The ORS is the online tool to provide educators and other stakeholders with timely, relevant score reports. The ORS for the Smarter Balanced assessments has been designed with stakeholders, who are not technical measurement experts, in mind, ensuring that test results are presented as easy to read and understand by using simple language so that users can quickly understand assessment results and make inferences about student achievement. The ORS 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.

Once authorized users log in to the ORS and select “Score Reports,” the online score reports are presented hierarchically. The ORS starts with presenting summaries on student performance by subject and grade at a selected aggregate level. To view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down menu with a list of aggregate units, e.g., schools within a district, or teachers within a school, to select. For more detailed student assessment results for a school, a teacher, or a roster, users can select the subject and grade on the online score reports.

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 40 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 *Online Reporting System User Guide*, located in a help button on the ORS.

Table 36. 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"> <li>• Number of students tested and percent of students with Level 3 or 4 (overall students and by subgroup)</li> <li>• Average scale score and standard error of average scale score (overall students and by subgroup)</li> <li>• Percent of students at each achievement level on overall test and by claims (overall students and by subgroup)</li> <li>• Performance category in each target (overall students)<sup>1</sup></li> <li>• Participation rate (overall students)<sup>2</sup></li> <li>• On-demand student roster report</li> </ul>
Student	<ul style="list-style-type: none"> <li>• Total scale score and standard error of measurement</li> <li>• Achievement level on overall and claim scores with achievement level descriptors</li> <li>• Average scale scores and standard errors of average scale scores for student’s school, and district</li> </ul>

*Note.*

1: Performance category in each target is provided for all aggregate levels.

2: Participation rate reports are provided at district and school level.

The aggregate score reports at a selected aggregate level are provided for overall students and by subgroups. Users can see student assessment results by any of the subgroups. Table 41 presents the types of subgroups and subgroup categories provided in ORS.

Table 37. 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	American Indian or Alaska Native Asian Black or African American Two or More Races Hispanic or Latino White Native Hawaiian or Other Pacific Islander

## 7.1.2 The Online Reporting System

### 7.1.2.1 Home Page

When users log in to the ORS and select “Score Reports”, the first page displays summaries of students’ performance across grades and subjects. District personnel see district summaries, school personnel see school summaries, and teachers see class summaries of their students. Using a drop-down menu with a list of aggregate units, users can see a summary of students’ performance for the lower aggregate unit as well. For example, the district personnel can see a summary of students’ performance for schools as well as the district.

The home page summarizes students’ performance including (1) number of students tested, and (2) percentage of students at Level 3 or above. Exhibit 1 presents a sample home page at a district level.

Exhibit 1. Home Page: District Level

### Home Page Dashboard

**Select Test and Year**

Test: Smarter Summative ▼

Administration: 2016-2017 ▼

Scores for students who were mine at the end of the selected administration  
 Scores for my current students  
 Scores for students who were mine when they tested during the selected administration

**Select**

Demo District (999) ▼

[Click on a grade and subject to view more information.](#)

#### Number of Students Tested and Percent of Students at Level 3 or Above for Students in Demo District, 2016-2017

ELA/Literacy

Grade	Number of Students Tested	Percent at Level 3 or Above
Grade 3	90	30%
Grade 4	90	28%
Grade 5	99	30%
Grade 6	169	33%
Grade 7	182	34%
Grade 8	102	40%

Mathematics

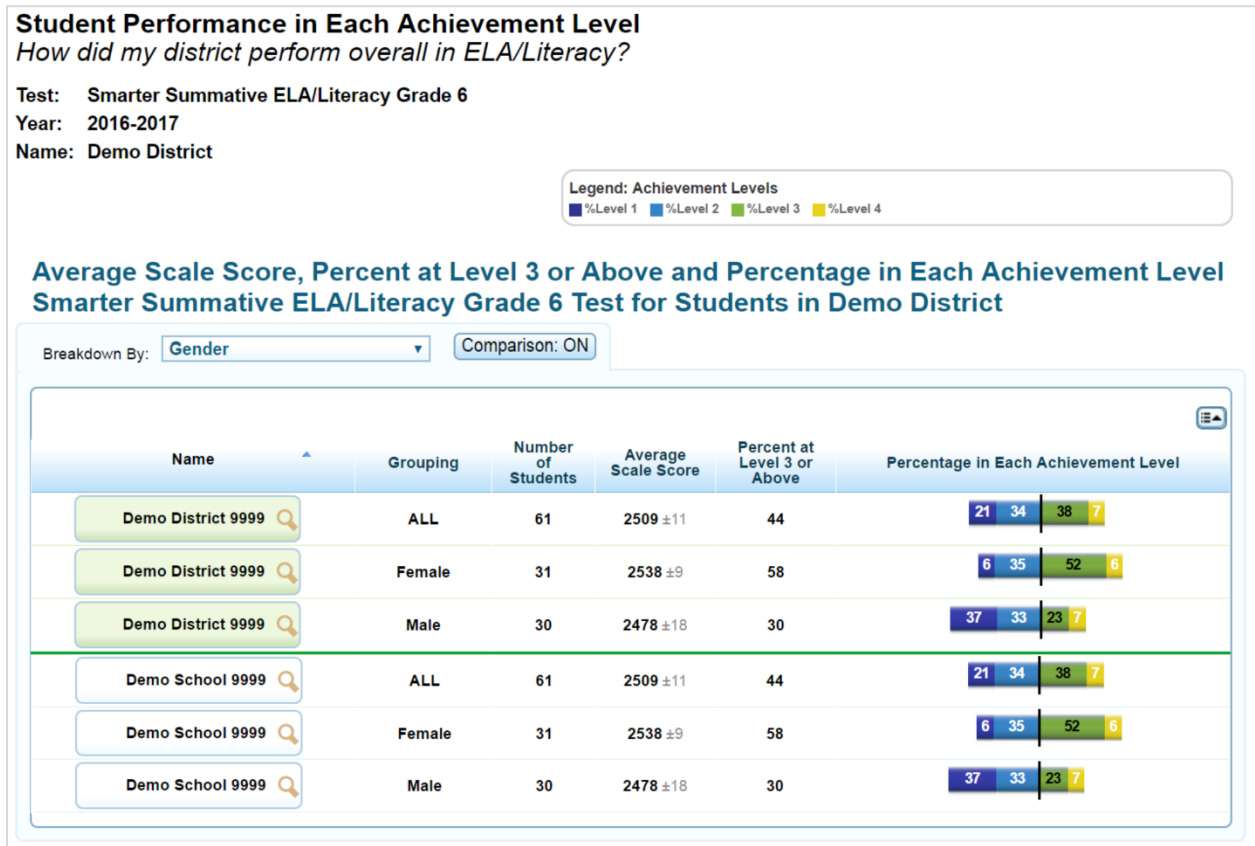
Grade	Number of Students Tested	Percent at Level 3 or Above
Grade 3	90	27%
Grade 4	89	18%
Grade 5	99	12%
Grade 6	298	24%
Grade 7	192	28%
Grade 8	253	26%

7.1.2.2 Subject Detail Page

More detailed summaries of student performance on each grade in a subject area for a selected aggregate level are presented when users select a grade within a subject on the home page. On each aggregate report, the summary report presents the summary results for the selected aggregate unit as well as the summary results for the aggregate unit above the selected aggregate. For example, if a school is selected on the subject detail page, the summary results of the district are provided above the school summary results as well, so that the school performance can be compared with the above aggregate levels.

The subject detail page provides the aggregate summaries on a specific subject area including (1) number of students, (2) average scale score and standard error of the average scale score, (3) percent of students at Level 3 or above, and (4) percent of students in each achievement level. The summaries are also presented for overall students and by subgroups. Exhibit 2 presents an example of a subject detail page for ELA/L at a district level when a user select a subgroup of gender.

Exhibit 2. Subject Detail Page for ELA/L by Gender: District Level

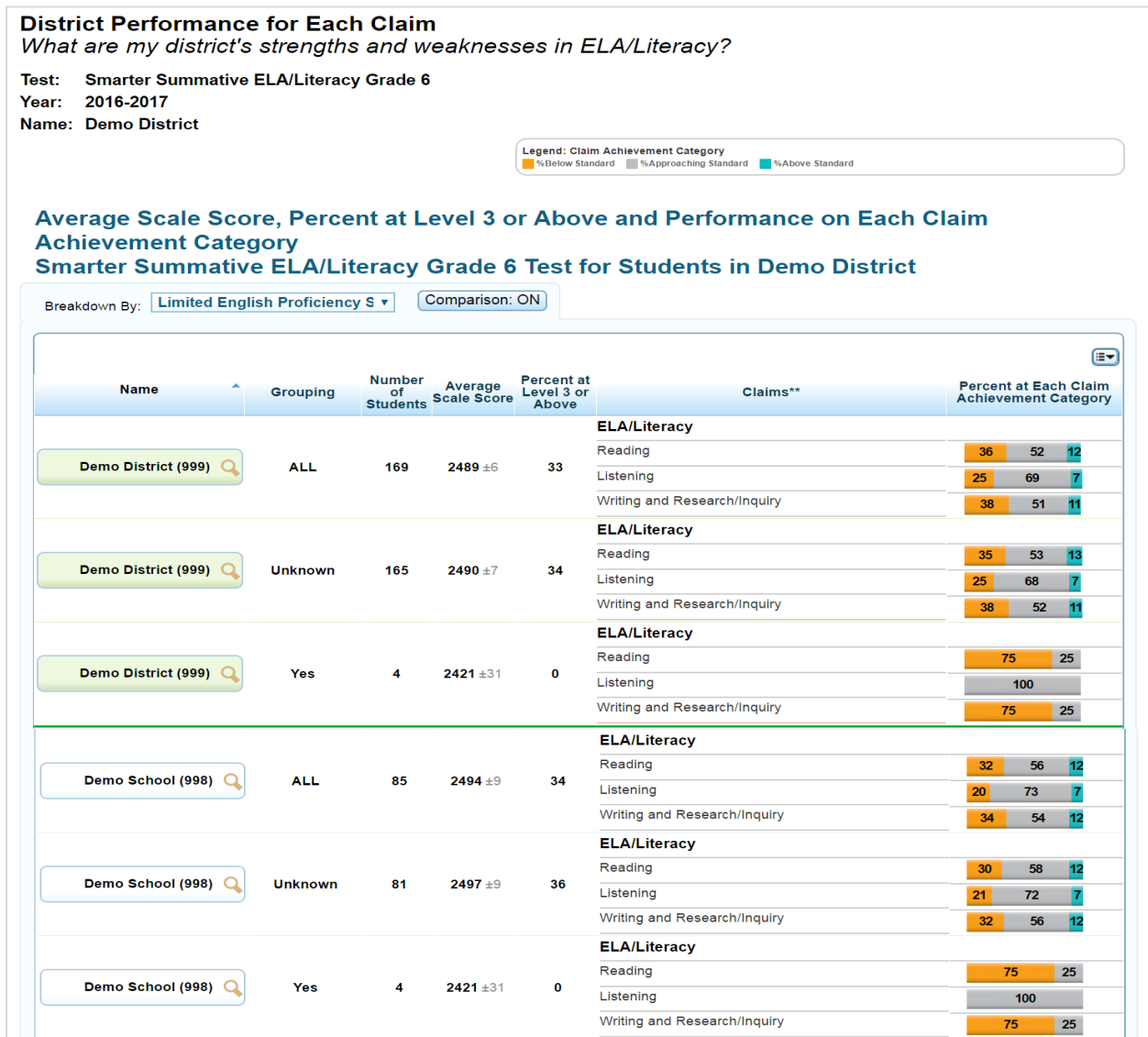


7.1.2.3 Claim Detail Page

The claim detail page provides the aggregate summaries on student performance in each claim for a particular grade and subject. The aggregate summaries on the claim detail page include (1) number of students, (2) average scale score and standard error of the average scale score, (3) percent of students at Level 3 or above, and (4) percent of students in each claim performance category.

Similar to the subject detail page, the summary report presents the summary results for the selected aggregate unit as well as the summary results for the aggregate unit above the selected aggregate. Also, the summaries on claim-level performance can be presented for overall students and by subgroup. Exhibit 3 presents an example of a claim detail page for ELA/L at a district level when users select a subgroup of LEP status.

Exhibit 3. Claim Detail Page for ELA/L by LEP Status: District Level



7.1.2.4 Target Detail Page

The target detail page provides the aggregate summaries on student performance in each target, including: (1) strength or weakness indicators in each target that are computed in two ways (i.e., performance relative to proficiency, performance relative to the test as a whole, and (2) average scale scores and standard errors of average scale scores for the selected aggregate unit and the aggregate unit above the selected aggregate. It should be noted that the summaries on target-level student performance are generated for overall students only. That is, the summaries on target-level student performance are not generated by subgroup. Exhibits 4–7 present examples of target detail pages for ELA/L and mathematics at the school level and the teacher level.

Exhibit 4. Target Detail Page for ELA/L: School Level

### Performance on Each Target for the ELA/Literacy Test

*What are my school's relative strengths and weaknesses in the ELA/Literacy Targets?*

**Test:** Smarter Summative ELA/Literacy Grade 6  
**Year:** 2016-2017  
**Name:** Demo School

**Legend: Performance Relative to Proficiency**

- + Performance is above the Proficiency Standard
- = Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- ★ Insufficient Information

**Legend: Performance Relative to the Test as a Whole**

- + Performance is better than on the rest of the test
- = Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- ★ Insufficient Information

**Comparison Scores**

Name	Average Scale Score
Demo District (999)	2489 ±6
Demo School (999)	2390 ±13

**Performance on Each Target**  
**Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo School**

Target	Performance Relative to Proficiency	Performance Relative to the Test as a Whole
<b>Reading</b>		
(Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	-	=
(Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide a summary of the text distinct from personal opinions or judgement.	-	=
(Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	-	=
(Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	-	=
(Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	=	=

Exhibit 5. Target Detail Page for ELA/L: Class Level

**Performance on Each Target for the ELA/Literacy Test**

*What are my roster's relative strengths and weaknesses in the ELA/Literacy Targets?*

**Test:** Smarter Summative ELA/Literacy Grade 6  
**Year:** 2016-2017  
**Name:** Demo Roster

**Legend: Performance Relative to Proficiency**

- + Performance is above the Proficiency Standard
- = Performance is near the Proficiency Standard
- Performance is below the Proficiency Standard
- ★ Insufficient Information

**Legend: Performance Relative to the Test as a Whole**

- + Performance is better than on the rest of the test
- = Performance is similar to performance on the test as a whole
- Performance is worse than on the rest of the test
- ★ Insufficient Information

**Comparison Scores**

Name	Average Scale Score
Demo District (999)	2569 ±11
Demo School (999)	2569 ±11
Demo Teacher	2594 ±21
Demo Roster	2594 ±21

**Performance on Each Target**  
**Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo Roster**

Target	Performance Relative to Proficiency	Performance Relative to the Test as a Whole
<b>Reading</b>		
(Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	<span style="color: green;">+</span>	<span style="color: blue;">=</span>
(Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide a summary of the text distinct from personal opinions or judgement.	<span style="color: blue;">=</span>	<span style="color: blue;">=</span>
(Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	<span style="color: blue;">=</span>	<span style="color: blue;">=</span>
(Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	<span style="color: green;">+</span>	<span style="color: blue;">=</span>
(Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and	<span style="color: blue;">=</span>	<span style="color: red;">-</span>

Exhibit 6. Target Detail Page for Mathematics: School Level

**Performance on Each Target for the Mathematics Test**





*What are my school's relative strengths and weaknesses in the Mathematics Targets?*

Test: Smarter Summative Mathematics Grade 6





Year: 2016-2017

Name: Demo School



**Legend: Performance Relative to Proficiency**

-  Performance is above the Proficiency Standard
-  Performance is near the Proficiency Standard
-  Performance is below the Proficiency Standard
-  Insufficient Information

**Legend: Performance Relative to the Test as a Whole**

-  Performance is better than on the rest of the test
-  Performance is similar to performance on the test as a whole
-  Performance is worse than on the rest of the test
-  Insufficient Information

**Comparison Scores**

Name	Average Scale Score
Demo District (999) 	2539 ±14
Demo School (999) 	2539 ±14

**Performance on Each Target**

**Smarter Summative Mathematics Grade 6 Test for Students in Demo School**





















Target	Performance Relative to Proficiency	Performance Relative to the Test as a Whole
<b>Concepts and Procedures</b>		
Understand ratio concepts and use ratio reasoning to solve problems.		
Apply and extend previous understandings of multiplication and division to divide fractions by fractions.		
Compute fluently with multi-digit numbers and find common factors and multiples.		
Apply and extend previous understandings of numbers to the system of rational numbers.		
Apply and extend previous understandings of arithmetic to algebraic expressions.		
Reason about and solve one-variable equations and inequalities.		
Represent and analyze quantitative relationships between dependent and independent variables.		
Solve real-world and mathematical problems involving area, surface area, and volume.		
Develop understanding of statistical variability.		
Summarize and describe distributions.		



Exhibit 7. Target Detail Page for Mathematics: Teacher Level

**Performance on Each Target for the Mathematics Test**

*What are my roster's relative strengths and weaknesses in the Mathematics Targets?*

**Test:** Smarter Summative Mathematics Grade 6

**Year:** 2016-2017

**Name:** Demo Roster

**Legend: Performance Relative to Proficiency**

- + Performance is above the Proficiency Standard
- ▬ Performance is near the Proficiency Standard
- ▬ Performance is below the Proficiency Standard
- ★ Insufficient Information

**Legend: Performance Relative to the Test as a Whole**

- + Performance is better than on the rest of the test
- ▬ Performance is similar to performance on the test as a whole
- ▬ Performance is worse than on the rest of the test
- ★ Insufficient Information

**Comparison Scores**

Name	Average Scale Score
Demo District (999)	2539 ±14
Demo School (999)	2539 ±14
Demo Teacher	2579 ±21
Demo Roster	2579 ±21

**Performance on Each Target**  
**Smarter Summative Mathematics Grade 6 Test for Students in Demo Roster**

Target	Performance Relative to Proficiency	Performance Relative to the Test as a Whole
<b>Concepts and Procedures</b>		
Understand ratio concepts and use ratio reasoning to solve problems.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Compute fluently with multi-digit numbers and find common factors and multiples.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Apply and extend previous understandings of numbers to the system of rational numbers.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Apply and extend previous understandings of arithmetic to algebraic expressions.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Reason about and solve one-variable equations and inequalities.	<span style="color: green;">+</span>	<span style="color: lightblue;">▬</span>
Represent and analyze quantitative relationships between dependent and independent variables.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Solve real-world and mathematical problems involving area, surface area, and volume.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Develop understanding of statistical variability.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>
Summarize and describe distributions.	<span style="color: lightblue;">▬</span>	<span style="color: lightblue;">▬</span>

*7.1.2.5 Student Detail Page*

When a student completes a test and the test is handscored, an online score report appears in the student detail page in the ORS. The student detail page provides individual student performance on the test. In each subject area, the student detail page provides (1) scale score and standard error of measurement (SEM), (2)

achievement level for overall test, (3) achievement category in each claim, (4) average scale scores for student’s district, and school.

On the top of the page, the student’s name, scale score with SEM, and achievement level are presented. On the left 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 was administered several times. Further, in the barrel chart, achievement-level descriptors with cut scores at each achievement level are provided, which defines the content area knowledge, skills, and processes that test-takers at each achievement level are expected to possess. On the right middle section, average scale scores and standard errors of the average scale scores for district, and school are displayed so that the student 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 reporting category is displayed along with a description of his/her performance on each reporting category. Exhibits 8 and 9 present examples of student detail pages for ELA/L and mathematics.

Exhibit 8. Student Detail Page for ELA/L

**Individual Student Report**

How did my student perform on the ELA/Literacy test?

Test: Smarter Summative ELA/Literacy Grade 6

Year: 2016-2017

Name: Demo, StudentA

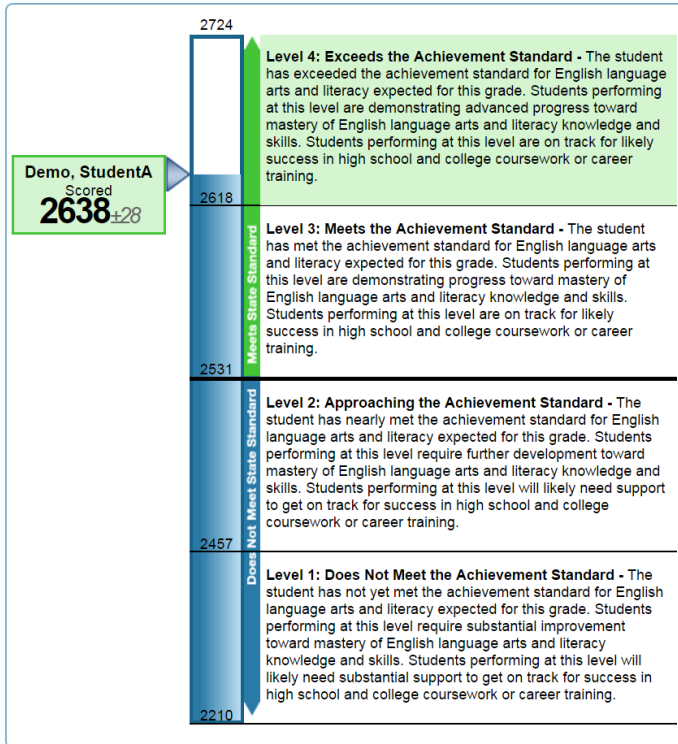
Legend: Claim Achievement Category

Below Standard Approaching Standard Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Demo, StudentA	999999999	2638 ±28	Level 4

Scale Score and Overall Performance



Comparison Scores

Name	Average Scale Score
Demo District (999)	2569 ±11
Demo School (999)	2569 ±11

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.

The table and the graph below indicate student performance on individual claims. The black line indicates the student's score on each claim. The green rectangle shows the range of likely scores your student would receive if he or she took the test multiple times.

Student Performance on Claims

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

Exhibit 9. Student Detail Page for Mathematics

**Individual Student Report**

How did my student perform on the Mathematics test?

Test: Smarter Summative Mathematics Grade 6

Year: 2016-2017

Name: Demo, StudentA

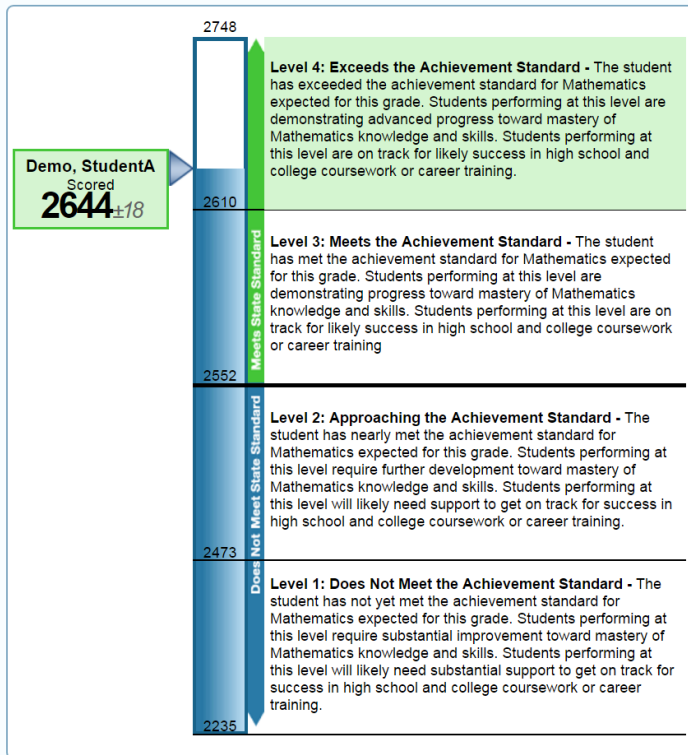
Legend: Claim Achievement Category

Below Standard Approaching Standard Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Demo, StudentA	999999999	2644 ±18	Level 4

Scale Score and Overall Performance



Comparison Scores

Name	Average Scale Score
Demo District (999)	2539 ±14
Demo School (999)	2539 ±14

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.

The table and the graph below indicate student performance on individual claims. The black line indicates the student's score on each claim. The green rectangle shows the range of likely scores your student would receive if he or she took the test multiple times.

Student Performance on Claims

Claim	Claim Performance	Claim Description
Concepts and Procedures	Above Standard	Student can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Problem Solving and Modeling & Data Analysis	Above Standard	Student can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. Student can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
Communicating Reasoning	Above Standard	Student can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

### 7.1.2.6 Participation Rate

In addition to online score reports, the ORS provides participation rate reports for districts and schools to help monitor the student participation rate. Participation data are updated each time students complete tests and these tests are handscored. Included in the participation table are (1) number and percent of students who are tested and not tested and (2) percent of students with achievement levels of 3 or above.

Exhibit 10 presents a sample participation rate report at a district level.

Exhibit 10. Participation Rate Report at District Level

#### Summary Statistics

**Step 1: Choose What**

Test:

Administration:

Test Name:

**Step 2: Choose Who**

District:

**Generate Report**

#### ELA Grade 6 Statistics of Students in Demo District

Smarter Summative: 2016-2017

**Legend**

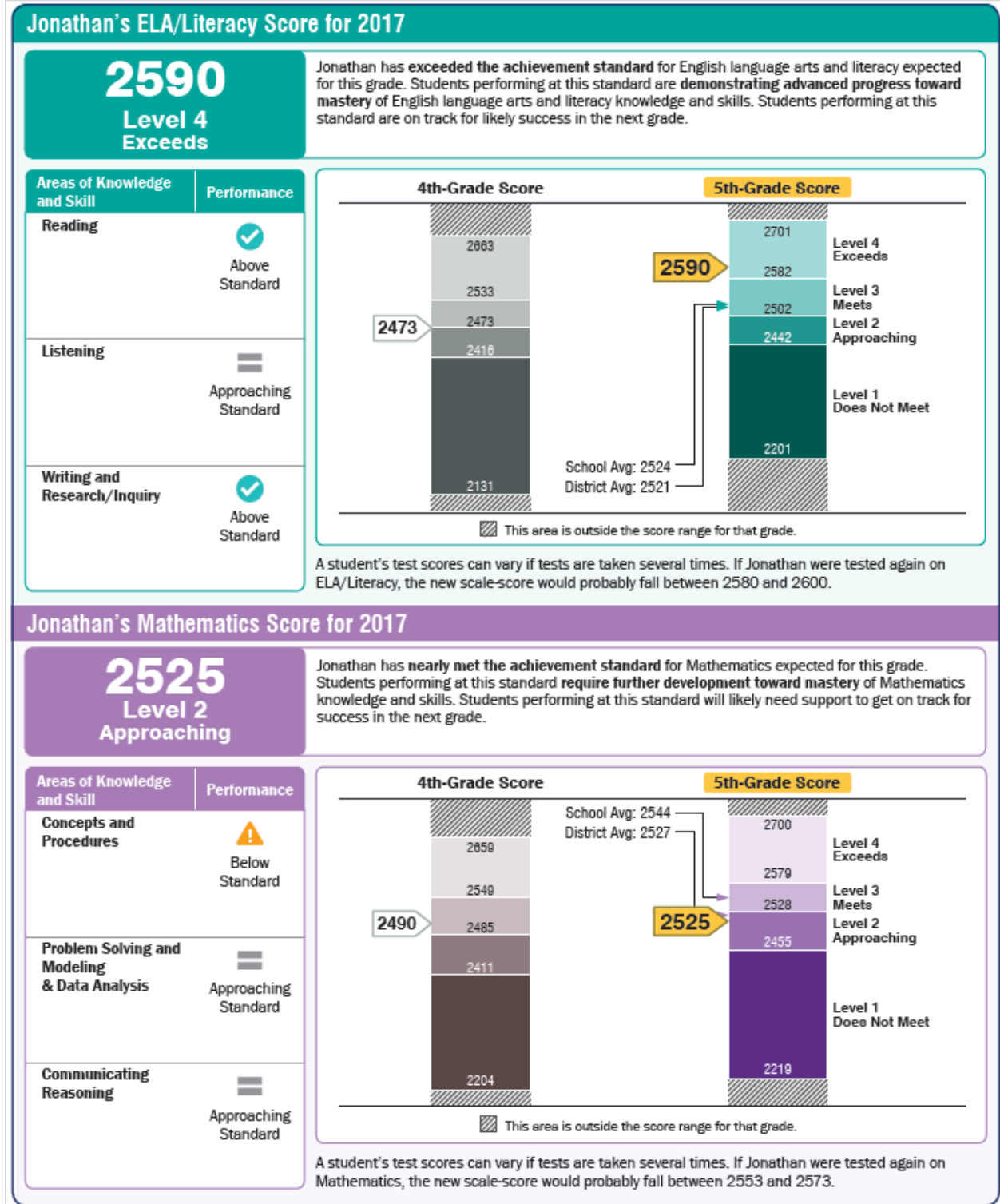
0 - not tested   1 - tested   **bold** - %   [ ] - count

Name	% Tested at each Opportunity & Count			% at Level 3 or Above by Opportunity
<b>Demo District (999)</b>	0	<b>99%</b>	[92]	N/A
	1	<b>1%</b>	[1]	0
<b>Demo School (999)</b>	0	<b>95%</b>	[18]	N/A
	1	<b>5%</b>	[1]	0
<b>Demo School (998)</b>	0	<b>99%</b>	[73]	N/A
	1	<b>1%</b>	[1]	0

## 7.2 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 ORS with additional guidance on how to interpret student achievement results in the family report. An example of a family report is shown in Exhibit 11.

Exhibit 11. Sample Paper Family Score Report



## **7.3 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 category for each claim. Students’ scores and achievement levels are also summarized at the aggregate levels. The next section describes how to interpret these scores.

### **7.3.1 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 from 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 sufficient 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.

### **7.3.2 Standard Error of Measurement**

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test several times, the resulting scale score would 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 several times. When interpreting scale scores, it is recommended to consider the range of scale scores incorporating the SEM of the scale score.

The “±” sign 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,  $2680 \pm 10$  indicates that if a student was tested again, it is likely that the student 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.

### **7.3.3 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 (i.e., Level 1, Level 2, Level 3, or 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 Level 3 in grade 6 ELA/L, for instance, achievement-level descriptors are described as “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 high school and college coursework or career training.” Generally, students performing at Levels 3 and 4 on Smarter Balanced assessments are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for college and career readiness.

### 7.3.4 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 of claims is evaluated with respect to the "Meets Standard" achievement standard. For students performing at either "Below Standard" or "Above Standard," this can be interpreted to mean that students' 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 students' performance does not provide enough information to tell whether students is clearly below or reached the "Meets Standard" mark for the specific claim.

### 7.3.5 Performance Category for Targets

In addition to the claim level reports, teachers and educators ask for additional reports on student performance for instructional needs. Target-level reports are produced for the aggregate units only, not for individual students, because each student is administered with too few items in a target to produce a reliable score for each target.

AIR reports relative strength and weakness scores for each target within a claim. The strengths and weaknesses report is generated for aggregate units of classroom, school, and district and provides information about how a group of students in a class, school, or district performed on the reporting target relative to their performance on the test as a whole. For each reporting element, we compare the observed performance on items within the reporting element with expected performance based on the overall ability estimate. At the aggregate level, when observed performance within a target is greater than expected performance, then the reporting unit (e.g., 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.

The performance on target shows how a group of students performed on each target relative to their overall subject performance on a test. The performance on target is mapped into three achievement levels: (1) better than performance on the test as a whole (higher than expected), (2) similar to performance on the test as a whole, and (3) worse than performance on the test as a whole (lower than expected). The "Worse than performance on the test as a whole" does not imply a lack of achievement. Instead, it can be interpreted to mean that student performance on that target was below their performance across all other targets put together. Although achievement categories for targets provide some evidence to help address students' strengths and weaknesses, they should not be over-interpreted because student performance on each target is based on relatively few items, especially for a small group.

### 7.3.6 Aggregated Score

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



## 7.4 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can be used to provide information on an individual student’s achievement on the test. Overall, assessment results tell what students know and are able to do in certain subject areas and further give information on whether students are on track to demonstrate 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 claims can be used to identify an individual student’s relative strengths and weaknesses among claims within a content area.

Assessment results on student achievement on the test can be used to help teachers or schools make decisions on how 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 performed very well overall, but it could be possible that they would not perform as well in several targets compared to their overall performance. In this case, teachers or schools can identify strengths and weaknesses of their students through the group performance by claim and target and promote instruction on specific claim or target areas that the group performance is below their overall performance. Further, by narrowing down the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from disadvantaged subgroups. For example, teachers can see student assessment results by LEP status and observe that LEP students are struggling with literary response and analysis in reading. Teachers can then provide additional instructions for these students to enhance their achievement of the benchmarks for literary response and analysis.

In addition, assessment results can be used to compare students’ performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts overall and by claim. Although all students are administered different sets of items in each computer adaptive test (CAT), scale scores are comparable across students. Furthermore, scale scores can be used to measure the growth of individual students over time if data are available. The scale score in the Smarter Balanced assessment is a vertical scale, which means scales are vertically linked across grades and scores across grades are on the same scale. Therefore, scale scores are comparable across grades so that scale scores from one grade can be compared with the next.

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 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 must 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 must 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. AIR implements a series of quality control steps to ensure error-free production of score reports in both online and paper 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 CAT, 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 specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (i.e., answer keys, item attributes, item parameters, and passage information). The accuracy of the information in the configuration file is checked and confirmed numerous times independently by multiple staff members before the testing window.

To verify the accuracy of the scoring engine, we use simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population (Smarter Balanced Assessment Consortium states). The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution. These simulations provide a rigorous test of the adaptive algorithm for adaptively administered tests and also provide a check of form distributions (if administering multiple test forms) and test scores in fixed-form tests.

Simulations are generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a wide range of student response patterns. The results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the Smarter Balanced summative assessments. The purpose of the simulations is to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability as well as checking the score accuracy.

After the adaptive test simulations, another set of simulations for the combined tests (adaptive test component plus a fixed-form performance task component) are performed to check scores. The simulated data are used to check whether the scoring specifications were applied accurately. The scores in the simulated data file are checked independently, following the scoring rules specified in the scoring specifications.

#### **8.1.1 Platform Review**

AIR's TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems like 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 in 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 Item Tracking System (ITS), and team members, each using a different platform, look at the same item to see 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 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 are scanned, a quality control sample of documents consisting of ten test cases per document type (normally between five and six hundred documents) was created so that all possible responses and all demographic grids were verified including various typical errors that required editing via MI's Data Inspection, Correction, and Entry (DICE) application program. This structured method of testing 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 results from the scanner, editing process (validation and data correction), and transfer to the AIR database are correct.

## **8.3 QUALITY ASSURANCE IN DATA PREPARATION**

AIR's TDS has a real-time quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to our Quality Assurance (QA) system. QA 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 operation items, and ensures that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitoring System (QMS) to the Database of Record (DoR), which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator (DEG) is the tool that is used to pull data from the DoR for delivery to the CSDE. AIR staff ensure that data in the extract files match the DoR before delivering to the CSDE.

## **8.4 QUALITY ASSURANCE IN HANDSCORING**

### **8.4.1 Double Scoring Rates, Agreement Rates, Validity Sets, and Ongoing Read-Behinds**

MI's scoring process is designed to employ a high level of quality control. All scoring activities are conducted anonymously; at no time do scorers have access to the demographic information of the students.

MI's Virtual Scoring Center (VSC) provides the infrastructure for extensive quality control procedures. Through the VSC platform, project leadership can: perform spot checks (read-behinds) of each scorer to evaluate scoring performance; provide feedback and respond to questions; deliver retraining and/or recalibration items on demand and at regularly scheduled intervals; and prevent scorers from scoring live responses in the event that they require additional monitoring.

Once scoring is underway, quality results are achieved by consistent monitoring of each scorer. The scoring director and team leaders read behind each scorer’s performance every day to ensure that he or she is on target, and they conduct one-on-one retraining sessions when necessary. MI’s QA procedures allow scoring staff to identify struggling scorers very early and begin retraining immediately.

If through read-behinds (or data monitoring) it becomes apparent that a scorer is experiencing difficulties, he or she is given interactive feedback and mentoring on the responses that have been scored incorrectly, and that scorer is expected to change the scores. Retraining is an ongoing process throughout the scoring effort to ensure more accurate scoring. Daily analyses of the scorer status reports alert management personnel to individual or group retraining needs.

In addition to using validity responses as a qualification threshold, other validity responses are presented throughout scoring as ongoing checks for quality. Validity responses can be culled from approved existing anchor or validity responses, but they also may be generated from live scoring and included in the pool following review and approval by the Smarter Balanced Assessment Consortium. MI periodically administers validity sets to each of MI’s scorers supporting the scoring effort. The VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whatever number of items is preferred by the state.

With the VSC program, the way in which the student responses are presented prevents scorers from having any knowledge about which responses are being single or double read, or which responses are validity set responses.

#### **8.4.2 Handscoring QA Monitoring Reports**

MI generates detailed scorer status reports for each scoring project using a comprehensive system for collecting and analyzing score data. The scores are validated and processed according to the specifications set out by Smarter Balanced. This allows MI to manage the quality of the scorers and take any corrective actions immediately. Updated real-time reports are available that show both daily and cumulative (project-to-date) data. These reports are available to Consortium states 24 hours a day via a secure website. Project leadership review these reports regularly. This mechanism allows project leadership to spot-check scores at any time and offer feedback to ensure that each scorer is on target.

#### **8.4.3 Monitoring by Connecticut State Department of Education**

The CSDE also directly observes MI activities, virtually. MI provides virtual access to the training activities through the online training interface. The CSDE monitors the scoring process through the Client Command Center (CCC) with access to view and run specific reports during the scoring process.

#### **8.4.4 Identifying, Evaluating, and Informing the State on Alert Responses**

MI implements a formal process for informing clients when student responses reflect a possibly dangerous situation for the test-takers. We also flag potential security breaches identified during scoring. For possible dangerous situations, scoring project management and staff employ a set of alert procedures to notify the client of responses indicating endangerment, abuse, or psychological and/or emotional difficulties.

This process is also used to notify each Consortium state of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he

or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow-up.

## **8.5 QUALITY ASSURANCE IN TEST SCORING**

To monitor the performance of the TDS during the test administration window, AIR 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 AIR contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our 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 us 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 are captured for each assessed student, such as data about how long it takes to load, view, or respond to an item. All of this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

A series of Quality Assurance 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 2.7.

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 of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and can be produced so that only items with statistics falling outside 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 blueprint and items are performing as anticipated.

Table 42 presents an overview of the QA reports.

Table 38. 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 unexpected 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

### 8.5.1 Score Report Quality Check

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

#### 8.5.1.1 Online Report Quality Assurance

Scores for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the IRT parameters), which can detect mis-keyed items, item score distribution, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are paired with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are checked by our quality assurance (QA) system. The integrated scores are sent to our test-scoring system, a mature, well-tested real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculating achievement-level indicators, subscale scores and other features, which then pass automatically to the reporting system and Database of Record (DoR). The scoring system is tested extensively before deployment, including hand checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

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. After scores have passed the QA checks and are uploaded to the DoR, they are passed to the ORS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system’s validation checks. All of the above processes take milliseconds to complete; within less than a second of handscores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

### *8.5.1.2 Paper Report Quality Assurance*

#### *Statistical Programming*

The family reports contain custom programming and require rigorous quality assurance processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in our 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 agreed-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

Much of the statistical processing is repeated, and AIR has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. We write small programs (called macros) that take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in our library for the grades 3–8 and 11 program 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 and the Director of Psychometrics, as well as by 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 do the many complicated calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. In addition, 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 VIPP and allows virtually infinite control of the visual appearance of the reports. After designers at AIR create backgrounds, our 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. AIR’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 run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables us to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the score reporting team to ensure that design elements are accurately reproduced and data are correctly displayed. Once we receive final data and VIPP programs, the AIR Score Reporting team reviews proofs that contain actual data based on our standard quality assurance documentation. In addition, we compare data independently calculated by AIR psychometricians with data on the reports. A large sample of reports is reviewed by several AIR staff members to make sure that all data are correctly placed on reports. This rigorous review typically is conducted over several days and takes place in a secure location in the AIR building. All reports containing actual data are stored in a locked storage area. Before printing the reports, AIR provides a live data file and individual student reports with sample districts for Department staff review. AIR works closely with the department to resolve questions and correct any problems. The reports are not 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.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 84–105.
- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* *20*, 37–46.
- 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).
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement*, *13*(4), 253–264.
- Linacre, J. M. (2011). *WINSTEPS Rasch-Model computer program*. Chicago: MESA Press.
- 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 Phillipine Statistician*, *52*(1–4), 81–92.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced. *Journal of Educational Measurement*, *13*(4), 265–276.



# APPENDICES

## Appendix A: Number of Students for Interim Assessments

The Interim Comprehensive Assessments (ICA) were fixed-form tests for each grade and subject. Most students took the ICA once, but some students took it more than twice. Table A–1 presents the number of students who took the ICA.

Table A–1. Number of Students Who Took ICAs

Grade	ELA/L				Mathematics			
	One	Two	Three	Total	One	Two	Three	Total
3	216	0	0	216	389	0	0	389
4	173	0	0	173	335	0	0	335
5	121	0	0	121	401	1	0	402
6	2	0	0	2	232	0	0	232
7	3	0	0	3	54	0	0	54
8	1	0	0	1	79	4	4	87

For the Interim Assessment Blocks (IAB), there were seven to nine IABs for ELA/L and five to six IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–2 presents the total number of students who took the IABs and the number of students by the number of IABs taken. For example, in grade 3 ELA/L, a total of 10,731 students took IABs, and among 10,731 students, 4,792 students took one IAB, 2,828 students took two IABs, and so on.

Tables A–3 and A–4 disaggregated the number of students in Table A–2 by each individual block. For example, 4,792 students in grade 3 ELA/L took one IAB only. Among 4,792 students, 38 of the students took the Brief Writes IAB.

Table A–2. Number of Students Who Took IABs

Grade	Total	Number of IABs Taken								
		1	2	3	4	5	6	7	8	9
<b>ELA/L</b>										
3	10,731	4,792	2,828	1,244	711	670	266	151	55	14
4	11,773	4,244	3,752	1,551	929	779	386	85	47	
5	11,128	4,662	3,550	1,795	481	259	221	160		
6	10,001	3,459	3,136	1,570	1,150	446	201	20	19	
7	11,185	3,276	3,793	2,528	922	441	131	94		
8	9,862	3,984	3,939	1,635	209	57	34	4		
11	2	1	1							
<b>Mathematics</b>										
3	15,580	5,361	4,374	3,784	1,964	97				
4	15,667	5,522	4,394	3,693	1,381	638	39			
5	14,353	5,642	3,930	2,962	946	863	10			
6	16,345	6,750	4,537	3,453	802	793	10			
7	16,667	6,972	4,355	3,986	707	647				
8	15,320	6,818	4,766	3,033	652	51				
11	16	16								

Table A–3: ELA/L Number of Students Who Took IABs by Block Labels (Grades 3–8)

Grade	Block	Number of IABs Taken								
		1	2	3	4	5	6	7	8	9
3	Brief Writes	38	61	216	187	129	24	53	55	14
	Editing	642	770	600	473	626	253	148	55	14
	Language and Vocabulary Use	1,517	555	388	357	423	158	150	55	14
	Listening and Interpretation	640	910	621	470	280	222	137	55	14
	Performance Task		2	38	4	1				14
	Reading Informational Text	1,151	1,509	562	414	586	250	151	55	14
	Reading Literary Text	567	1,381	644	379	488	244	137	55	14
	Research	117	204	232	234	475	202	133	55	14
	Revision	120	264	431	326	342	243	148	55	14
4	Brief Writes	26	85	164	222	99	88	51	47	
	Editing	833	636	707	595	651	296	85	47	
	Language and Vocabulary Use	584	1,628	714	371	608	161	79	47	
	Listening and Interpretation	611	740	727	661	383	370	45	47	
	Performance Task	4	59	4	54	13		1		
	Reading Informational Text	1,500	2,516	876	587	642	383	85	47	
	Reading Literary Text	441	1,426	775	493	585	360	82	47	
	Research	176	223	337	428	577	369	83	47	
	Revision	69	191	349	305	337	289	84	47	
5	Brief Writes	31	75	153	56	1				
	Editing	803	839	1,018	296	217	218	160		
	Language and Vocabulary Use	652	1,718	943	242	151	148	160		
	Listening and Interpretation	456	685	929	415	252	219	160		
	Performance Task	65	7	61	6					
	Reading Informational Text	1,418	2,140	651	287	160	202	160		
	Reading Literary Text	734	941	629	202	147	157	160		
	Research	88	234	380	110	152	190	160		
	Revision	415	461	621	310	215	192	160		
6	Brief Writes	6	105	156	114	6	2	19	19	
	Editing	898	829	1,100	1,000	418	201	20	19	
	Language and Vocabulary Use	414	1,634	850	552	376	200	20	19	
	Listening and Interpretation	370	264	505	578	402	199	20	19	
	Performance Task									
	Reading Informational Text	832	1,992	523	531	133	116	20	19	
	Reading Literary Text	646	810	456	437	155	93	2	19	
	Research	181	186	181	617	375	195	20	19	
	Revision	112	452	939	771	365	200	19	19	
7	Brief Writes		38	238	188	38	34	10		
	Editing	425	1,530	1,368	692	431	131	94		
	Language and Vocabulary Use	745	1,670	1,103	564	352	83	94		
	Listening and Interpretation	370	1,085	711	295	246	96	94		
	Performance Task									
	Reading Informational Text	553	1,675	1,174	421	251	71	84		
	Reading Literary Text	611	760	1,159	569	278	110	94		
	Research	420	382	1,097	406	273	130	94		
	Revision	152	446	734	553	336	131	94		

Grade	Block	Number of IABs Taken								
		1	2	3	4	5	6	7	8	9
8	Brief Writes	16	68	203	35	7	34	4		
	Editing and Revising	1,388	3,213	1017	171	57	34	4		
	Listening and Interpretation	396	1,442	574	173	56	34	4		
	Performance Task			8	7	1	3	4		
	Reading Informational Text	493	1,657	1,104	132	50	31	4		
	Reading Literary Text	1,064	546	1263	191	57	34	4		
	Research	627	952	736	127	57	34	4		
11	Brief Writes	1								
	Editing and Revising									
	Listening and Interpretation									
	Performance Task									
	Reading Informational Text		1							
	Reading Literary Text		1							
	Research									

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

Grade	Block	Number of IABs Taken					
		1	2	3	4	5	6
3	Measurement and Data	633	1,025	874	1,921	97	
	Number and Operations in Base Ten	1,728	2,327	3,517	1,950	97	
	Number and Operations – Fractions	1,277	2,300	3,504	1,939	97	
	Operational and Algebraic Thinking	1,680	2,923	3,424	1,963	97	
	Performance Task	43	173	33	83	97	
4	Measurement and Data	157	261	399	762	637	39
	Number and Operations in Base Ten	2,022	3,302	3,410	1,366	638	39
	Number and Operations – Fractions	1,393	2,227	3,280	1,334	637	39
	Operational and Algebraic Thinking	1,732	2,486	3,423	1,290	635	39
	Geometry	171	458	544	680	635	39
	Performance Task	47	54	23	92	8	39
5	Measurement and Data	238	666	2,094	774	862	10
	Number and Operations in Base Ten	2,264	2,966	2,715	882	862	10
	Number and Operations – Fractions	1,724	2,401	2,794	787	863	10
	Geometry	296	433	539	710	863	10
	Operations and Algebraic Thinking	1,114	1,330	698	622	852	10
	Performance Task	6	64	46	9	13	10
6	Expressions and Equations	1,385	1,013	2,874	662	790	10
	Geometry	138	878	365	541	791	10
	Number System	2,142	3,060	3,067	764	790	10
	Statistics and Probability	115	377	562	432	790	10
	Performance Task	82	73	206	54	11	10
	Ratios and Proportional Relationships	2,888	3,673	3,285	755	793	10
7	Expressions and Equations	2,564	2,167	3,467	586	647	
	Number System	2,616	3,392	3,818	683	647	
	Geometry	190	543	552	622	647	
	Statistics and Probability	210	61	151	255	646	
	Performance Task	167	14	146	15	1	
	Ratios and Proportional Relationships	1,225	2,533	3,824	667	647	
8	Expressions and Equations I	506	1,613	2,189	602	51	
	Expressions and Equations II	2,733	2,760	2,682	593	51	
	Functions	2,334	2,800	1,774	634	51	
	Geometry	1,229	2,262	2,295	568	51	
	Performance Task	16	97	159	211	51	
11	Algebra – Linear Functions	16					
	Algebra – Quadratic Functions						
	Geometry – Right Triangles and Trigonometric Ratios						
	Performance Task						
	Statistics and Probability						

## Appendix B: Percentage of Proficient Students in 2014–2015, 2015–2016, and 2016–2017 for All Students and by Subgroups

Table B–1. ELA/L Percentages of Proficient Students Across Years (Grades 3–5)

Group	2015–2016	2016–2017
<b>Grade 3</b>		
All Students	54	52
Female	58	56
Male	50	48
American Indian/Alaska Native	48	37
Asian	74	71
African American	31	30
Hispanic/Latino	33	31
Native Hawaiian/Pacific Islander	38	61
White	67	65
Multiple Ethnicities	57	55
LEP	16	18
IDEA Eligible	17	16
<b>Grade 4</b>		
All Students	56	54
Female	59	58
Male	52	50
American Indian/Alaska Native	42	47
Asian	74	76
African American	31	32
Hispanic/Latino	33	33
Native Hawaiian/Pacific Islander	55	43
White	70	67
Multiple Ethnicities	59	58
LEP	14	15
IDEA Eligible	17	17
<b>Grade 5</b>		
All Students	59	56
Female	64	61
Male	53	52
American Indian/Alaska Native	54	38
Asian	77	75
African American	33	31
Hispanic/Latino	37	34
Native Hawaiian/Pacific Islander	63	69
White	72	71
Multiple Ethnicities	62	62
LEP	13	9
IDEA Eligible	17	16

Table B–2. ELA/L Percentages of Proficient Students Across Years (Grades 6–8)

Group	2015–2016	2016–2017
<b>Grade 6</b>		
All Students	55	54
Female	60	59
Male	50	49
American Indian/Alaska Native	47	47
Asian	73	74
African American	31	31
Hispanic/Latino	31	31
Native Hawaiian/Pacific Islander	50	45
White	68	67
Multiple Ethnicities	56	57
LEP	6	5
IDEA Eligible	15	14
<b>Grade 7</b>		
All Students	55	55
Female	61	60
Male	50	50
American Indian/Alaska Native	43	46
Asian	77	74
African American	29	30
Hispanic/Latino	32	32
Native Hawaiian/Pacific Islander	56	59
White	67	68
Multiple Ethnicities	59	56
LEP	5	5
IDEA Eligible	15	15
<b>Grade 8</b>		
All Students	55	54
Female	62	60
Male	49	48
American Indian/Alaska Native	44	44
Asian	76	76
African American	32	30
Hispanic/Latino	33	32
Native Hawaiian/Pacific Islander	58	61
White	67	65
Multiple Ethnicities	59	57
LEP	4	3
IDEA Eligible	15	14

Table B–3. Mathematics Percentages of Proficient Students Across Years (Grades 3–5)

Group	2014–2015	2015–2016	2016–2017
<b>Grade 3</b>			
All Students	48	53	53
Female	47	52	53
Male	49	53	54
American Indian/Alaska Native	36	51	42
Asian	71	78	76
African American	21	27	29
Hispanic/Latino	24	31	33
Native Hawaiian/Pacific Islander	34	46	52
White	62	67	66
Multiple Ethnicities	49	56	58
LEP	11	20	24
IDEA Eligible	15	18	18
<b>Grade 4</b>			
All Students	44	48	50
Female	43	47	49
Male	45	49	51
American Indian/Alaska Native	34	36	43
Asian	70	73	77
African American	17	21	25
Hispanic/Latino	21	24	29
Native Hawaiian/Pacific Islander	46	55	46
White	57	62	64
Multiple Ethnicities	46	51	53
LEP	11	12	15
IDEA Eligible	11	13	15
<b>Grade 5</b>			
All Students	37	41	43
Female	35	40	42
Male	38	42	44
American Indian/Alaska Native	20	32	32
Asian	60	68	70
African American	11	14	16
Hispanic/Latino	15	18	21
Native Hawaiian/Pacific Islander	33	37	48
White	49	54	57
Multiple Ethnicities	35	43	46
LEP	5	6	7
IDEA Eligible	7	9	10



Table B–4. Mathematics Percentages of Proficient Students Across Years (Grades 6–8)

Group	2014-2015	2015–2016	2016–2017
<b>Grade 6</b>			
All Students	37	41	44
Female	37	41	44
Male	37	41	43
American Indian/Alaska Native	21	31	37
Asian	65	66	71
African American	12	14	18
Hispanic/Latino	15	17	20
Native Hawaiian/Pacific Islander	53	41	39
White	48	53	57
Multiple Ethnicities	39	40	45
LEP	4	4	5
IDEA Eligible	7	7	8
<b>Grade 7</b>			
All Students	39	42	43
Female	38	42	42
Male	39	42	43
American Indian/Alaska Native	18	29	27
Asian	68	71	70
African American	14	14	16
Hispanic/Latino	16	19	20
Native Hawaiian/Pacific Islander	32	44	48
White	50	54	56
Multiple Ethnicities	40	44	40
LEP	4	5	5
IDEA Eligible	7	9	9
<b>Grade 8</b>			
All Students	37	40	42
Female	38	42	43
Male	36	39	40
American Indian/Alaska Native	23	20	28
Asian	64	69	72
African American	12	15	15
Hispanic/Latino	15	17	19
Native Hawaiian/Pacific Islander	32	31	59
White	48	52	54
Multiple Ethnicities	35	43	43
LEP	4	3	4
IDEA Eligible	6	7	8

## Appendix C: Classification Accuracy and Consistency Index by Subgroups

Table C–1. ELA/L Classification Accuracy and Consistency by Achievement Levels (Grades 3–5)

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
<b>Grade 3</b>											
All Students	38,097	79	89	70	67	88	71	83	59	56	83
Female	18,506	79	88	70	67	88	71	81	59	56	83
Male	19,591	79	89	70	67	87	72	84	59	56	82
American Indian/Alaska Native	97	78	88	71	67	84	70	84	58	58	72
Asian	2,049	80	86	70	67	90	72	78	57	57	86
African American	4,841	80	90	70	67	85	72	85	59	56	75
Hispanic or Latino	9,847	80	90	70	66	85	72	86	60	55	77
Native Hawaiian/Pacific Islander	33	77	88*	70*	68	83	68	81*	57*	57	78
White	19,903	79	87	70	67	88	71	78	59	56	84
Multiple	1,327	80	89	71	67	89	73	82	60	56	85
LEP	4,011	82	91	70	67	81	75	87	60	54	69
IDEA	4,490	85	92	70	66	83	79	90	58	53	73
<b>Grade 4</b>											
All Students	39,228	77	89	61	63	87	69	83	49	53	81
Female	19,281	77	88	61	64	88		81	49	53	82
Male	19,947	77	90	61	63	87	70	84	49	53	80
American Indian/Alaska Native	86	74	91	62	65	77	65	80	53	51	73
Asian	2,109	80	86	61	63	90	73	77	48	52	87
African American	4,939	78	91	61	63	83	70	86	49	52	73
Hispanic or Latino	10,078	78	91	61	63	83	70	86	50	52	74
Native Hawaiian/Pacific Islander	42	79	87	65*	63*	91	70	83	50*	50*	83
White	20,623	76	87	61	64	88	68	77	49	53	83
Multiple	1,351	76	87	61	64	88	69	80	49	53	82
LEP	3,372	82	93	61	63	77	76	90	50	50	61
IDEA	5,006	83	93	61	64	82	78	91	48	51	70
<b>Grade 5</b>											
All Students	38,748	79	90	64	72	87	71	84	52	63	80
Female	19,028	79	89	64	72	87	71	82	52	64	81
Male	19,720	79	90	64	72	86	72	85	52	63	79
American Indian/Alaska Native	104	82	90	68	74	92	74	86	56	63	82
Asian	1,992	81	87	64	72	90	74	80	51	62	87
African American	5,019	80	91	64	72	84	73	87	53	63	70
Hispanic or Latino	9,580	79	91	64	73	83	72	87	53	63	71
Native Hawaiian/Pacific Islander	29	80	91*	67*	79	85*	70	81*	49*	71	80*
White	20,830	78	87	64	72	87	70	79	51	64	81
Multiple	1,194	79	88	64	73	87	71	82	51	65	80
LEP	2,779	86	93	65	71	71	80	91	53	56	53
IDEA	5,464	85	93	64	72	81	79	91	52	60	68

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

Table C–2. ELA/L Classification Accuracy and Consistency by Achievement Levels (Grades 6–8)

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
<b>Grade 6</b>											
All Students	39,180	78	88	68	73	85	69	81	57	65	76
Female	19,355	77	87	68	73	85	69	79	57	65	77
Male	19,825	78	89	68	73	84	70	82	57	65	75
American Indian/Alaska Native	105	77	85	69	75	85	68	78	57	66	77
Asian	1,980	79	86	68	73	88	71	74	57	65	82
African American	4,889	78	89	68	73	82	70	84	58	63	67
Hispanic or Latino	9,438	79	90	68	73	81	71	85	58	64	67
Native Hawaiian/Pacific Islander	44	77	89	65	70	84	70	82	55	59	80
White	21,699	77	85	68	73	85	68	75	56	65	77
Multiple	1,025	78	86	68	73	86	69	79	58	64	79
LEP	2,315	87	94	67	72	74	82	91	56	54	54
IDEA	5,415	84	92	68	73	82	77	89	57	61	67
<b>Grade 7</b>											
All Students	39,212	78	89	67	76	85	70	82	56	68	76
Female	19,056	78	88	67	76	85	70	80	55	69	76
Male	20,156	79	89	67	76	84	71	83	56	68	75
American Indian/Alaska Native	100	78	90	65	77	81	70	83	55	67	71
Asian	1,982	80	88	67	76	87	72	79	55	68	81
African American	4,933	79	90	67	76	81	72	85	57	66	66
Hispanic or Latino	8,956	80	91	67	75	83	72	86	57	67	68
Native Hawaiian/Pacific Islander	34	79	85*	68*	82*	78	72	84*	56*	66*	77
White	22,182	78	86	67	76	85	69	75	55	69	76
Multiple	1,025	78	86	66	77	85	69	79	54	69	76
LEP	2,110	88	94	66	74	75*	83	92	54	56	57*
IDEA	5,368	84	92	66	75	81	77	89	56	62	67
<b>Grade 8</b>											
All Students	40,139	79	88	70	77	83	70	81	59	69	74
Female	19,440	78	86	70	76	84	70	78	59	69	75
Male	20,699	79	89	70	77	83	71	83	59	69	73
American Indian/Alaska Native	108	76	90	69	74	74	68	81	60	66	63
Asian	1,973	80	86	70	77	87	72	77	58	69	80
African American	4,978	80	89	70	76	78	72	84	60	68	64
Hispanic or Latino	9,068	80	89	70	76	80	72	85	60	69	65
Native Hawaiian/Pacific Islander	41	78	86*	71	75	87*	70	74*	60	70	76*
White	22,921	78	85	70	77	83	69	75	59	70	74
Multiple	1,050	79	89	71	77	85	71	82	61	70	74
LEP	1,857	89	94	71	74	82*	85	92	57	55	53*
IDEA	5,358	84	91	70	75	81	77	88	59	64	68

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

Table C–3. Mathematics Classification Accuracy and Consistency by Achievement Levels (Grades 3–5)

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
<b>Grade 3</b>											
All Students	38,016	83	90	73	79	90	76	84	64	71	84
Female	18,464	82	89	73	79	89	75	83	63	71	83
Male	19,552	83	90	73	79	90	76	85	64	71	85
American Indian/Alaska Native	96	80	88	68	77	93	72	83	58	70	85
Asian	2,042	85	89	73	79	93	79	78	64	71	90
African American	4,826	83	91	74	78	87	77	87	64	69	80
Hispanic or Latino	9,817	83	91	73	79	87	76	86	64	70	79
Native Hawaiian/Pacific Islander	33	80	90*	75*	73	88*	73	86*	64*	64	79*
White	19,881	82	87	73	79	90	75	79	64	72	85
Multiple	1,321	83	90	73	79	90	77	84	63	72	87
LEP	4,005	83	92	73	78	85	77	87	64	69	75
IDEA	4,486	87	94	73	78	87	82	92	62	69	78
<b>Grade 4</b>											
All Students	39,162	84	90	80	79	90	77	84	73	71	85
Female	19,254	83	89	80	79	89	77	83	73	71	84
Male	19,908	84	90	80	79	90	78	84	73	71	86
American Indian/Alaska Native	86	85	94	86	75	86	78	86	81	68	78
Asian	2,106	87	88	80	79	93	81	77	72	70	91
African American	4,927	84	91	80	78	86	78	86	73	69	79
Hispanic or Latino	10,055	84	91	80	79	87	78	86	73	70	78
Native Hawaiian/Pacific Islander	41	82	97*	80	70	95*	76	85*	71	66	89*
White	20,598	83	87	80	79	90	77	78	73	72	85
Multiple	1,349	84	90	80	78	91	77	82	74	70	87
LEP	3,370	86	92	79	78	85	80	88	72	67	75
IDEA	4,999	88	93	79	79	87	83	91	72	68	80
<b>Grade 5</b>											
All Students	38,656	83	91	77	71	90	76	86	68	61	85
Female	18,990	83	90	77	72	89	76	85	69	61	84
Male	19,666	84	91	77	71	90	77	87	68	61	86
American Indian/Alaska Native	101	84	92	76	73	89	78	89	67	62	83
Asian	1,987	86	89	77	72	94	80	81	69	61	92
African American	4,994	85	92	77	71	85	80	89	67	59	76
Hispanic or Latino	9,545	84	92	77	71	86	78	88	68	61	78
Native Hawaiian/Pacific Islander	29	79	88*	75*	78*	79*	71	78*	68*	68*	73*
White	20,805	82	88	77	72	90	74	81	69	62	85
Multiple	1,195	83	89	78	71	89	76	85	69	60	86
LEP	2,770	88	94	76	71	85	83	92	65	58	75
IDEA	5,455	89	95	75	71	86	84	93	65	58	78

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

Table C–4. Mathematics Classification Accuracy and Consistency by Achievement Levels (Grades 6–8)

Group	N	%Accuracy					%Consistency				
		All	L1	L2	L3	L4	All	L1	L2	L3	L4
<b>Grade 6</b>											
All Students	39,031	83	92	77	72	90	76	87	70	62	84
Female	19,287	82	91	77	72	89	76	86	70	62	83
Male	19,744	84	92	78	72	90	77	88	70	63	85
American Indian/Alaska Native	103	83	90	80	75	87	77	87	72	64	79
Asian	1,976	85	89	76	73	93	80	81	69	62	91
African American	4,864	85	93	77	71	84	79	90	69	61	72
Hispanic or Latino	9,397	85	93	77	72	86	79	90	70	61	77
Native Hawaiian/Pacific Islander	44	87	93	75	77*	95	82	90	69	64*	91
White	21,627	82	90	78	72	90	74	82	70	63	84
Multiple	1,020	83	89	77	72	92	76	84	69	62	87
LEP	2,307	91	96	76	69	84	88	94	67	57	74
IDEA	5,392	90	96	76	72	89	86	94	68	60	77
<b>Grade 7</b>											
All Students	39,033	83	91	76	75	90	76	85	68	65	85
Female	18,969	83	90	76	75	89	76	84	68	65	84
Male	20,064	84	91	76	75	91	77	86	68	65	86
American Indian/Alaska Native	100	83	91	76	75	85	76	85	71	59	83
Asian	1,983	86	88	75	75	94	80	80	66	66	92
African American	4,906	85	92	75	73	85	79	89	66	62	76
Hispanic or Latino	8,883	85	92	75	74	88	79	89	67	64	80
Native Hawaiian/Pacific Islander	33	84	91*	75*	69*	91	79	85*	71*	59*	87
White	22,106	82	88	76	75	90	75	80	68	66	85
Multiple	1,022	83	89	76	75	92	76	84	68	65	87
LEP	2,091	91	95	74	72	91	87	94	62	58	85
IDEA	5,335	90	95	76	75	88	85	93	65	63	80
<b>Grade 8</b>											
All Students	39,955	82	91	72	72	91	76	86	62	62	86
Female	19,350	82	90	72	72	90	74	84	62	62	85
Male	20,605	83	91	72	72	91	77	87	61	62	87
American Indian/Alaska Native	109	82	91	70	74	93	75	84	61	67	82
Asian	1,970	85	88	71	72	94	80	80	61	63	92
African American	4,950	85	93	71	71	86	79	90	59	59	77
Hispanic or Latino	9,008	84	92	71	71	88	78	89	61	60	80
Native Hawaiian/Pacific Islander	41	82	90	70*	75	87	75	84	58*	66	84
White	22,831	81	88	72	72	90	73	80	63	62	86
Multiple	1,046	83	91	72	73	91	76	85	62	63	87
LEP	1,845	92	95	69	69	93	88	94	54	55	83
IDEA	5,300	90	95	71	72	88	86	94	59	58	80

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