Connecticut Smarter Balanced Summative Assessments

2017–2018 Technical Report

Addendum to the Smarter Balanced Technical Report

CSDE
CONNECTICUT STATE DEPARTMENT OF EDUCATION

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

LIST OF APPENDICES

1. OVERVIEW

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

The Connecticut State Board of Education formally adopted the CCSS in ELA/L and mathematics on July 7, 2010. All students in Connecticut, including students with significant cognitive disabilities who are eligible to take the Connecticut Alternate Assessment, 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 are composed of the end-of-year summative assessment designed for accountability purposes and the optional interim assessments designed to support teaching and learning throughout the year. The summative assessments are used to determine student achievement based on the CCSS and track student progress toward college and career readiness in ELA/L and mathematics. The summative assessments consist of two parts: a computer-adaptive test (CAT) and a performance task (PT).

- **Computer-Adaptive Test:** 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, none of which can 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.
- Reported scores for combining claim 2 (writing) and 4 (research/inquiry) in ELA/L.

Optional interim assessments allow teachers to check student progress throughout the year and provide information teachers can use to improve their instruction and learning. These tools are used at the discretion of schools and districts, and teachers can employ them to check students' progress in 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)** test the same content and report scores on the same scale as the summative assessments.
- **Interim Assessment Blocks (IABs)** focus on smaller sets of related concepts and provide more detailed information about student learning.

This report provides a technical summary of the 2017–2018 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 2017–2018 operational test administration; validity, reliability, scoring, reporting and interpreting scores; and the quality control procedures. The data included in this report are based on Connecticut data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs and a summary of their performance are provided in Appendix A.

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

Smarter Balanced produces a technical report 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. TEST ADMINISTRATION

2.1 TESTING WINDOWS

The 2017–2018 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. 2017–2018 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	03/26/2018	06/08/2018	Online Computer-Adaptive Tests
	3–8	03/26/2018	06/08/2018	Paper-Pencil Fixed-Form Tests
Interim Comprehensive Assessments	3–8, 11	09/26/2017	06/15/2018	Online Fixed-Form Tests
Interim Assessment Blocks	3–8, 11	09/26/2017	06/15/2018	Online Fixed-Form Tests

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 2017–2018 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2017–2018. A testing option is selected by content area. Once a testing option is selected, it applies to all tests in the content area.

Table 2. Summary of Tests and Testing Options in 2017–2018

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

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

To ensure standardized administration conditions, teachers (TEs) and test administrators (TAs) follow procedures outlined in the *Smarter Balanced ELA/L and Mathematics Online, Summative Test Administration Manual* (TAM). TEs and TAs must review the TAM prior to the beginning of testing to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures should be established for any students who are absent on 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 in the following subsections. More detailed descriptions can be found in the TAM provided online at this URL: http://ct.portal.airast.org/resources/.

District Administrator

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

District Test Coordinator

The District Test Coordinator (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 (including STCs, TEs, and TAs) into TIDE.
- Verifying all student information and eligibility in TIDE.
- Scheduling and administering training sessions for all STCs, TEs, TAs, and TCs.
- Ensuring that all personnel are trained on how to 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

The School Test Coordinator (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 the CSDE. STC responsibilities include 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

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

Test Administrator

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

TAs are responsible for the following:

- Completing Smarter Balanced test administration training.
- Reviewing all Smarter Balanced policy and test administration documents before administering any Smarter Balanced assessments.
- Viewing student information before testing to ensure that a student receives the proper test with the appropriate supports and reporting any potential data errors to STCs and DTCs, as appropriate.

- Administering the Smarter Balanced assessments.
- Reporting all potential test security incidents to the STCs and DTCs in a manner consistent with Smarter Balanced, CSDE, and district policies.

2.2.2 Online Test Administration

Within Connecticut's testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long 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 required to complete AIR's 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 TE or TA must first enter the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the assessment with the TE or TA must enter their State Student Identification Number (SSID), 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 settings. The TA or TE then reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the student(s) and guides them through the login process.

Once an assessment is started, the student must answer all of the 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.

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

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

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

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. Then the TAs or TEs must collect and send for secure shredding any handouts or scratch paper that students used during the assessment.

2.2.3 Paper-Pencil Test Administration

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

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

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

The total number of students who took paper-pencil tests is presented in Table 3.

Table 3. Number of Students Who Took Paper-Pencil Tests in the 2017–2018 Summative Test Administration

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

2.2.4 Braille Test Administration

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

embossed materials through the CSDE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described below:

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

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

2.3 TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

All DAs, DTCs, and STCs oversee all aspects of testing at their schools and serve as the main points of contact, and TEs and TAs administer the online assessments. The online AIR 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 required to complete AIR's 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.

Office Hour Webinars

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

Practice and Training Test Site

In January 2015, separate practice and training sites were opened for TEs/TAs and students, and these sites were refreshed before the 2016–2017 school year. In the fall of 2017, UEB braille forms were also offered for the practice and training tests. TEs and 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 levels of difficulty (approximately 30 items each in ELA/L and mathematics), as well as practice the performance task.

The training tests are designed to provide students and teachers with opportunities to quickly familiarize themselves with the software and navigational tools they will use for the upcoming Smarter Balanced assessments for ELA/L and mathematics. Training tests are available for both ELA/L and mathematics, and the tests are organized by grade bands (grades 3–5, grades 6–8, and grade 11), with each test containing 5–10 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 or TA in the TA Training Site. The student training test includes all item types in the operational item pool, including multiple-choice items, grid items, and natural language items. Teachers can also use these training tests to help students become familiar with the online platform and question types.

Manuals and User Guides

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

The *Test Coordinator Manual* provides information for DCs and STCs regarding policies and procedures for the 2018 Smarter Balanced assessments in ELA/L and mathematics.

The *Smarter Balanced Summative Assessment Test Administration Manual* provides information for TEs and TAs administering the Smarter Balanced online summative assessments in ELA/L and mathematics. It includes 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 the 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 *AIRWays User Guide* provides instructions and support for users viewing student interim assessment performance reports in AIRWays.

All manuals and user guides pertaining to the 2017–2018 online testing were available on the portal, and DAs, DTCs, and STCs used the manuals and user guides to train TAs and TEs in test administration policies and procedures.

Brochures and Quick Guides

The following brochures and quick guides are available on the 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-pencil tests.

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

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

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

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

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

User Role Permissions for Online Systems Brochure: This brochure outlines the user roles and permissions for each secure online testing system used to administer the online assessments for the Connecticut Comprehensive Assessment Program. These systems include: Test Information Distribution Engine (TIDE), Online Reporting System (ORS), Test Administration (TA) Interface, Data Entry Interface (DEI), Teacher Hand Scoring System (THSS), Assessment Viewing Application (AVA), and the AIRWays Reporting System.

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

2.3.2 District Test Coordinator Training Workshops

District Test Coordinator (DTC) training workshops were held on January 17−29, 2018, 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 secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal laws. Secure transmission and password-protected access are basic features of the current system and 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 students are accessing appropriate test content:

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

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

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

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

All students, including home-schooled students, must be enrolled or registered at their testing schools in order to take the online, paper-pencil, or braille assessments. Student enrollment information, including demographic data, is generated using a CSDE file and uploaded nightly via a 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 DA, DTC, STC, or TE can view their students' scores. TAs do not have access to student scores.

2.4.2 System Security

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

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

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

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 interruption are important factors to consider when selecting testing rooms.

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

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

Room Preparation

The room should be prepared prior to the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts that students might use to help answer test 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 in the testing room out of sight. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances in order to promote optimum testing conditions; they should also post "TESTING—DO NOT DISTURB" signs on the doors of testing rooms.

Seating Arrangements

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

After the Test

At the end of the test session, TEs or TAs must walk through the classroom to pick up any scratch paper that students used and any papers that display students' SSID numbers and names together. These materials should be securely shredded or stored in a locked area immediately. The printed reading passages and 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 on how to package and secure the test booklets to be returned to the testing contractor′s office are provided in the *Paper and Pencil Test Administration Manual*.

2.4.4 Test Security Violations

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

Impropriety: This is a test security incident that has a low impact on the individual or group of students who are testing and has a low risk of potentially affecting student performance on the test, test security, or test validity (for 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 (for 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 (for 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 Homeschooled 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

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 *Connecticut Assessment Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. They focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments of ELA/L and mathematics. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

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

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

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

For additional information about the availability of designated supports and accommodations, refer to the Connecticut's Assessment Guidelines for complete information at: https://ct.portal.airast.org/core/fileparse.php/51/urlt/2017-18_Assessment_Guidelines_LIVE_.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 pre-set in TIDE. In the 2017–2018 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

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

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

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

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

Mark 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.

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

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

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 PT(s) must be completed within 20 calendar days of the starting date.

Writing Tools. Selected writing tools (i.e., bold, italic, bullets, undo/redo) are available for all student-generated responses.

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

Non-Embedded Universal Tools

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

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

2.6.2 Designated Supports and Accommodations

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

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

Embedded Designated Supports

Color contrast: Students 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.

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

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

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, Filipino, Korean, Mandarin, Punjabi, Russian, Spanish, Ukrainian, and Vietnamese.

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

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

Non-Embedded Designated Supports

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

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

Color overlays: Color transparencies may be placed over a paper-pencil 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 the student to increase the size of test content 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 (for mathematics): Spanish text is read aloud to the student by a trained and qualified human reader who follows the administration guidelines provided in the *Smarter Balanced Test Administration Manual* and the read-aloud guidelines. All or portions of the content may be read aloud.

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

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

Embedded Accommodations

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

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

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 an adapted keyboard, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, 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.

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

Multiplication table (grade 4 and above mathematics tests): A paper-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 what they dictate verbatim. The scribe must be trained and qualified and must follow the administration guidelines provided in the *Smarter Balanced Online Summative Test Administration Manual*.

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 4 presents a list of universal tools, designated supports, and accommodations that were offered in the 2017–2018 administration. Tables 5–10 provide the number of students who were offered the accommodations and designated supports.

Table 4. 2017–2018 Universal Tools, Designated Supports, and Accommodations

	Universal Tools	Designated Supports	Accommodations
Embedded	Breaks Calculator[1] Digital Notepad English Dictionary[2] English Glossary Expandable Passages Global Notes Highlighter Keyboard Navigation Line Reader Mark for Review Mathematics Tools[3] Strikethrough Writing Tools[4] Zoom	Color Contrast Masking Mouse Pointer Print Size Text-to-Speech[5] Translated Test Directions[6] Translations (Glossary)[6] Translations (Stacked)[7] Turn off Any Universal Tools	American Sign Language[8] Braille Closed Captioning[9] Streamline Text-to-Speech[10]
Non-embedded	Breaks Scratch Paper	Amplification Color Contrast Color Overlay Magnification Noise Buffers Read Aloud[11] Read Aloud in Spanish[6] Scribe[12] Separate Setting Simplified Test Directions Translated Test Directions Translations (Glossary)[6]	100s Number Table[13] Abacus Alternate Response Options[14] Calculator[1] Multiplication Table[6] Paper Test (Large Print and Braille) Print-on-Demand Read Aloud[15] Scribe[16] Speech-to-Text

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

[1] For calculator-allowed items only in grades 6–8
[2] For ELA/L performance task full-writes
[3] Includes embedded ruler, embedded protractor
[4] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo
[5] For ELA/L PT stimuli, ELA/L PT and CAT items (not ELA/L CAT reading passages), and mathematics stimuli and items: must be set in TIDE before test begins
[6] For mathematics items
[7] For mathematics test
[8] For ELA/L listening items and mathematics items
[9] For ELA/L listening items
[10] For ELA/L reading passages; must be set in TIDE by state-level user
[11] For ELA/L items (not ELA/L reading passages) and mathematics items
[12] For ELA/L non-writing items and mathematics items
[13] For grade 4 and above mathematics tests
[14] Includes adapted keyboards, large keyboard, Sticky Keys, Mouse Keys, Filter Keys, adapted mouse, touch screen, head wand, and switches
[15] For ELA/L reading passages, all grades
[16] For ELA/L performance task writing items, all grades

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

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	3	4	10	5	4	8
Closed Captioning	19	23	33	31	28	21
Streamlined Mode	109	125	125	104	81	46
Text-to-Speech: Passages and Items	945	983	948	919	811	748
Non-Embedded Accommodations						
Alternate Response Options	5	5	9	7	6	4
Braille Paper Booklet - Uncontracted				1		
Speech-to-Text	130	124	126	99	67	53

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

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	8	13	18	32	20	37
	LEP	1	2	1	1		
	Special Ed	4	8	10	12	12	8
Masking	Overall	125	149	134	133	77	74
	LEP	29	33	30	11	10	17
	Special Ed	87	95	105	93	59	66
Mouse Pointer	Overall		1				
	LEP						
	Special Ed		1				
Print Size	Overall	109	121	142	119	121	114
	LEP	22	14	15	18	13	7
	Special Ed	79	109	113	105	95	93
Text-to-Speech: Items	Overall	5,437	5,223	4,720	3,401	2,967	2,455
	LEP	2,603	2,360	1,990	1,164	1,095	890
	Special Ed	2,019	2,195	2,181	2,041	1,675	1,347

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

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Color Contrast	Overall	4	2		3	1	4
	LEP	1	1				
	Special Ed	2			3	1	2
Color Overlay	Overall	5	10	6	4	6	5
	LEP						
	Special Ed	4	3	3	3	4	4
Magnification	Overall	7	4	6	8	6	4
	LEP						
	Special Ed	1	2	2	5	1	1
Noise Buffers	Overall	27	17	17	8	7	6
	LEP	8	5		1		
	Special Ed	11	11	15	6	1	1
Read-Aloud Items	Overall	182	157	125	63	66	66
	LEP	87	69	54	26	18	24
	Special Ed	109	93	85	52	56	61
Separate Setting	Overall	3,403	3,587	3,652	3,163	2,871	2,637
	LEP	760	730	701	451	446	403
	Special Ed	2,405	2,594	2,676	2,502	2,273	2,103
Simplified Test Directions	Overall	1,013	1,051	616	425	371	339
	LEP	336	326	266	190	214	186
	Special Ed	387	462	361	276	234	216
Translated Test Directions	Overall	159	143	132	278	267	238
	LEP	156	138	131	275	264	238
	Special Ed	21	20	23	38	25	25

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

Accommodations	Grade					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	2	4	10	5	4	8
Braille	1					
Streamlined Mode	12	8	12	7	2	6
Non-Embedded Accommodations						
100s Number Table	225	606	326	167	99	76
Abacus	4	9	3	1	4	
Alternate Response Options	5	5	8	6	5	3
Calculator	6	15	25	211	218	301
Multiplication Table		1,915	2,406	2,385	1,998	1,512
Speech-to-Text	116	113	118	90	57	49

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	7	12	18	32	20	38
	LEP		1	1	1		
	Special Ed	3	7	10	12	12	8
Masking	Overall	124	146	132	141	90	78
	LEP	29	34	29	19	17	17
	Special Ed	86	92	105	94	67	70
Mouse Pointer	Overall		1				
	LEP						
	Special Ed		1				
Print Size	Overall	108	119	143	120	120	111
	LEP	22	14	15	18	12	6
	Special Ed	77	107	114	106	94	89
Text-to-Speech: Stimuli and Items	Overall	6,991	6,847	6,347	5,107	4,412	3,801
	LEP	2,899	2,669	2,287	1,406	1,291	1,054
	Special Ed	3,167	3,399	3,435	3,306	2,759	2,344
Translation (Glossary): Spanish	Overall	686	677	554	620	618	611
	LEP	679	670	543	614	611	601
	Special Ed	63	70	54	65	73	85
Translation (Glossary): Other Languages	Overall	50	47	48	36	54	49
	LEP	50	45	48	35	54	49
	Special Ed	2	1	1		1	

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

Designated Supports	Subgroup	Grade					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	3	3		3	1	3
	LEP	1	1				
	Special Ed	1	1		3	1	1
Color Overlay	Overall	4	9	6	4	5	5
	LEP						
	Special Ed	3	3	3	3	4	4
Magnification	Overall	8	4	7	7	6	4
	LEP						
	Special Ed	2	2	3	4	1	1
Noise Buffers	Overall	27	16	17	10	8	5
	LEP	8	5		2	1	
	Special Ed	11	10	15	7	1	
Read Aloud Stimuli & Items	Overall	239	205	177	150	130	172
	LEP	51	45	37	29	16	19
	Special Ed	131	113	98	73	76	94
Read Aloud Stimuli & Items (Spanish)	Overall	38	41	39	27	17	34
	LEP	37	39	34	27	16	34
	Special Ed	6	9	13	10	8	10
Separate Setting	Overall	3,400	3,600	3,640	3,185	2,887	2,644
	LEP	758	731	697	451	445	393
	Special Ed	2,399	2,588	2,655	2,528	2,292	2,104
Simplified Test Directions	Overall	1,017	1,090	627	436	352	334
	LEP	318	335	258	170	181	168
	Special Ed	405	497	376	304	243	229
Translated Test Directions	Overall	104	104	92	217	195	182
	LEP	103	102	91	216	193	182
	Special Ed	9	14	15	23	19	16
Translation (Glossary): Spanish	Overall	49	49	57	56	79	81
	LEP	49	48	55	55	78	80
	Special Ed	8	10	10	12	6	8
Translation (Glossary): Other Languages	Overall	9	10	10	12	12	6
	LEP	8	10	10	11	11	6
	Special Ed					2	

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

Online test administration allows the collection of information that was impossible using 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 is 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 among administrations, testing times, 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 the student level and are summarized for each aggregate unit, including by 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 the past year and the 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., 2017–2018) and the year before (e.g., 2016–2017).

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 residuals of students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations and/or years based on the average 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|,

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

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, $var\big(E(\hat{e}_i|e_i)\big) = s^2$ and $E\big(var(\hat{e}_i|e_i)\big) = \sigma^2(1 - h_{ii})$. Following the law of total variance (Billingsley, 1995, page 456),

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

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

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 from one to five 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. For discrete items, each item appears on the screen one item at a time, whereas stimulus-based items appear on the screen together. The page time is the time spent on one item for discrete items and the time spent on all items associated with a stimulus for stimulus-based items. For each student, the total time taken to complete the test is computed by adding up the page time for all items and item groups (stimulus-based items).

The 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.

2.7.4 Inconsistent Item Response Pattern (Person Fit)

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

If a test taker has prior knowledge of some test items (or is provided answers during the 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, though the item response time index might flag such a student.

The person-fit index is based on all item responses in a test. An unlikely response to a single test 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, Pornel, and Vallejo (2003), aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items, *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 l_z for systematic flagging of aberrant response patterns. Students with l_z values greater than |3| are flagged. Aggregate units are flagged with *t* greater than |3|,

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

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

2.8 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 (TDS) 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 in the following subsections, is designed to recover from a 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 few hours prior or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them *before* a failure. On multiple occasions, this has enabled us to adjust 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 by text message our executive and technical staff, who then immediately join a call to understand the problem.

The following subsection 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 the 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 to prevent student data loss at every point in the process.

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

Student Machine

Student responses are conveyed to 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 serves as a backup hub for every four satellites. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon malfunction, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described 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 earlier. 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 a notification immediately 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 the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

2.8.3 Other Disruption Prevention and Recovery Systems

We have designed our system to be extremely fault-tolerant. The system can withstand failure of any component with little to 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 2017–2018 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All Connecticut 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 11–12 present the demographic composition of Connecticut students who meet attemptedness requirements for scoring and reporting of the Smarter Balanced summative assessments.

Table 11. Number of Students in Summative ELA/L Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	37,525	38,376	39,594	39,019	39,391	39,427
Female	18,417	18,646	19,454	19,152	19,421	19,178
Male	19,108	19,730	20,140	19,866	19,970	20,245
African American	4,764	4,854	5,034	5,034	4,895	4,932
American Indian/Alaskan	110	105	82	119	95	98
Asian	2,022	2,010	2,109	1,931	1,942	1,975
Hispanic/Latino	10,287	10,195	10,458	9,938	9,757	9,258
Pacific Islander	46	37	49	32	46	37
White	18,889	19,781	20,476	20,706	21,546	22,056
Two or More Races	1,407	1,394	1,386	1,259	1,110	1,071
LEP	4,153	3,776	3,186	2,502	2,410	2,112
Special Education	4,871	5,174	5,520	5,839	5,632	5,557

Table 12. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	37,472	38,307	39,540	38,946	39,265	39,294
Female	18,393	18,618	19,439	19,115	19,382	19,100
Male	19,079	19,689	20,101	19,830	19,883	20,190
African American	4,751	4,839	5,031	5,020	4,873	4,909
American Indian/Alaskan	110	104	82	118	95	98
Asian	2,024	2,007	2,107	1,929	1,939	1,975
Hispanic/Latino	10,270	10,178	10,442	9,918	9,719	9,209
Pacific Islander	46	37	49	32	46	37
White	18,866	19,747	20,449	20,674	21,486	21,997
Two or More Races	1,405	1,395	1,380	1,255	1,107	1,069
LEP	4,158	3,773	3,188	2,495	2,405	2,101
Special Education	4,865	5,169	5,511	5,832	5,607	5,527

3.2 SUMMARY OF STUDENT PERFORMANCE

Tables 13–16 present a summary of overall student performance in the 2017–2018 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 show the percentage of proficient students in four years for all students (cohort comparisons). Figures 3–4 show the average scale scores in four years for all students. The average and the standard deviation of scale scores, as well as the percentage of proficient students for each test administration, are provided in

Appendix B. In ELA/L, student performance is compared for three years because ELA/L scores in 2014–2015 were based on both CAT and PT components while ELA/L scores from 2015–2016 were based on CAT component only.

Table 13. ELA/L Percentage of Students in Achievement Levels 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
Grade 3								
All Students	37,525	2435	90	24	23	23	30	53
Female	18,417	2443	88	20	23	25	32	57
Male	19,108	2427	91	27	24	22	27	49
African American	4,764	2395	84	39	28	19	14	33
American Indian/Alaskan	110	2422	87	27	23	27	23	50
Asian	2,022	2479	85	10	16	25	48	73
Hispanic/Latino	10,287	2392	84	40	27	19	13	32
Pacific Islander	46	2438	85	17	37	22	24	46
White	18,889	2464	80	12	21	27	40	67
Two or More Races	1,407	2445	90	21	20	25	33	58
LEP	4,153	2360	76	54	28	13	5	18
Special Education	4,871	2355	78	60	24	11	6	16
Grade 4								
All Students	38,376	2479	97	27	18	23	32	55
Female	18,646	2488	95	24	18	24	35	59
Male	19,730	2470	99	30	18	23	29	52
African American	4,854	2431	90	46	20	19	15	34
American Indian/Alaskan	105	2451	85	30	29	26	15	41
Asian	2,010	2525	89	12	13	26	49	75
Hispanic/Latino	10,195	2432	93	44	21	20	15	35
Pacific Islander	37	2502	93	24	11	27	38	65
White	19,781	2509	87	15	16	26	42	68
Two or More Races	1,394	2490	100	25	17	22	36	59
LEP	3,776	2392	83	62	20	14	4	18
Special Education	5,174	2388	86	66	16	11	6	17
Grade 5								
All Students	39,594	2517	98	23	19	31	28	58
Female	19,454	2528	95	19	18	32	31	63
Male	20,140	2506	100	27	19	30	24	54
African American	5,034	2467	90	39	24	25	11	36
American Indian/Alaskan	82	2489	100	32	13	33	22	55
Asian	2,109	2571	90	9	12	29	50	79
Hispanic/Latino	10,458	2470	94	39	23	26	13	38
Pacific Islander	49	2495	101	31	27	20	22	43
White	20,476	2547	87	13	16	34	37	72
Two or More Races	1,386	2529	95	19	18	32	31	63
LEP	3,186	2410	79	65	22	11	2	13
Special Education	5,520	2423	86	62	20	14	5	18

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

Table 14. ELA/L Percentage of Students in Achievement Levels 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
Grade 6								
All Students	39,019	2534	101	23	23	33	22	54
Female	19,152	2546	97	19	22	34	25	59
Male	19,866	2522	103	27	23	31	19	50
African American	5,034	2484	92	40	28	24	8	32
American Indian/Alaskan	119	2498	99	36	28	24	12	36
Asian	1,931	2591	93	10	13	34	43	77
Hispanic/Latino	9,938	2482	95	40	27	25	7	32
Pacific Islander	32	2533	91	22	22	34	22	56
White	20,706	2565	89	12	20	38	30	68
Two or More Races	1,259	2542	99	22	20	33	25	58
LEP	2,502	2406	73	75	19	5	0	6
Special Education	5,839	2436	89	62	22	12	3	15
Grade 7								
All Students	39,391	2556	104	23	22	35	20	55
Female	19,421	2572	100	18	21	38	24	61
Male	19,970	2541	107	28	23	32	17	49
African American	4,895	2501	97	41	28	25	6	31
American Indian/Alaskan	95	2544	107	25	23	36	16	52
Asian	1,942	2612	94	9	15	36	40	76
Hispanic/Latino	9,757	2502	101	41	26	26	7	33
Pacific Islander	46	2560	122	30	11	30	28	59
White	21,546	2588	92	12	19	41	27	68
Two or More Races	1,110	2564	104	20	23	34	22	57
LEP	2,410	2421	79	77	18	5	0	5
Special Education	5,632	2454	92	62	23	13	2	15
Grade 8								
All Students	39,427	2575	103	21	23	36	20	56
Female	19,178	2591	99	16	22	39	24	62
Male	20,245	2560	104	26	24	34	16	50
African American	4,932	2522	95	37	30	26	7	33
American Indian/Alaskan	98	2546	96	26	37	23	14	38
Asian	1,975	2629	95	9	15	38	38	76
Hispanic/Latino	9,258	2522	98	38	28	27	7	34
Pacific Islander	37	2595	109	22	16	30	32	62
White	22,056	2605	92	11	20	42	27	69
Two or More Races	1,071	2581	102	18	26	35	22	56
LEP	2,112	2437	72	77	18	4	0	5
Special Education	5,557	2476	89	58	26	14	3	16

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 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
Grade 3								
All Students	37,472	2440	84	24	22	29	25	54
Female	18,393	2439	81	24	23	30	23	53
Male	19,079	2442	87	24	22	28	26	55
African American	4,751	2395	79	43	27	21	9	30
American Indian/Alaskan	110	2427	77	30	25	27	18	45
Asian	2,024	2496	77	8	13	29	50	79
Hispanic/Latino	10,270	2400	78	40	27	23	10	33
Pacific Islander	46	2441	72	20	30	28	22	50
White	18,866	2467	74	12	19	34	34	68
Two or More Races	1,405	2448	84	22	22	29	27	56
LEP	4,158	2380	77	51	25	18	6	24
Special Education	4,865	2361	83	61	20	13	5	19
Grade 4								
All Students	38,307	2484	85	20	29	28	23	51
Female	18,618	2482	80	19	30	29	21	50
Male	19,689	2485	90	21	27	27	25	52
African American	4,839	2434	79	38	35	19	8	26
American Indian/Alaskan	104	2462	80	25	33	32	11	42
Asian	2,007	2541	78	6	16	29	49	78
Hispanic/Latino	10,178	2443	79	35	35	20	10	30
Pacific Islander	37	2491	88	11	41	24	24	49
White	19,747	2511	75	9	25	34	31	65
Two or More Races	1,395	2491	87	17	30	26	28	53
LEP	3,773	2418	76	47	34	14	4	19
Special Education	5,169	2402	82	56	28	11	4	16
Grade 5								
All Students	39,540	2510	92	28	27	20	25	45
Female	19,439	2510	89	28	28	21	24	44
Male	20,101	2510	96	29	25	20	26	46
African American	5,031	2453	82	53	28	12	7	19
American Indian/Alaskan	82	2488	78	34	37	16	13	29
Asian	2,107	2577	85	9	16	20	54	74
Hispanic/Latino	10,442	2466	85	46	30	14	10	24
Pacific Islander	49	2475	99	37	31	22	10	33
White	20,449	2539	82	16	26	25	34	59
Two or More Races	1,380	2520	90	24	27	21	27	48
LEP	3,188	2425	77	66	24	6	3	9
Special Education	5,511	2422	82	69	20	7	4	12

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

Table 16. Mathematics Percentage of Students in Achievement Levels 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
Grade 6								
All Students	38,946	2527	107	28	28	21	22	44
Female	19,115	2531	102	26	29	23	22	45
Male	19,830	2523	112	30	27	20	23	43
African American	5,020	2464	100	51	30	13	7	19
American Indian/Alaskan	118	2495	107	45	24	16	15	31
Asian	1,929	2608	100	9	18	21	52	73
Hispanic/Latino	9,918	2472	101	47	30	15	7	22
Pacific Islander	32	2532	93	22	31	34	13	47
White	20,674	2561	92	15	28	27	31	58
Two or More Races	1,255	2536	108	26	27	20	26	46
LEP	2,495	2407	88	77	18	4	1	5
Special Education	5,832	2415	100	71	20	6	3	9
Grade 7								
All Students	39,265	2542	113	30	26	22	22	44
Female	19,382	2546	110	28	27	22	22	45
Male	19,883	2539	117	31	25	22	22	44
African American	4,873	2473	100	54	28	13	6	18
American Indian/Alaskan	95	2521	112	40	22	23	15	38
Asian	1,939	2628	106	10	16	21	52	73
Hispanic/Latino	9,719	2481	104	51	28	14	7	21
Pacific Islander	46	2550	141	30	30	9	30	39
White	21,486	2578	99	17	26	28	30	58
Two or More Races	1,107	2550	113	28	28	19	24	44
LEP	2,405	2417	91	79	16	4	2	5
Special Education	5,607	2427	99	73	17	7	3	9
Grade 8								
All Students	39,294	2558	120	33	24	19	24	43
Female	19,100	2564	115	30	25	21	24	44
Male	20,190	2553	125	36	23	18	24	42
African American	4,909	2483	105	59	23	11	6	18
AmerIndian/Alaskan	98	2518	107	46	31	11	12	23
Asian	1,975	2646	114	12	16	19	52	72
Hispanic/Latino	9,209	2493	108	55	25	12	8	20
Pacific Islander	37	2589	127	24	19	22	35	57
White	21,997	2595	107	20	24	24	31	56
Two or More Races	1,069	2563	118	33	24	20	23	43
LEP	2,101	2426	89	82	13	3	1	4
Special Education	5,527	2438	100	76	16	6	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

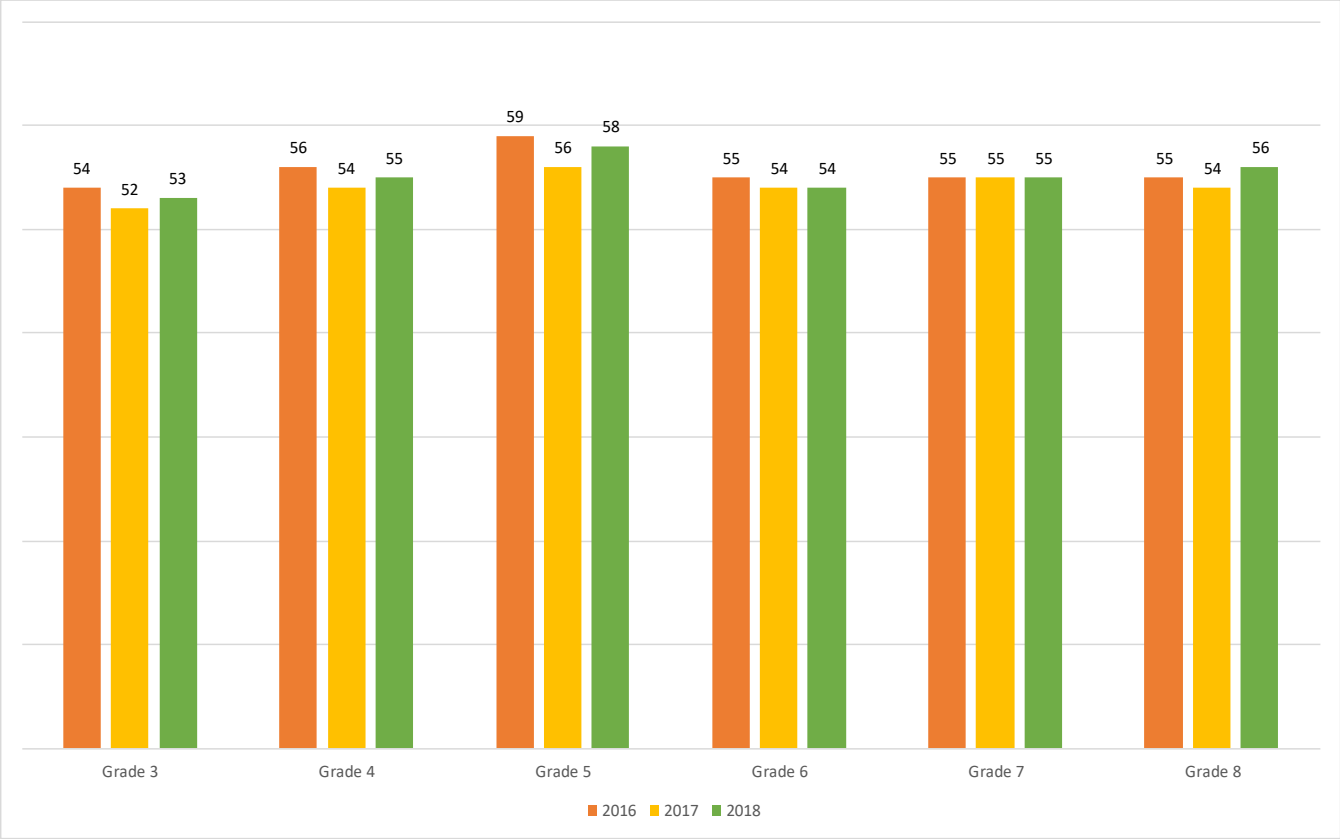


Figure 2. Mathematics %Proficient Across Years

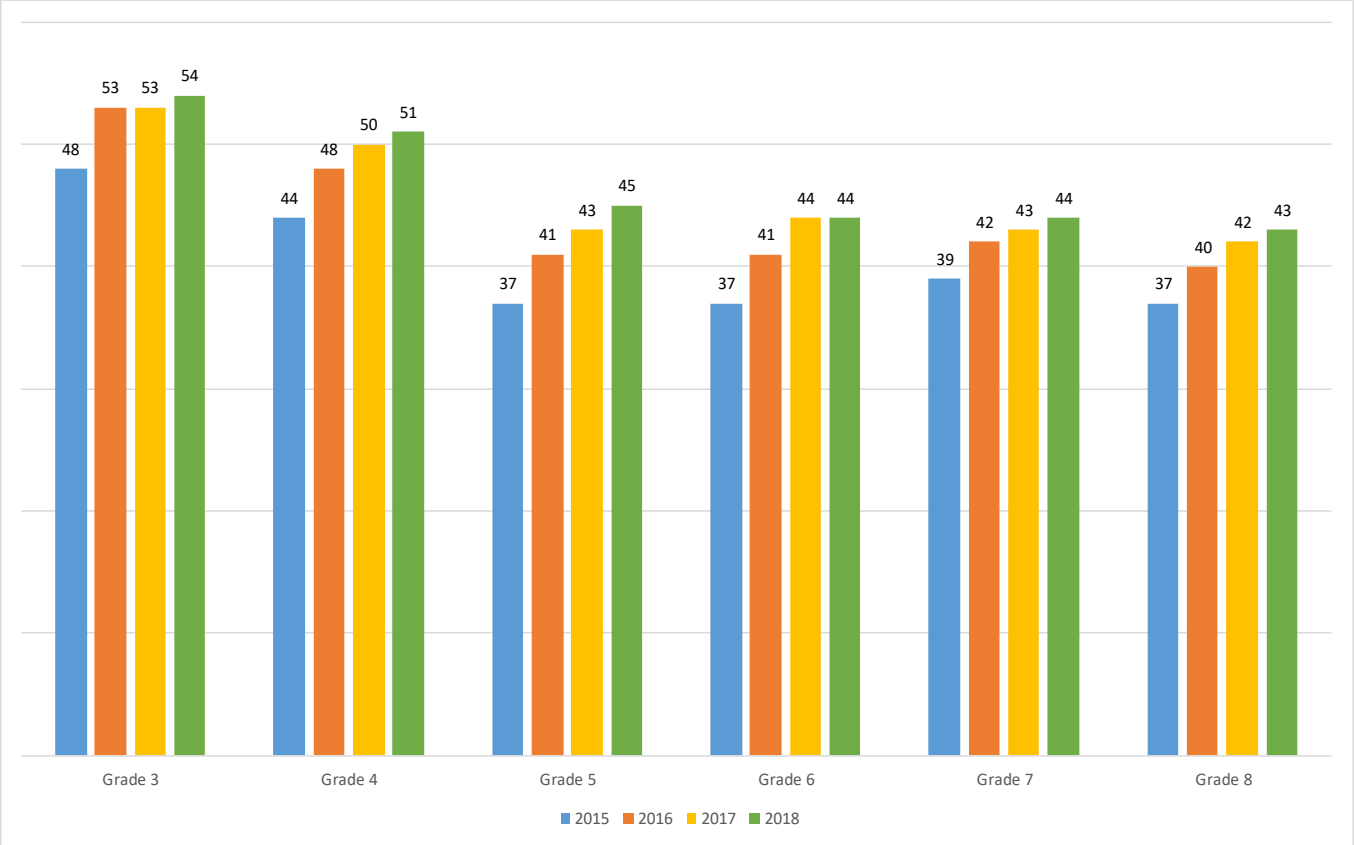


Figure 3. ELA/L Average Scale Score Across Years

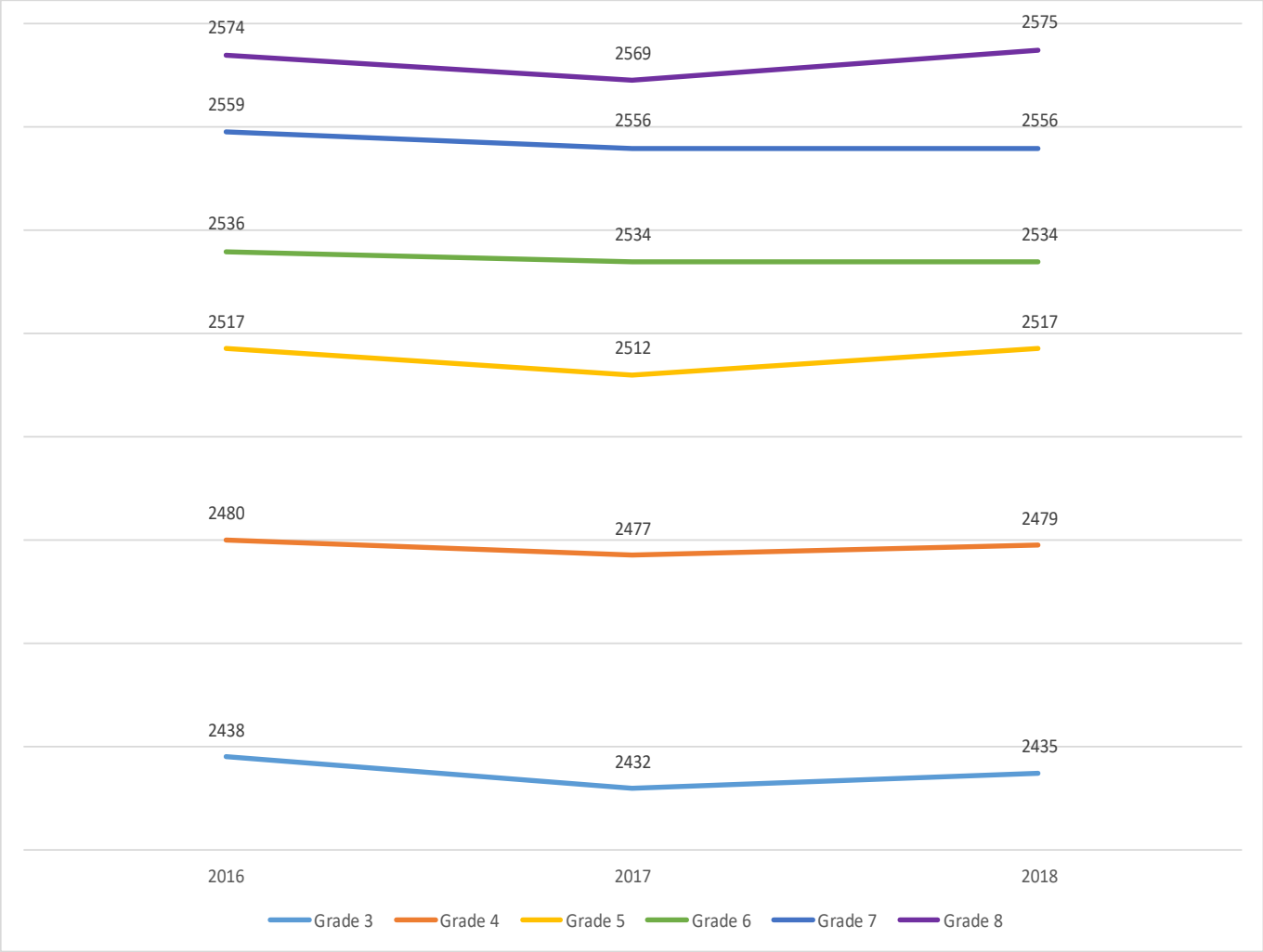
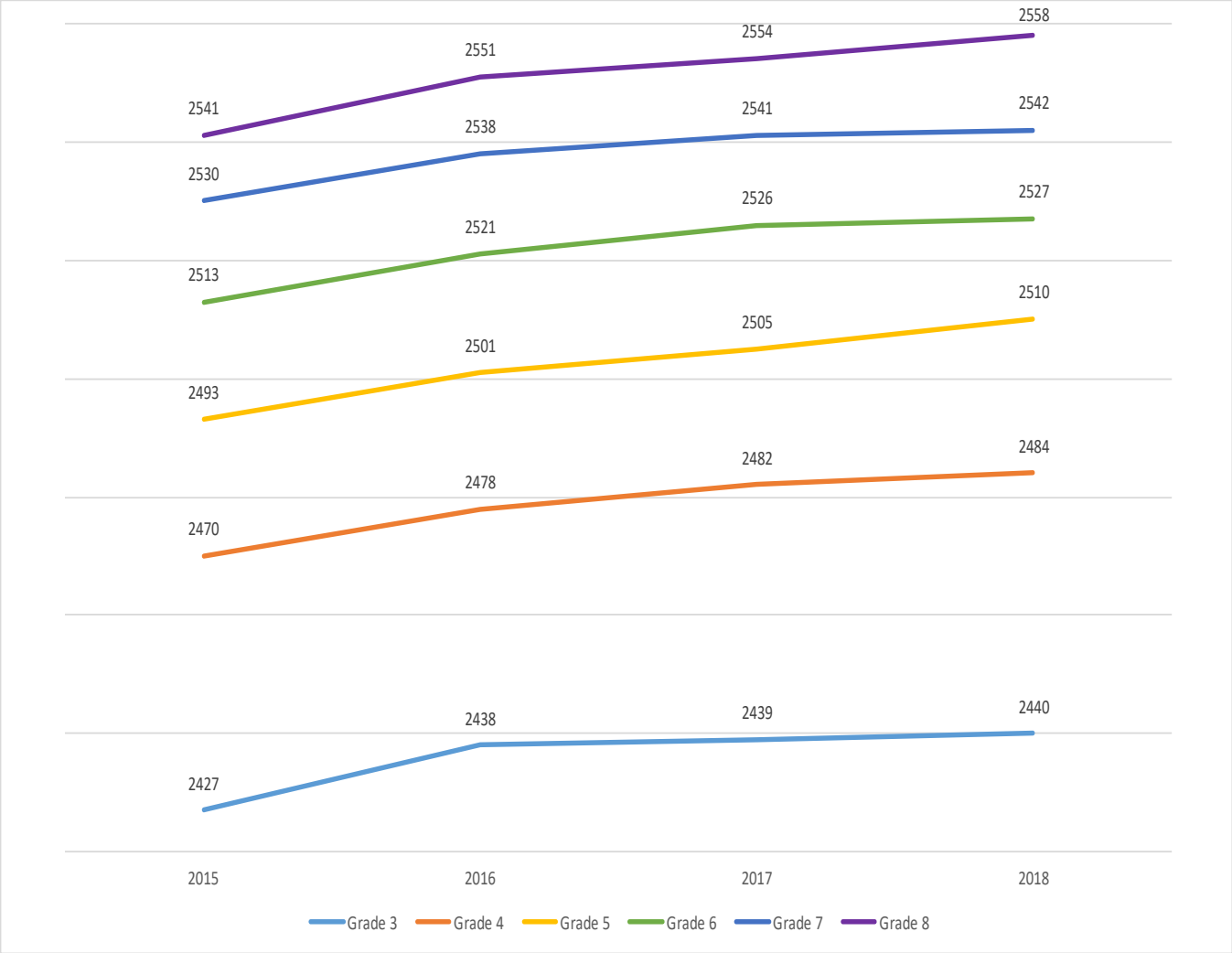


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three performance categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 17 and 18 present the distribution of performance categories for each claim. The number of claims is three in both ELA/L and mathematics, combining claims 2 and 4.

Table 17. ELA/L Percentage of Students in Performance Categories for Claims

Grade	Performance Category	Claim 1 Reading	Claims 2 & 4: Writing & Research	Claim 3 Listening
3	Below	24	30	15
	At/Near	45	42	61
	Above	30	28	24
4	Below	22	29	14
	At/Near	47	44	61
	Above	31	27	25
5	Below	21	25	16
	At/Near	46	42	61
	Above	33	33	23
6	Below	27	26	15
	At/Near	45	45	63
	Above	29	29	21
7	Below	25	24	18
	At/Near	44	47	66
	Above	31	29	17
8	Below	25	26	14
	At/Near	43	44	63
	Above	32	30	23

Table 18. Mathematics Percentage of Students in Performance Categories for Claims

Grade	Performance Category	Claim 1	Claims 2 & 4	Claim 3
3	Below	30	27	21
	At/Near	32	42	46
	Above	38	31	33
4	Below	32	27	25
	At/Near	33	45	44
	Above	35	27	31
5	Below	37	31	30
	At/Near	32	43	47
	Above	31	26	24
6	Below	37	33	33
	At/Near	35	44	43
	Above	29	23	24
7	Below	39	31	23
	At/Near	31	44	54
	Above	30	25	23
8	Below	37	24	29
	At/Near	34	49	48
	Above	29	27	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 testing 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 19 and 20 present an average testing time and the testing time at percentiles for the overall test, the CAT component, and the PT component.

Table 19. ELA/L Test-Taking Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test (CAT Component)							
3	1:44	0:54	2:01	2:10	2:20	2:36	3:05
4	1:47	0:53	2:05	2:13	2:23	2:39	3:08
5	1:47	0:46	2:06	2:14	2:24	2:39	3:05
6	1:45	0:46	2:04	2:12	2:23	2:38	3:05
7	1:35	0:40	1:53	2:00	2:09	2:22	2:45
8	1:31	0:39	1:48	1:54	2:03	2:16	2:40

Table 20. Mathematics Test-Taking Time

Grade	Average Testing Time (hh:mm)	SD of Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)				
			75th	80th	85th	90th	95th
Overall Test							
3	2:11	1:03	2:39	2:51	3:07	3:28	4:04
4	2:11	1:04	2:38	2:49	3:05	3:25	4:02
5	2:29	1:10	3:02	3:15	3:32	3:56	4:37
6	2:21	1:04	2:49	3:00	3:15	3:36	4:13
7	1:53	0:50	2:17	2:26	2:38	2:54	3:24
8	2:05	0:57	2:31	2:42	2:55	3:14	3:49
CAT Component							
3	1:28	0:45	1:47	1:56	2:06	2:22	2:49
4	1:31	0:48	1:50	1:59	2:10	2:25	2:53
5	1:31	0:43	1:52	2:00	2:10	2:24	2:48
6	1:33	0:42	1:51	1:58	2:08	2:23	2:47
7	1:23	0:37	1:41	1:48	1:57	2:09	2:31
8	1:29	0:42	1:48	1:56	2:06	2:20	2:46
PT Component							
3	0:43	0:25	0:54	0:59	1:05	1:14	1:29
4	0:40	0:23	0:50	0:55	1:00	1:08	1:21
5	0:58	0:36	1:12	1:19	1:28	1:41	2:03
6	0:48	0:30	1:00	1:05	1:12	1:21	1:40
7	0:30	0:19	0:38	0:42	0:46	0:53	1:04
8	0:35	0:21	0:45	0:49	0:54	1:01	1:14

3.4 DISTRIBUTION OF STUDENT ABILITY AND ITEM DIFFICULTY

Figures 5 and 6 display the empirical distribution of the Connecticut student scale scores in the 2017–2018 administration and the distribution of the administered summative item difficulty parameters in the operational pool. The student ability distribution is shifted to the left in all grades and subjects but is 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 intended to measure high

performing students accurately, but the pool 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 to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth-of-Knowledge [DoK], item type, and item difficulties) to better measure low performing students.

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

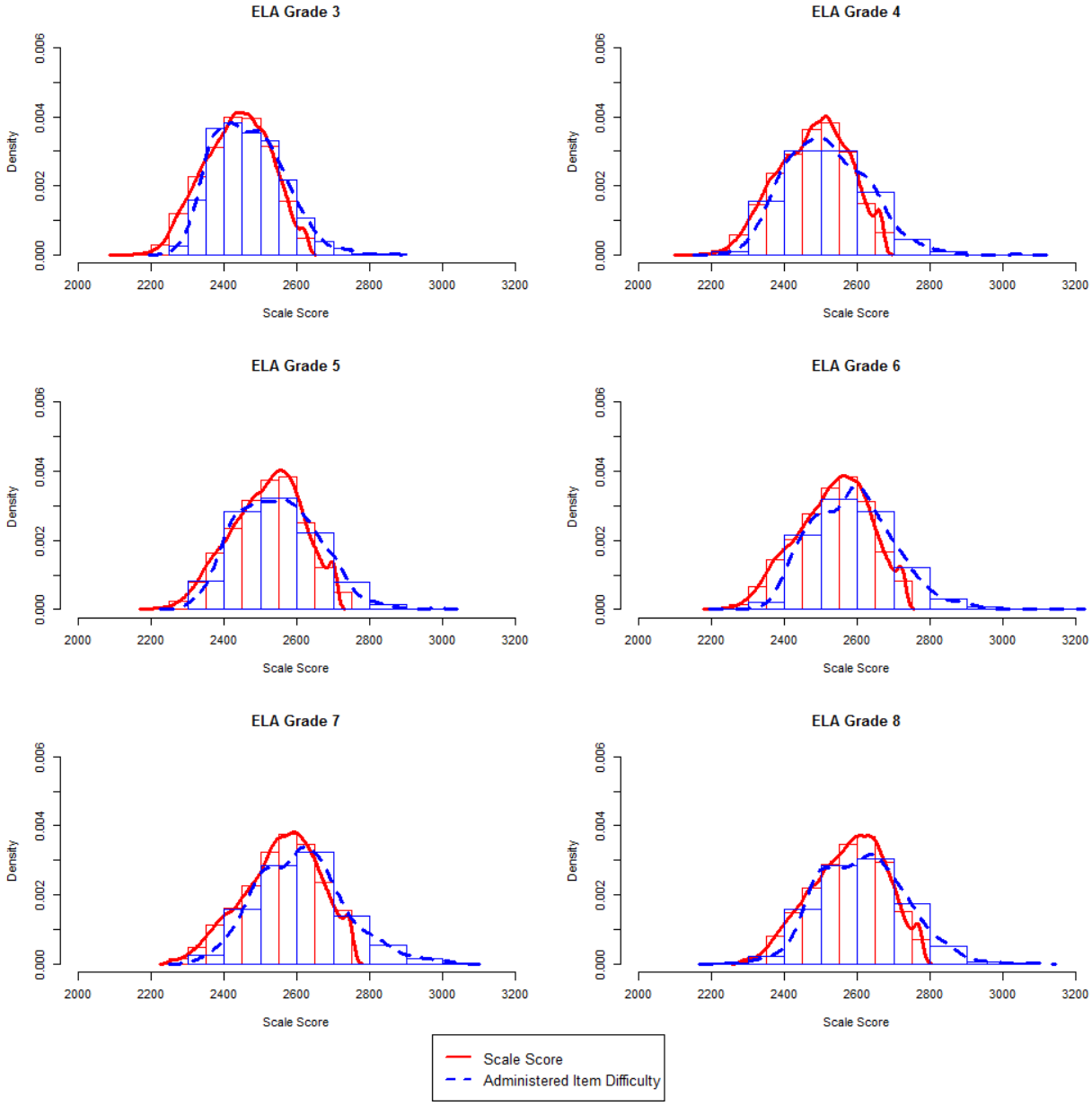
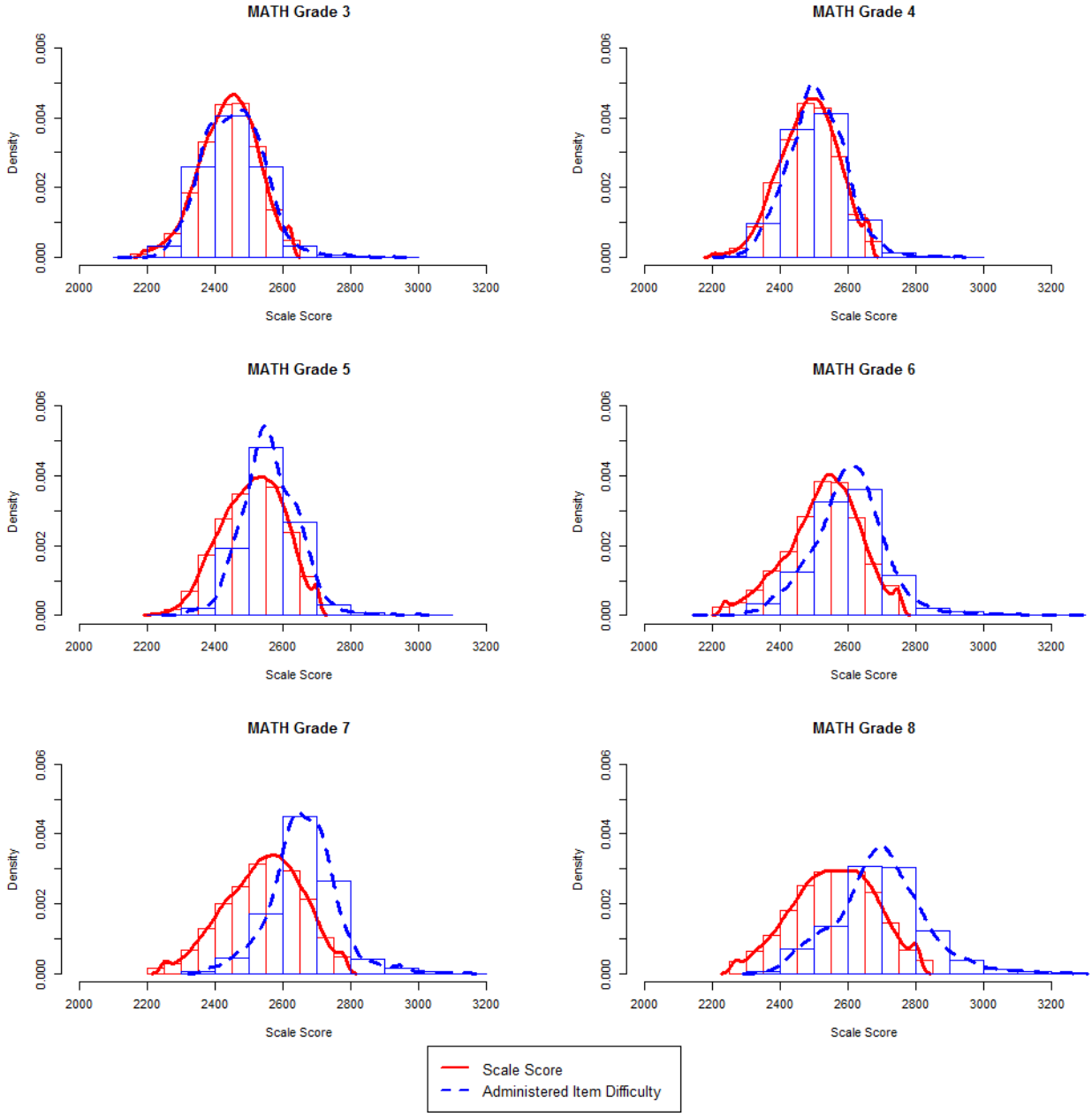


Figure 6. Student Ability–Item Difficulty Distribution for Mathematics



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 claim scores.

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

4.1 EVIDENCE ON TEST CONTENT

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

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints (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 21–22 present the percentages of tests aligned with the test blueprint constraints for ELA/L CAT. Table 21 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 22 presents the percentages of tests that satisfied the DOK and item type constraints for each claim. All tests met the requirements.

Tables 23–24 provide the percentages of tests aligned with the test blueprint constraints for the mathematics CAT, the blueprint match rates for claims, DOK, and target constraints. In mathematics, the tests met the blueprint requirements except for grades 3, 6, and 8. In mathematics grade 3, the violation was in claim 3 for target sets of A and D, which administered one item fewer than required. In mathematics grade 6, the violation was in the claim 1 for target sets of E and F and target sets of B and G, and claim 3 calculator segment for target sets of A and D; each administered fewer or more items than required. In mathematics

grade 8, the violation was in the claim 1 for target sets of C and D, and target sets of B, E, and G; each administered one item fewer or one item more than required.

Table 21. 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 Requirement	%BP Match for Passage Requirement
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; 3-L: Listening; 4-CR: Research

Table 22. ELA/L Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Item Type

DoK and Item Type Constraints	Required Items	%Blueprint Match					
		G3	G4	G5	G6	G7	G8
Claim 1 DOK2	≥ 7	100	100	100	100	100	100
Claim 1 DOK3 or higher	≥ 2	100	100	100	100	100	100
Claim 1 Short Answer in Target 2 or 4	0-1	100	100	100	100	100	100
Claim 1 Short Answer in Target 9 or 11	0-1	100	100	100	100	100	100
Claim 2 DOK2	≥ 4	100	100	100	100	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 23. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Target: Grades 3–5 Mathematics

Claim	Content Domain	Grade 3		Grade 4		Grade 5	
		Required Items	%BP Match	Required Items	%BP Match	Required Items	%BP Match
1	Overall	17-20	100	17-20	100	17-20	100
	DOK 2 or higher	≥ 7	100	≥ 7	100	≥ 7	100
	Priority Cluster	13-15	100				
	Targets B, C, G, I	5-6	100				
	Targets D, F	5-6	100				
	Target A	2-3	100				
	Supporting Cluster	4-5	100				
	Targets E. J, K	3-4	100				
	Target H	1	100				
	Priority Cluster			13-15	100		
	Targets A, E, F			8-9	100		
	Target G			2-3	100		
	Target D			1-2	100		
	Target H			1	100		
	Supporting Cluster			4-5	100		
	Targets I, K			2-3	100		
	Targets B, C, J			1	100		
	Target L			1	100		
	Priority Cluster					13-15	100
	Targets E, I					5-6	100
	Target F					4-5	100
	Targets C, D					3-4	100
	Supporting Cluster					4-5	100
	Targets J, K					2-3	100
	Targets A, B, G, H					2	100
2&4	Overall	6	100	6	100	6	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100

3	Overall	8	99	8	100	8	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	3	99	3	100	3	100
	Targets B, E	3	100	3	100	3	100
	Targets C, F	2	100	2	100	2	100

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

Claim	Content Domain	Grade 6		Grade 7		Grade 8	
		Required Items	**%BP Match**	**Required Items**	**%BP Match**	**Required Items**	**%BP Match**
1	Overall	16-20	100	16-20	100	16-20	100
	DOK 2 or higher	≥ 7	100	≥ 7	100	≥ 7	100
	Priority Cluster	12-15	100				
	Targets E, F	5-6	99				
	Target A	3-4	100				
	Targets G, B	2	99				
	Target D	2	100				
	Supporting Cluster	4-5	100				
	Targets C, H, I, J	4-5	100				
	Priority Cluster			12-15	100		
	Targets A, D			8-9	100		
	Targets B, C			5-6	100		
	Supporting Cluster			4-5	100		
	Targets E, F			2-3	100		
	Targets G, H, I			1-2	100		
	Priority Cluster					12-15	100
	Targets C, D					5-6	87
	Targets B, E, G					5-6	87
	Targets F, H					2-3	100
	Supporting Cluster					4-5	100
	Targets A, I, J					4-5	100
2&4	Overall	6	100	6	100	6	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3-Calc	Overall	7	100	8	100	8	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	3	99	3	100	3	100
	Targets B, E	2-3	100	3	100	3	100
	Targets C, F, G	2	100	2	100	2	100
3-No Calc	Overall	1	100				

Table 25 summarizes the target coverage by claim that includes the number of unique targets administered in each delivered test. The table includes the number of targets specified in the blueprints and the mean and range of the number of targets administered to students. Because the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered varies across tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level, across all tests combined.

Table 25. 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
ELA/L												
3	14	5	1	3	11	5	1	3	8-14	4-5	1-1	3-3
4	14	5	1	3	11	5	1	3	8-13	4-5	1-1	3-3
5	14	5	1	3	11	5	1	3	8-14	3-5	1-1	3-3
6	14	5	1	3	11	5	1	3	9-11	4-5	1-1	3-3
7	14	5	1	3	10	5	1	3	8-11	3-5	1-1	3-3
8	14	5	1	3	10	5	1	3	8-11	3-5	1-1	3-3
Mathematics												
3	11	4	6	6	11	2	6	3	8-11	2-2	3-6	3-4
4	12	4	6	6	10	2	5	3	9-11	2-2	3-6	3-3
5	11	4	6	6	9	2	5	3	9-9	2-2	3-6	3-4
6	10	4	7	6	10	2	5	3	8-10	1-2	3-7	2-3
7	9	3	7	6	8	2	5	3	8-8	2-2	3-6	3-3
8	10	4	7	6	10	2	5	3	10-10	2-2	3-6	2-4

An adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty). 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 claim measured. The evidence on the internal structure is examined based on the correlations among claim scores.

The correlations among claim scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 26 and 27. The correction for attenuation indicates what the correlation would be if claim scores could be measured with perfect reliability, corrected (adjusted) for measurement error estimates. The observed correlation between two claim scores with measurement errors can be corrected for attenuation as $r_{x\prime y\prime} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$, where $r_{x\prime y\prime}$ is the correlation between *x* and *y* corrected for

attenuation, r_{xy} is the observed correlation between *x* and *y*, r_{xx} is the reliability coefficient for *x*, and r_{yy} is the reliability coefficient for *y*.

When corrected for attenuation (above diagonal), the correlations among claim scores are higher than observed correlations. The disattenuated correlations are quite high. 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 reliability for claim scores are low, the performance of all the claim scores is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 17 and 18, Section 3.2. Scale scores are not reported for claims.

Table 26. Correlations among Claims for ELA/L

Grade	Claim	Observed and Disattenuated Correlation		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1: Reading		0.96	0.98
	Claims 2 & 4: Writing & Research	0.77		0.97
	Claim 3: Listening	0.65	0.67	
4	Claim 1: Reading		0.97	1
	Claims 2 & 4: Writing & Research	0.76		0.98
	Claim 3: Listening	0.65	0.66	
5	Claim 1: Reading		0.98	0.98
	Claims 2 & 4: Writing & Research	0.77		0.97
	Claim 3: Listening	0.65	0.68	
6	Claim 1: Reading		0.98	1
	Claims 2 & 4: Writing & Research	0.77		1
	Claim 3: Listening	0.67	0.69	
7	Claim 1: Reading		1	1
	Claims 2 & 4: Writing & Research	0.79		1
	Claim 3: Listening	0.65	0.66	
8	Claim 1: Reading		0.99	1
	Claims 2 & 4: Writing & Research	0.79		1
	Claim 3: Listening	0.68	0.69	

Table 27. Correlations among Claims for Mathematics

Grade	Claim	Observed and Disattenuated Correlation		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		0.97	0.96
	Claims 2 & 4	0.82		0.99
	Claim 3	0.80	0.76	
4	Claim 1		0.99	0.99
	Claims 2 & 4	0.81		1
	Claim 3	0.82	0.76	
5	Claim 1		1	0.98
	Claims 2 & 4	0.77		1
	Claim 3	0.78	0.72	
6	Claim 1		1	0.99
	Claims 2 & 4	0.84		1
	Claim 3	0.82	0.78	
7	Claim 1		1	1
	Claims 2 & 4	0.81		1
	Claim 3	0.78	0.73	
8	Claim 1		1	1
	Claims 2 & 4	0.77		1
	Claim 3	0.80	0.71	

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 σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

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

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

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

Table 28 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores.

Table 28. 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				
ELA/L							
3	37,525	38	40	0.91	2435	90	27
4	38,376	38	40	0.90	2479	97	30
5	39,594	38	41	0.91	2517	98	30
6	39,019	38	41	0.91	2534	101	31
7	39,391	38	41	0.90	2556	104	32
8	39,427	40	41	0.91	2575	103	31
Mathematics							
3	37,472	39	40	0.95	2440	84	19
4	38,307	37	40	0.95	2484	85	20
5	39,540	38	40	0.94	2510	92	23
6	38,946	39	39	0.94	2527	107	26
7	39,265	38	40	0.93	2542	113	29
8	39,294	38	39	0.93	2558	120	31

5.2 STANDARD ERROR CURVES

Figures 7 and 8 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 7. Conditional Standard Error of Measurement for ELA/L

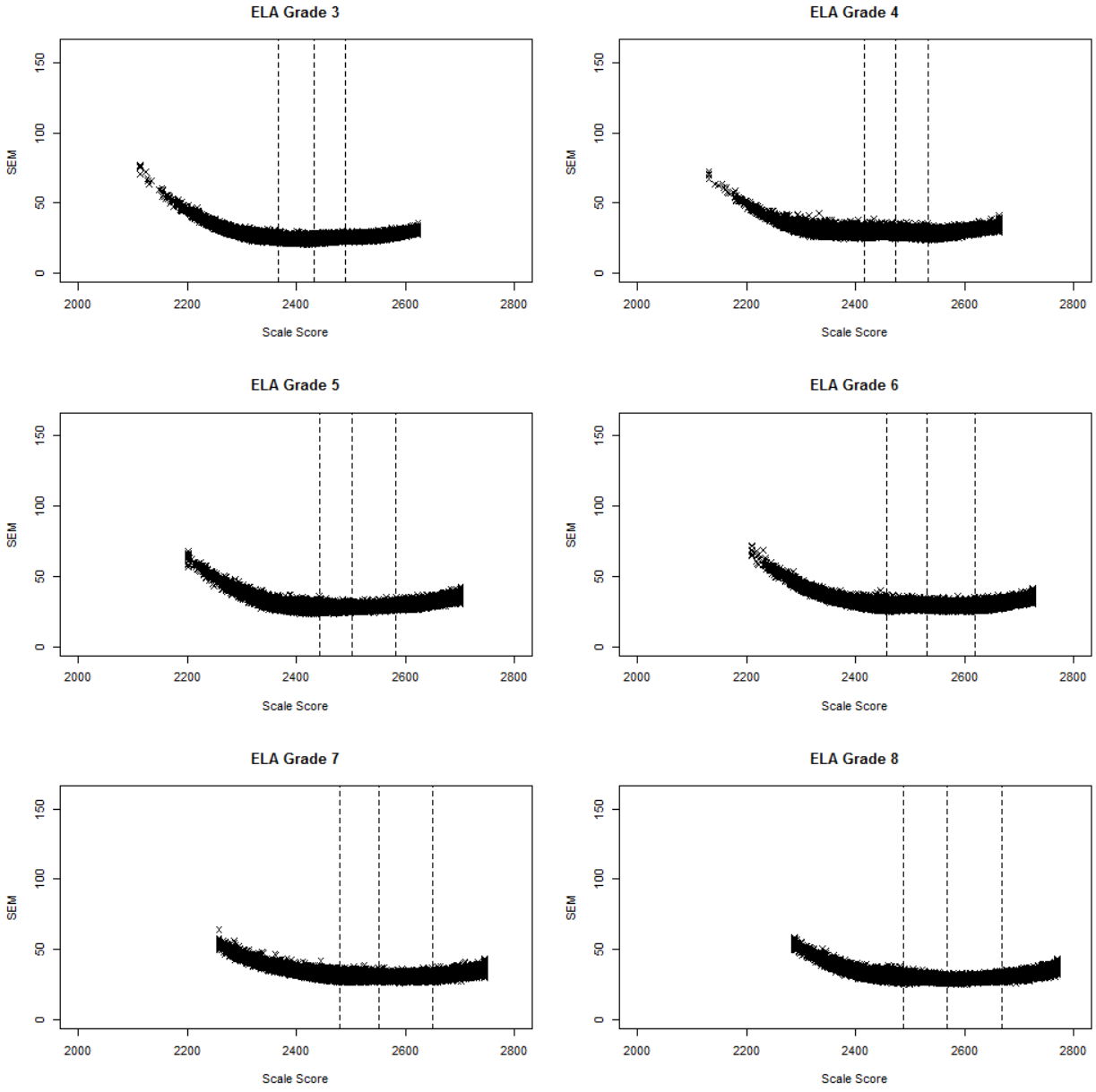
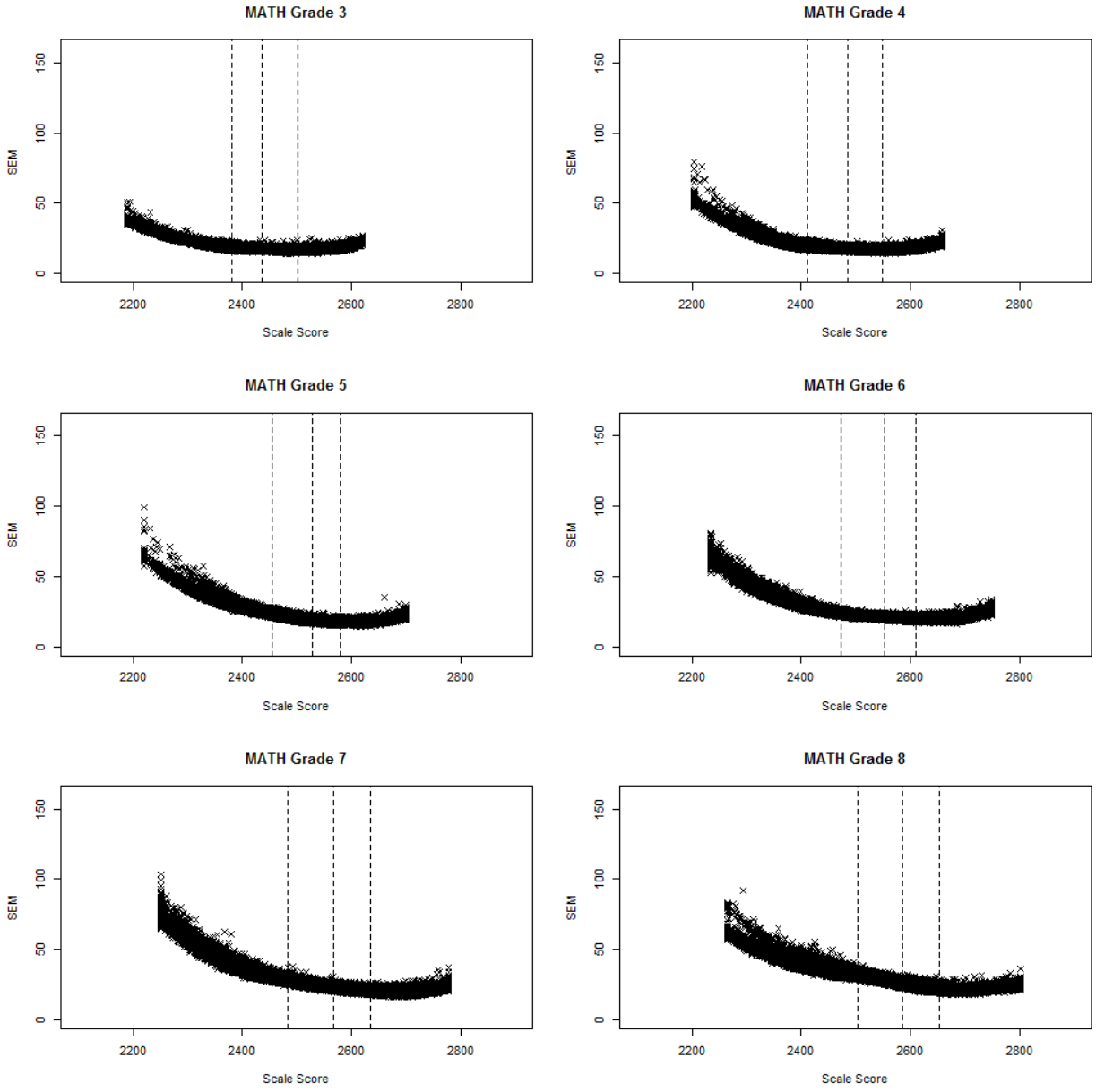


Figure 8. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures are summarized in Tables 29 and 30. Table 29 provides the average conditional SEM for all scores and scores in each achievement level. Table 30 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 7 and 8, 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 they are larger in Level 2 cut scores in mathematics.

Table 29. Average Conditional Standard Error of Measurement by Achievement Levels

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	30	24	25	27	27
4	32	29	29	30	30
5	31	27	28	31	30
6	34	29	29	31	31
7	37	30	30	33	32
8	35	29	29	33	31
Mathematics					
3	23	18	17	18	19
4	25	18	17	18	20
5	30	21	18	18	23
6	35	22	20	21	26
7	40	25	22	21	29
8	40	29	24	22	31

Table 30. 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|
ELA/L						
3	25	25	25	1	1	0
4	29	29	28	0	1	1
5	27	27	29	0	2	2
6	29	29	29	0	0	0
7	31	30	31	0	1	0
8	30	29	30	2	1	0
Mathematics						
3	19	17	17	1	1	2
4	20	17	17	2	1	3
5	24	19	18	5	1	6
6	24	21	20	3	1	4
7	28	23	20	5	3	7
8	32	26	22	6	4	11

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

$$p_{il} = p(c_{l-1} \le \theta_i < c_l) = p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \le \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) = p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \le \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right)$$
$$= \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right).$$

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, 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, a 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 ($l = 1,2,\cdots,L$) based on the cut scores cut_{l-1} and cut_l, given the student's item scores $\mathbf{z}_i = (z_{i1},\cdots,z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1,\cdots,\mathbf{b}_J)$ and using the *J* administered items, can be estimated as

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

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

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

where the likelihood function based on general IRT models is

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

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 $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ 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, $c_j = 0$), and *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 ($l = 1,\cdots,L$) 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}, which is 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 $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$. 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 ($l = 1, \cdots, L$) is estimated by

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

and the overall classification consistency is

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

The analysis of the classification index is performed based on overall scale scores. Table 31 provides the proportion of classification accuracy and consistency both 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 fact that the intervals used to compute the classification probabilities for students in L1 and L4 [$-\infty$, L2 cut; L4 cut, ∞] are wider than the intervals used to compute the classification probabilities for students in L2 and L3 [L2 cut, L3 cut; L3 cut, L4 cut]. The misclassification probability tends to be higher for narrow intervals.

Accuracy of classifications is higher than the consistency of classifications at all achievement levels. The consistency of classification rates can be lower because the consistency is based on two tests with measurement errors, but 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 31. Classification Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	79	71	83	77
	L1	89	83	90	85
	L2	70	59	74	64
	L3	66	55	80	72
	L4	88	82	90	85
4	Overall	77	69	84	77
	L1	89	83	90	84
	L2	60	47	80	73
	L3	62	52	79	71
	L4	88	82	90	85
5	Overall	78	70	83	76
	L1	89	83	90	85
	L2	64	52	77	68
	L3	72	63	71	61
	L4	86	79	90	86
6	Overall	78	70	83	77
	L1	89	83	92	87
	L2	68	57	78	70
	L3	74	65	72	62
	L4	85	77	90	85
7	Overall	78	70	83	77
	L1	89	83	91	86
	L2	67	56	76	67
	L3	75	67	74	65
	L4	84	75	90	86
8	Overall	79	71	82	75
	L1	88	81	90	85
	L2	70	59	71	61
	L3	77	70	72	61
	L4	84	75	91	86

5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores is also computed by subgroups. Tables 32 and 33 present the marginal reliability coefficients by the subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for Limited English Proficiency (LEP) and Special Education subgroups, a large percentage of whom received Level 1 with large SEMs.

Table 32. Marginal Reliability Coefficients Overall and by Subgroups for ELA/L

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.91	0.90	0.91	0.91	0.90	0.91
Female	0.91	0.90	0.90	0.90	0.90	0.90
Male	0.91	0.91	0.91	0.91	0.91	0.91
African American	0.90	0.89	0.89	0.89	0.88	0.89
American Indian/Alaskan	0.91	0.88	0.91	0.90	0.87	0.90
Asian	0.90	0.88	0.89	0.89	0.88	0.89
Hispanic/Latino	0.90	0.89	0.90	0.89	0.89	0.89
Pacific Islander	0.91	0.90	0.91	0.89	0.93	0.92
White	0.90	0.88	0.89	0.89	0.88	0.89
Two or More Races	0.91	0.91	0.90	0.91	0.91	0.91
LEP	0.86	0.85	0.83	0.79	0.79	0.77
Special Education	0.86	0.86	0.87	0.86	0.85	0.86

Table 33. Marginal Reliability Coefficients Overall and by Subgroups for Mathematics

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.95	0.95	0.94	0.94	0.93	0.93
Female	0.95	0.94	0.93	0.94	0.93	0.93
Male	0.95	0.95	0.94	0.94	0.94	0.94
African American	0.94	0.93	0.90	0.91	0.88	0.88
American Indian/Alaskan	0.94	0.94	0.91	0.93	0.93	0.91
Asian	0.94	0.94	0.94	0.95	0.95	0.95
Hispanic/Latino	0.94	0.93	0.91	0.92	0.89	0.89
Pacific Islander	0.94	0.95	0.93	0.93	0.95	0.95
White	0.94	0.94	0.93	0.94	0.93	0.93
Two or More Races	0.95	0.95	0.94	0.95	0.94	0.93
LEP	0.93	0.91	0.85	0.84	0.79	0.77
Special Education	0.93	0.91	0.87	0.88	0.84	0.84

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 insufficient 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 34 and 35 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 34. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Claim	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1: Reading	14	16	0.76	2440	99	48
	Claims 2 & 4: Writing & Research	16	16	0.83	2426	98	41
	Claim 3: Listening	8	8	0.58	2440	118	76
4	Claim 1: Reading	14	16	0.75	2480	105	53
	Claims 2 & 4: Writing & Research	16	16	0.81	2468	105	46
	Claim 3: Listening	8	8	0.57	2488	126	83
5	Claim 1: Reading	14	16	0.75	2521	106	53
	Claims 2 & 4: Writing & Research	16	16	0.82	2512	106	44
	Claim 3: Listening	8	9	0.58	2511	125	81
6	Claim 1: Reading	14	16	0.77	2526	116	56
	Claims 2 & 4: Writing & Research	16	16	0.80	2529	106	47
	Claim 3: Listening	8	9	0.55	2549	124	84
7	Claim 1: Reading	14	16	0.78	2560	112	52
	Claims 2 & 4: Writing & Research	16	16	0.79	2551	114	52
	Claim 3: Listening	8	9	0.53	2549	126	86
8	Claim 1: Reading	16	16	0.78	2574	112	52
	Claims 2 & 4: Writing & Research	16	16	0.81	2568	112	50
	Claim 3: Listening	8	9	0.55	2589	120	80

Table 35. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	Claim	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1	20	20	0.91	2443	90	27
	Claims 2 and 4	8	11	0.78	2431	96	45
	Claim 3	9	11	0.76	2438	93	45
4	Claim 1	20	20	0.91	2485	89	27
	Claims 2 and 4	8	10	0.74	2476	99	51
	Claim 3	9	10	0.75	2479	98	49
5	Claim 1	20	20	0.89	2512	96	32
	Claims 2 and 4	8	10	0.61	2491	122	76
	Claim 3	9	10	0.71	2501	111	60
6	Claim 1	19	19	0.89	2530	113	37
	Claims 2 and 4	9	10	0.72	2514	124	65
	Claim 3	10	11	0.76	2522	117	57
7	Claim 1	20	20	0.89	2543	119	40
	Claims 2 and 4	10	10	0.66	2525	135	79
	Claim 3	8	10	0.66	2537	129	75
8	Claim 1	20	20	0.89	2560	126	43
	Claims 2 and 4	8	10	0.60	2540	148	93
	Claim 3	9	10	0.71	2548	136	73

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

6. SCORING

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

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

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

Depending on the item score points, the probability $p_{ij}(z_{ij} \mid \theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ 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 $m_i = 1$,

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

where m_i is the maximum possible score point (starting from 0) for the i^{th} item, and *D* is the scale factor, 1.7. The SE is calculated based only on the answered 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 36 presents the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 36. 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, SS_{θ} is the standard error of the ability estimate on the Θ scale, and a is the slope of the scaling constant that transforms Θ into the reporting scale.

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 37 provides three achievement standards for each grade and content area.

Table 37. 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 38 presents the lowest obtainable score (LOT or LOSS) and the highest obtainable score (HOT or HOSS) in both theta and scale score metrics. Estimated thetas lower than LOT or higher than HOT are truncated to the LOT and HOT values, and are assigned LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and claim scores). The standard error for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 38. 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 the IRT maximum likelihood (ML) ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all correct and all incorrect cases, the highest obtainable scores (HOT and HOSS) or the lowest obtainable scores (LOT and LOSS) were assigned.

6.5 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR CLAIM SCORES

In both ELA/L and mathematics, claim scores are computed for claim 1, claims 2 and 4 combined, and claim 3. For each claim score, three performance categories relative strengths and weaknesses are produced.

If the difference between the proficiency cut score and the claim score is greater (or less) than 1.5 times the standard error of the claim score, 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 claim; SS_p is the proficiency scale score cut (Level 3 cut); and $SE(SS_{rc})$ is the standard error of the student's scale score on the claim. HOSS and LOSS are automatically assigned to *Above Standard* and *Below Standard*, respectively.

6.6 TARGET SCORES

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

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

6.6.1 Target Scores Relative to Student's Overall Estimated Ability

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

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

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:

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging the individual student target scores for the target across students of different abilities receiving different items and measuring the same target at different levels of difficulty,

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

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:

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

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

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

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

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

For an aggregate unit, a target score is computed by averaging the individual student target scores for the target across students of different abilities receiving different items and 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 Connecticut Smarter Balanced summative assessments. In ELA/L, short-answer (SA) items and Full Write items are scored by human reader; this is also referred to as "handscoring." 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 that 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. 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. 2017–2018 was the fourth consecutive year that MI scored operational Smarter Balanced assessments, and the majority of readers recruited to score the 2017–2018 summative assessment had previous experience scoring Smarter Balanced assessments.

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 and provide 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 MI also considers 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 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. Additionally, MI conducts an annual review of the reader agreement and scoring materials to inform the development of item-specific, supplemental training materials. Supplemental materials are developed each summer and implemented in the following operational administration.

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 are assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual reader scores is minimized so that the reader becomes highly experienced in scoring responses to a given set of 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, functions as the portal to the VSC interface, and serves as the data repository for 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, how to reference the scoring guide, how to develop the flexibility needed to handle a variety of responses, and how to retain the consistency needed to accurately score all responses. In addition to completing all 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 (e.g., 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 except 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 through the 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 assessments, 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, NC.

More than 20 reports are available and can be customized to meet the information needs of the client and MI's scoring department. These reports provide 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 of 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 further utilized dynamic "threshold" reports which, based on inputted criteria, immediately identify potential scoring performance issues. These reports allow scoring leadership to pinpoint areas of concern and to take corrective action with great efficiency. MI scoring directors are experienced in examining these reports and using the information to determine a need for retraining of individual readers or the group. 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 addressing any problems. 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 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 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 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 randomly selected for second reads 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.

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, as well as at-risk responses that are alerted to the client state for action.

6.7.5 Reader Validity Checks

Approved responses are loaded into the VSC system as validity responses. A small set of validity responses is provided by Smarter Balanced for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The "true" scores for these responses are entered in 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 the 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) that are scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) that 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 for handscored items range from 1–3.

Tables 39–40 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 average of % exact agreement, minimum and maximum % exact agreements, combined % exact and % adjacent agreement, and quadratic weighted Kappa (QWK).

Table 39. ELA/L Reader Agreements for Short-Answer Items

Grade	# of Items	%Exact			%(Exact+ Adjacent)	QWK
		Average	Min	Max		
3	16	79	65	91	100	0.64
4	30	82	68	95	100	0.73
5	21	75	55	92	100	0.68
6	22	74	62	88	100	0.63
7	25	74	59	87	100	0.65
8	23	73	61	81	100	0.65

Table 40. Mathematics Reader Agreements

Grade	Score Points	# of Items	%Exact			%(Exact+ Adjacent)	QWK
			Average	Min	Max		
3	1	12	94	91	97	100	0.86
3	2	26	92	80	100	100	0.93
3	3	4	96	95	97	100	0.98
4	1	8	87	81	94	100	0.68
4	2	36	91	79	99	100	0.90
4	3	4	87	80	91	99	0.92
5	1	4	92	88	99	100	0.61
5	2	41	90	74	98	100	0.89
5	3	8	89	82	98	97	0.84
6	1	14	97	84	99	100	0.91
6	2	32	90	83	98	100	0.90
7	1	8	96	93	99	100	0.79
7	2	25	88	77	94	100	0.84
7	3	2	75	74	75	97	0.81
8	1	14	93	87	98	100	0.84
8	2	26	91	80	99	100	0.91

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that includes the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete a test with handscored items. Because the score reports on student performance are updated each time that students complete tests and the tests are hand-scored, authorized users (e.g., school principals, teachers) can have quickly available information on 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' performance in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year. Additionally, the ORS provides participation data that helps monitor the student participation rates.

This section contains a description of the types of scores reported in the ORS and a description on 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 about how 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 in order to make score reports to be easy to read and understand. This is achieved 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 by 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 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 41 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 via a help button on the ORS.

Table 41. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
District School Teacher Roster	• Number of students tested and percentage of students with Level 3 or 4 (for overall students and by subgroup) • Average scale score and standard error of average scale score (for overall students and by subgroup) • Percentage of students at each achievement level on the overall test and by claims (for overall students and by subgroup) • Performance category in each target (overall students)[1] • Participation rate (for overall students)[2] • On-demand student roster report
Student	• Total scale score and standard error of measurement • Achievement level on overall and claim scores with achievement-level descriptors • Average scale scores and standard errors of average scale scores for student's school, and district

1: Performance category in each target is provided for all aggregate levels.
2: Participation rate reports are provided at the district and school level.

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 42 presents the types of subgroups and subgroup category provided in ORS.

Table 42. 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 Hispanic or Latino Native Hawaiian or Other Pacific Islander White Demographic Race Two or More Races

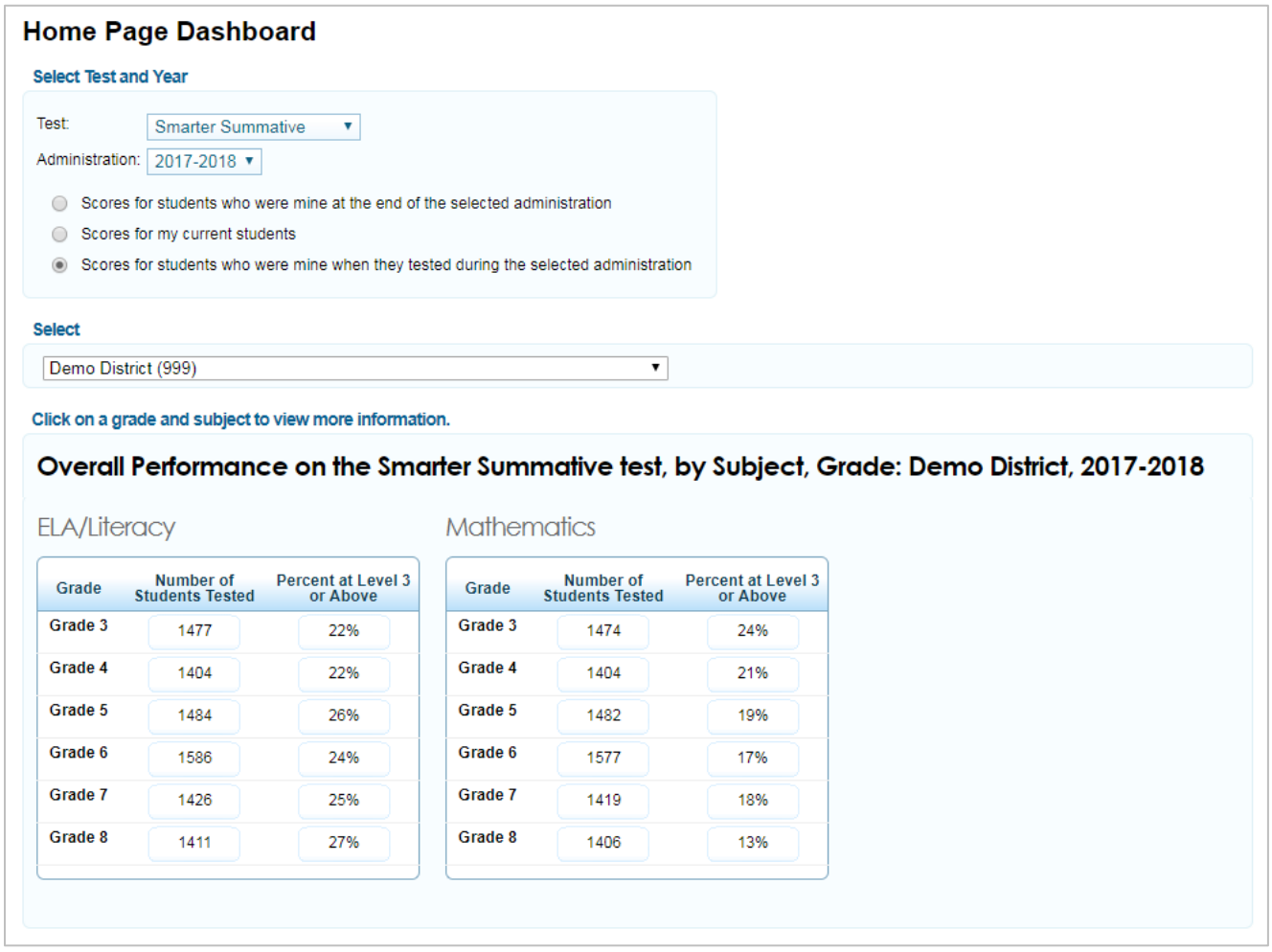
7.1.2 The Online Reporting System

7.1.2.1 Home Page

When users log in to the ORS and select "Score Reports," the first page displays summaries of student 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 student performance for the lower aggregate unit, as well. For example, the district personnel can see a summary of student performance for schools as well as the district.

The home page summarizes student 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

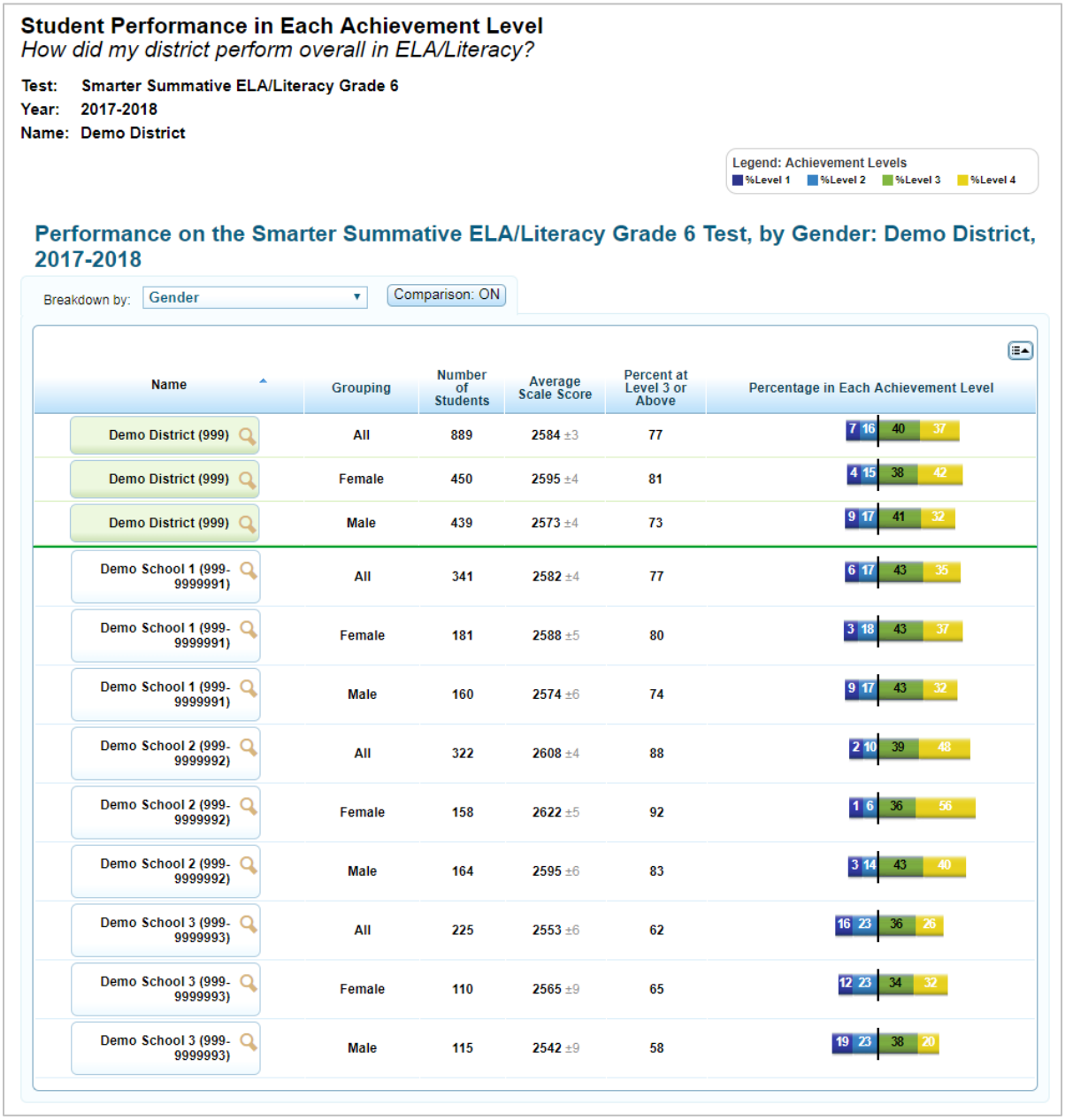


7.1.2.2 Subject Detail Page

More detailed summaries of student performance for 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 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) percentage of students at Level 3 or above, and (4) percentage 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 selects a subgroup of gender.

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



Student Performance in Each Achievement Level

How did my district perform overall in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2017-2018
Name: Demo District

Legend: Achievement Levels
%Level 1 %Level 2 %Level 3 %Level 4

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Gender: Demo District, 2017-2018

Breakdown by: Gender | Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Percentage in Each Achievement Level
Demo District (999)	All	889	2584 ±3	77	7 16 40 37
Demo District (999)	Female	450	2595 ±4	81	4 15 38 42
Demo District (999)	Male	439	2573 ±4	73	9 17 41 32
Demo School 1 (999-9999991)	All	341	2582 ±4	77	6 17 43 35
Demo School 1 (999-9999991)	Female	181	2588 ±5	80	3 18 43 37
Demo School 1 (999-9999991)	Male	160	2574 ±6	74	9 17 43 32
Demo School 2 (999-9999992)	All	322	2608 ±4	88	2 10 39 48
Demo School 2 (999-9999992)	Female	158	2622 ±5	92	1 6 36 56
Demo School 2 (999-9999992)	Male	164	2595 ±6	83	3 14 43 40
Demo School 3 (999-9999993)	All	225	2553 ±6	62	16 23 36 26
Demo School 3 (999-9999993)	Female	110	2565 ±9	65	12 23 34 32
Demo School 3 (999-9999993)	Male	115	2542 ±9	58	19 23 38 20

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) percentage of students at Level 3 or above, and (4) percentage 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 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 mathematics at a district level when users select a subgroup of LEP status.

Exhibit 3. Claim Detail Page for Mathematics by LEP Status: District Level

District Performance for Each Claim

What are my district's strengths and weaknesses in Mathematics?

Test: Smarter Summative Mathematics Grade 6
Year: 2017-2018
Name: Demo District

Legend: Claim Achievement Category
%Below Standard | %Approaching Standard | %Above Standard

Performance on the Smarter Summative Mathematics Grade 6 Test, by Claim: Demo District, 2017-2018

Breakdown by: All | Comparison: ON

Name	Number of Students	Average Scale Score	Percent at Level 3 or Above	Claims**	Claim Average Scale Score	Percent at Each Claim Achievement Category
Demo District (999)	887	2582 ±3	69	**Mathematics**		
				Concepts and Procedures	2583 ±3	14 40 46
				Problem Solving and Modeling & Data Analysis	2578 ±3	13 45 42
				Communicating Reasoning	2579 ±3	14 45 41
Demo School 1 (999-9999991)	340	2585 ±4	67	**Mathematics**		
				Concepts and Procedures	2588 ±4	14 39 47
				Problem Solving and Modeling & Data Analysis	2578 ±5	12 45 43
				Communicating Reasoning	2578 ±5	16 42 42
Demo School 2 (999-9999992)	225	2554 ±6	56	**Mathematics**		
				Concepts and Procedures	2550 ±6	25 44 31
				Problem Solving and Modeling & Data Analysis	2551 ±7	19 50 31
				Communicating Reasoning	2557 ±7	20 47 33

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 of 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 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: 2017-2018
Name: Demo School

Legend: Areas of Strongest and Weakest Performance
- Area of Strengths
- Performance is similar to performance on the test as a whole
- Area of Weakness
- Insufficient Information

Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Approaching Proficiency Standard
- Below the Proficiency Standard
- Insufficient Information

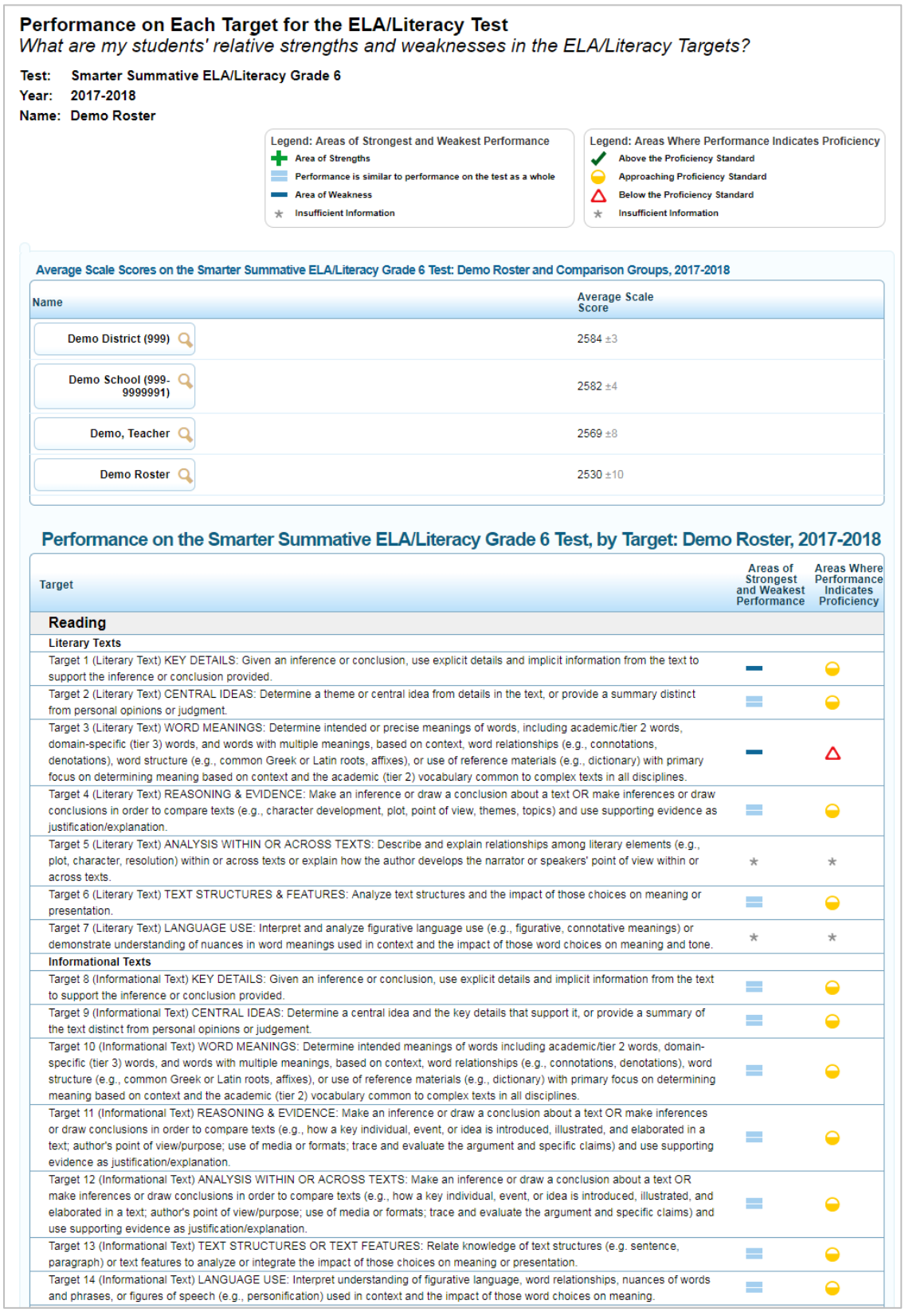
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2584 ±3
Demo School (999-9999991)	2582 ±4

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: Demo School, 2017-2018

Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Reading		
Literary Texts		
Target 1 (Literary Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	=	◒
Target 2 (Literary Text) CENTRAL IDEAS: Determine a theme or central idea from details in the text, or provide a summary distinct from personal opinions or judgment.	=	◒
Target 3 (Literary Text) WORD MEANINGS: Determine intended or precise 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.	=	✓
Target 4 (Literary 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., character development, plot, point of view, themes, topics) and use supporting evidence as justification/explanation.	=	✓
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Describe and explain relationships among literary elements (e.g., plot, character, resolution) within or across texts or explain how the author develops the narrator or speakers' point of view within or across texts.	*	*
Target 6 (Literary Text) TEXT STRUCTURES & FEATURES: Analyze text structures and the impact of those choices on meaning or presentation.	=	◒
Target 7 (Literary Text) LANGUAGE USE: Interpret and analyze figurative language use (e.g., figurative, connotative meanings) or demonstrate understanding of nuances in word meanings used in context and the impact of those word choices on meaning and tone.	+	✓
Informational Texts		
Target 8 (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.	=	✓
Target 9 (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.	=	✓
Target 10 (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.	=	✓
Target 11 (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.	=	◒
Target 12 (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.	=	✓
Target 13 (Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. sentence, paragraph) or text features to analyze or integrate the impact of those choices on meaning or presentation.	=	◒
Target 14 (Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., personification) used in context and the impact of those word choices on meaning.	=	◒

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



Performance on Each Target for the ELA/Literacy Test

What are my students' relative strengths and weaknesses in the ELA/Literacy Targets?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2017-2018
Name: Demo Roster

Legend: Areas of Strongest and Weakest Performance
- Area of Strengths
- Performance is similar to performance on the test as a whole
- Area of Weakness
- Insufficient Information

Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Approaching Proficiency Standard
- Below the Proficiency Standard
- Insufficient Information

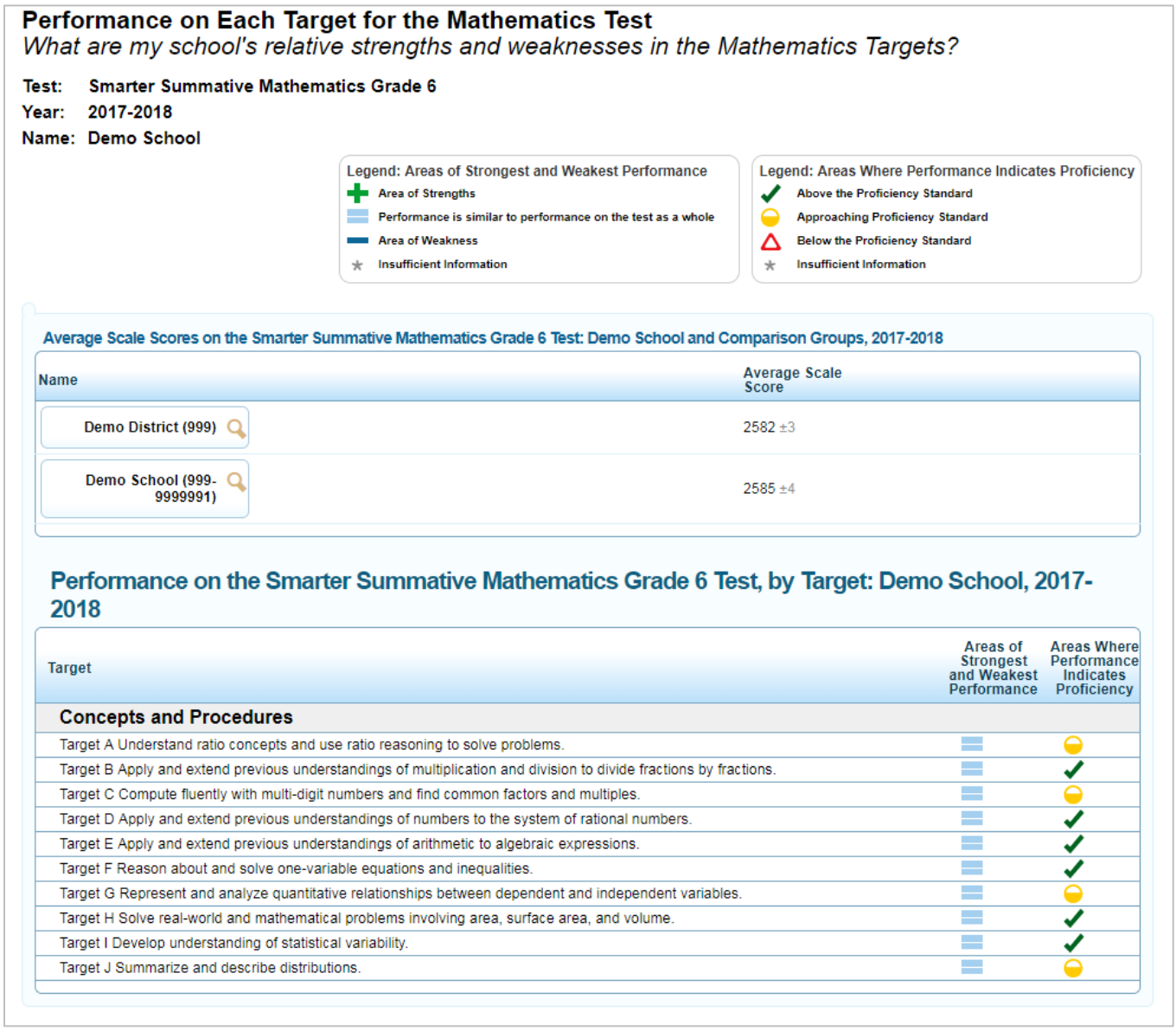
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo Roster and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2584 ±3
Demo School (999-9999991)	2582 ±4
Demo, Teacher	2569 ±8
Demo Roster	2530 ±10

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Target: Demo Roster, 2017-2018

Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Reading		
Literary Texts		
Target 1 (Literary Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	Area of Weakness	Approaching Proficiency Standard
Target 2 (Literary Text) CENTRAL IDEAS: Determine a theme or central idea from details in the text, or provide a summary distinct from personal opinions or judgment.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 3 (Literary Text) WORD MEANINGS: Determine intended or precise 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.	Area of Weakness	Below the Proficiency Standard
Target 4 (Literary 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., character development, plot, point of view, themes, topics) and use supporting evidence as justification/explanation.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Describe and explain relationships among literary elements (e.g., plot, character, resolution) within or across texts or explain how the author develops the narrator or speakers' point of view within or across texts.	Insufficient Information	Insufficient Information
Target 6 (Literary Text) TEXT STRUCTURES & FEATURES: Analyze text structures and the impact of those choices on meaning or presentation.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 7 (Literary Text) LANGUAGE USE: Interpret and analyze figurative language use (e.g., figurative, connotative meanings) or demonstrate understanding of nuances in word meanings used in context and the impact of those word choices on meaning and tone.	Insufficient Information	Insufficient Information
Informational Texts		
Target 8 (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.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 9 (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.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 10 (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.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 11 (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.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 12 (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.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 13 (Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. sentence, paragraph) or text features to analyze or integrate the impact of those choices on meaning or presentation.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target 14 (Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., personification) used in context and the impact of those word choices on meaning.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard

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: 2017-2018
Name: Demo School

Legend: Areas of Strongest and Weakest Performance
- Area of Strengths
- Performance is similar to performance on the test as a whole
- Area of Weakness
- Insufficient Information

Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Approaching Proficiency Standard
- Below the Proficiency Standard
- Insufficient Information

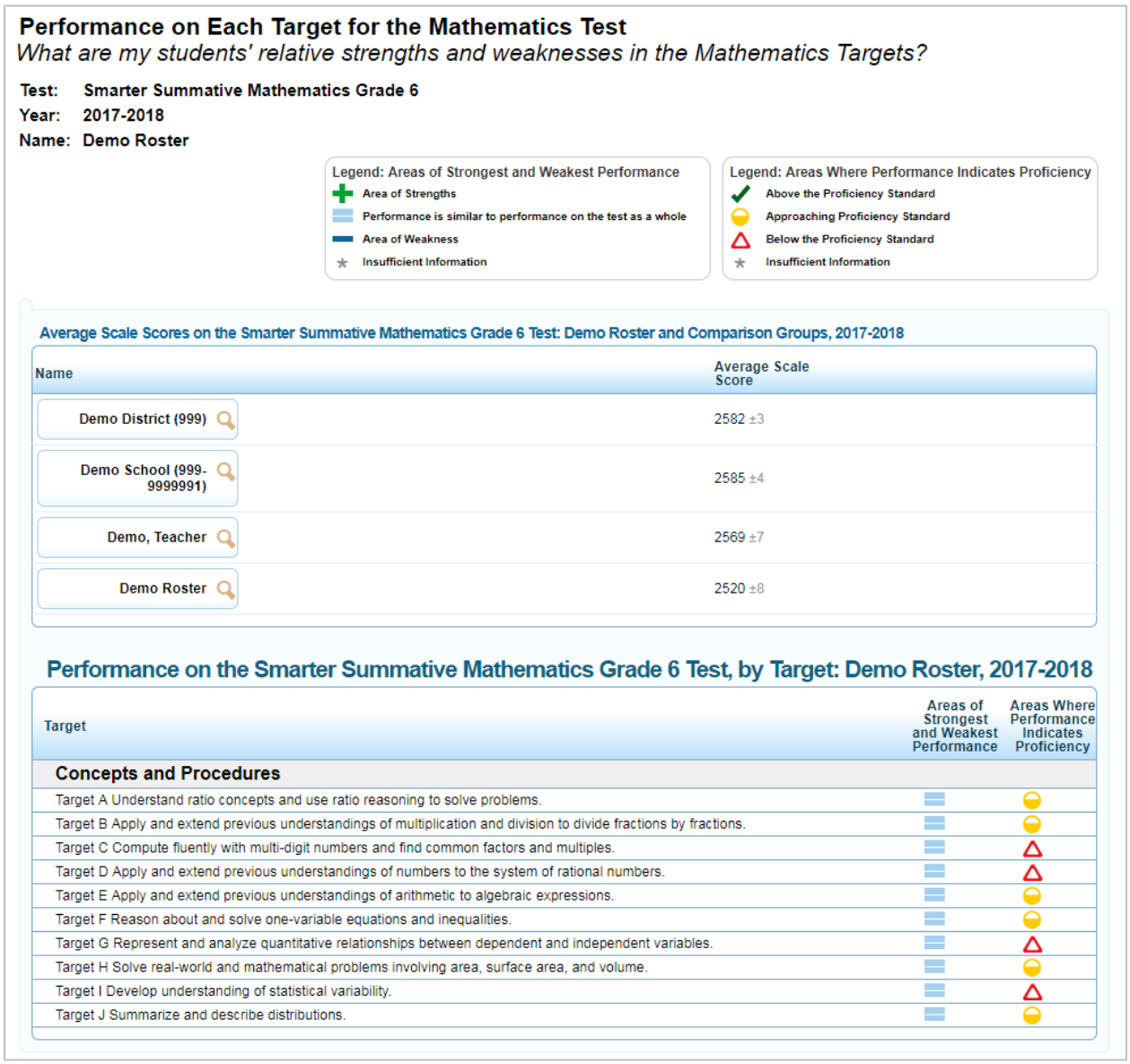
Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2582 ±3
Demo School (999-9999991)	2585 ±4

Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: Demo School, 2017-2018

Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Concepts and Procedures		
Target A Understand ratio concepts and use ratio reasoning to solve problems.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target B Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target C Compute fluently with multi-digit numbers and find common factors and multiples.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target D Apply and extend previous understandings of numbers to the system of rational numbers.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target E Apply and extend previous understandings of arithmetic to algebraic expressions.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target F Reason about and solve one-variable equations and inequalities.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target G Represent and analyze quantitative relationships between dependent and independent variables.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard
Target H Solve real-world and mathematical problems involving area, surface area, and volume.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target I Develop understanding of statistical variability.	Performance is similar to performance on the test as a whole	Above the Proficiency Standard
Target J Summarize and describe distributions.	Performance is similar to performance on the test as a whole	Approaching Proficiency Standard

Exhibit 7. Target Detail Page for Mathematics: Teacher Level



Performance on Each Target for the Mathematics Test

What are my students' relative strengths and weaknesses in the Mathematics Targets?

Test: Smarter Summative Mathematics Grade 6
Year: 2017-2018
Name: Demo Roster

Legend: Areas of Strongest and Weakest Performance
- Area of Strengths
- Performance is similar to performance on the test as a whole
- Area of Weakness
- Insufficient Information

Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Approaching Proficiency Standard
- Below the Proficiency Standard
- Insufficient Information

Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: Demo Roster and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2582 ±3
Demo School (999-9999991)	2585 ±4
Demo, Teacher	2569 ±7
Demo Roster	2520 ±8

Performance on the Smarter Summative Mathematics Grade 6 Test, by Target: Demo Roster, 2017-2018

Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Concepts and Procedures		
Target A Understand ratio concepts and use ratio reasoning to solve problems.		
Target B Apply and extend previous understandings of multiplication and division to divide fractions by fractions.		
Target C Compute fluently with multi-digit numbers and find common factors and multiples.		
Target D Apply and extend previous understandings of numbers to the system of rational numbers.		
Target E Apply and extend previous understandings of arithmetic to algebraic expressions.		
Target F Reason about and solve one-variable equations and inequalities.		
Target G Represent and analyze quantitative relationships between dependent and independent variables.		
Target H Solve real-world and mathematical problems involving area, surface area, and volume.		
Target I Develop understanding of statistical variability.		
Target J Summarize and describe distributions.		

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 shows 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.

Specifically, the student's name, scale score with SEM, and achievement level shown at the top of the page. On the left middle section, the student's performance is described in detail using a barrel chart. In the 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 multiple 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 claim is displayed alongside a description of his/her performance on each claim. 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

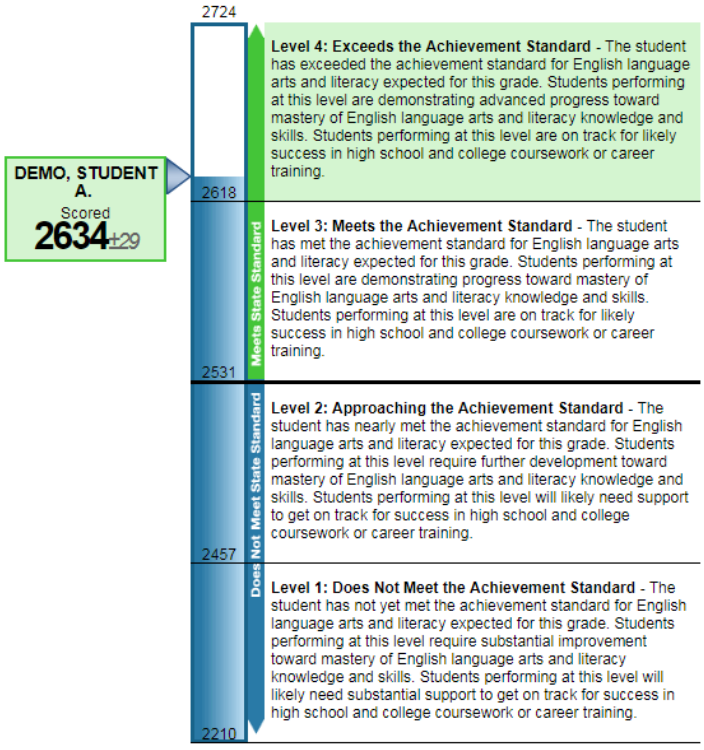
How did my student perform on the ELA/Literacy test?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2017-2018
Name: DEMO, STUDENT A.

Overall Performance on the Smarter Summative ELA/Literacy Grade 6 Test: DEMO, STUDENT A., 2017-2018

Name	SSID	Scale Score	Achievement Level
DEMO, STUDENT A.	9999999991	2634 ±29	Level 4

Scale Score and Performance on the Smarter Summative ELA/Literacy Grade 6 Test: DEMO, STUDENT A., 2017-2018



Average Scale Scores on the Smarter Summative ELA/Literacy Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2584 ±3
Demo School (999-9999991)	2582 ±4

Information on Standard Error of Measurement

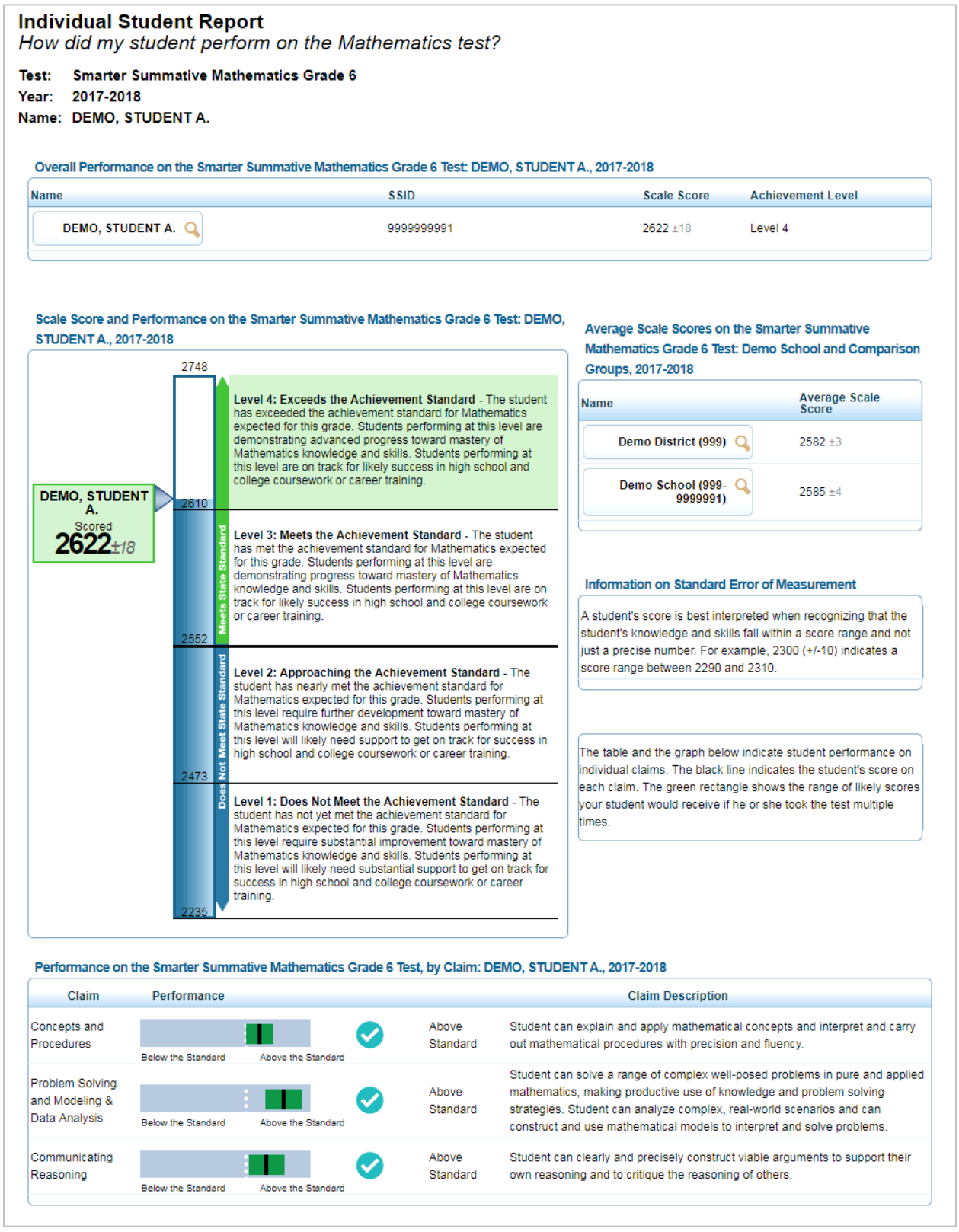
A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (+/-10) indicates a score range between 2290 and 2310.

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.

Performance on the Smarter Summative ELA/Literacy Grade 6 Test, by Claim: DEMO, STUDENT A., 2017-2018

Claim	Performance		Claim Description
Reading	Below the Standard / Above the Standard	Above Standard	Student can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Listening	Below the Standard / Above the Standard	Above Standard	Student can employ effective listening skills for a range of purposes and audiences.
Writing and Research/Inquiry	Below the Standard / Above the Standard	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: 2017-2018
Name: DEMO, STUDENT A.

Overall Performance on the Smarter Summative Mathematics Grade 6 Test: DEMO, STUDENT A., 2017-2018

Name	SSID	Scale Score	Achievement Level
DEMO, STUDENT A.	9999999991	2622 ±18	Level 4

Scale Score and Performance on the Smarter Summative Mathematics Grade 6 Test: DEMO, STUDENT A., 2017-2018

DEMO, STUDENT A. Scored 2622±18

2748

Level 4: Exceeds the Achievement Standard - The student has exceeded the achievement standard for Mathematics expected for this grade. Students performing at this level are demonstrating advanced progress toward mastery of Mathematics knowledge and skills. Students performing at this level are on track for likely success in high school and college coursework or career training.

2610

Meets State Standard

Level 3: Meets the Achievement Standard - The student has met the achievement standard for Mathematics expected for this grade. Students performing at this level are demonstrating progress toward mastery of Mathematics knowledge and skills. Students performing at this level are on track for likely success in high school and college coursework or career training.

2552

Does Not Meet State Standard

Level 2: Approaching the Achievement Standard - The student has nearly met the achievement standard for Mathematics expected for this grade. Students performing at this level require further development toward mastery of Mathematics knowledge and skills. Students performing at this level will likely need support to get on track for success in high school and college coursework or career training.

2473

Level 1: Does Not Meet the Achievement Standard - The student has not yet met the achievement standard for Mathematics expected for this grade. Students performing at this level require substantial improvement toward mastery of Mathematics knowledge and skills. Students performing at this level will likely need substantial support to get on track for success in high school and college coursework or career training.

2235

Average Scale Scores on the Smarter Summative Mathematics Grade 6 Test: Demo School and Comparison Groups, 2017-2018

Name	Average Scale Score
Demo District (999)	2582 ±3
Demo School (999-9999991)	2585 ±4

Information on Standard Error of Measurement

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

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.

Performance on the Smarter Summative Mathematics Grade 6 Test, by Claim: DEMO, STUDENT A., 2017-2018

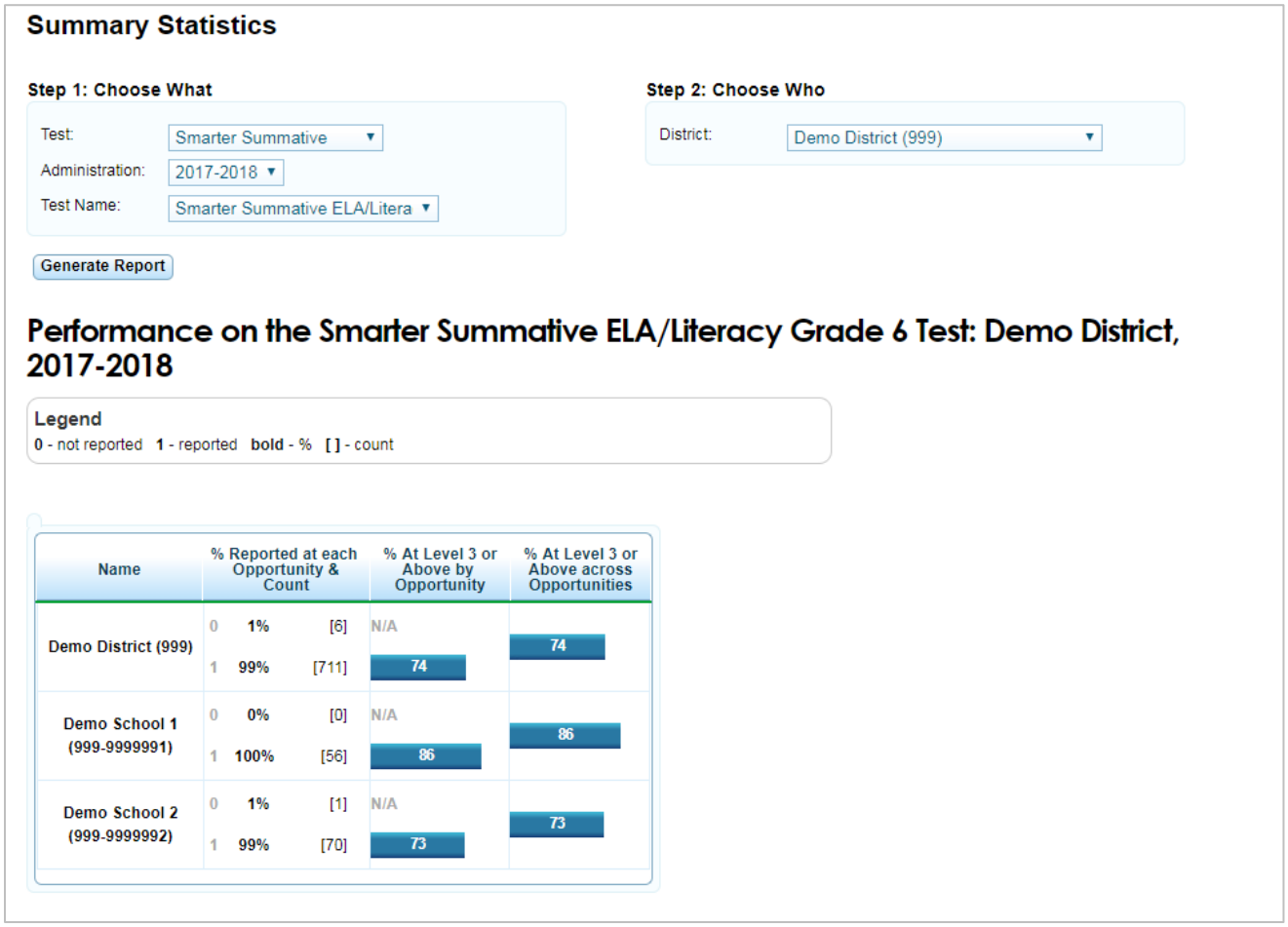
Claim	Performance		Claim Description
Concepts and Procedures	Below the Standard / Above the Standard	Above Standard	Student can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Problem Solving and Modeling & Data Analysis	Below the Standard / Above the Standard	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	Below the Standard / Above the Standard	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 the students complete tests, and these tests are handscored. Included in the participation table are (1) the number and percentage of students who are tested and not tested and (2) the percentage of students with achievement levels of 3 or above.

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

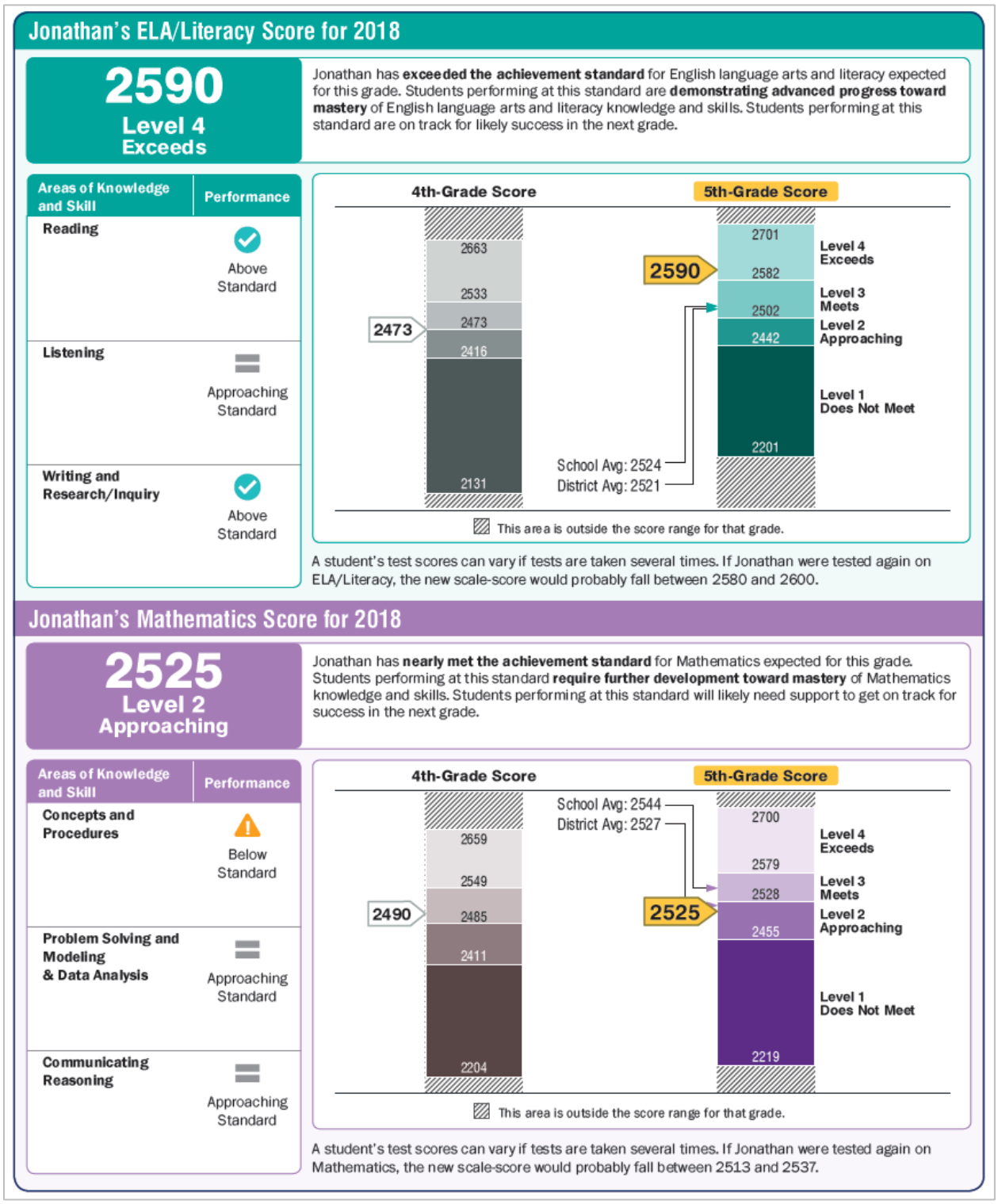
Exhibit 10. Participation Rate Report at District Level



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 at 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. 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 multiple times, the resulting scale score would vary across administrations, sometimes a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered multiple times. When interpreting scale scores, it is recommended to consider the range of scale scores, incorporating the SEM of the scale score.

The "±" 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 ± 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 the 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 n 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

Student 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 the 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 student 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 are 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 two types of relative strength and weakness scores for each target within a claim. The strengths and weaknesses reports are generated for aggregate units of classroom, school, and district and provide information about how a group of students in a class, school, or district performed on each target, either relative to their performance on the test as a whole or relative to the proficiency cut set by Smarter Balanced. Specifically, for target performance relative to the test as a whole, students' observed performance on items within the reporting element is compared 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., roster, teacher, school, or district) shows a relative strength in that target. Conversely, when observed performance within a target is below the level expected based on overall achievement, then the reporting unit shows a relative weakness in that target. For target performance relative to proficiency, students' observed performance on items within the reporting element is compared with proficiency cut (i.e., Achievement Level 3 cut). At the aggregate level, when observed performance within a target is greater than the proficiency cut, the reporting unit shows a relative strength in that target. Conversely, when observed performance within a target is below the proficiency cut, the reporting unit shows a relative weakness in that target.

The performance on target shows how a group of students performed on each target either relative to their overall subject performance on a test or relative to proficiency standard. The performance on target is mapped into three performance categories: (1) better than performance on the test as a whole (higher than expected) or relative to proficiency standard, (2) similar to performance on the test as a whole or relative to proficiency standard, and (3) worse than performance on the test as a whole (lower than expected) or relative to proficiency standard. "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 performance 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

Student 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 the knowledge and skills that a group of students possesses. 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 percentage of students in each achievement level for the overall test and by claim are reported at the aggregate level to represent how well a group of students performs on the overall test, and by claim.

7.4 Appropriate Uses for Scores and Reports

Assessment results can be used to provide information about 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 give further information on whether students are on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can be used to identify students' relative strengths and weaknesses in certain content areas. For example, performance categories for claims can be used to identify an individual student's relative strengths and weaknesses among claims within a content area.

Assessment results for 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 could perform very well in the overall test, but it is possible that they would not perform as well in several targets compared to their overall performance. In this case, teachers and schools can identify the 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. Furthermore, 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 in a specific target in a claim.

In addition, assessment results can be used to compare student performance among different students and among different groups. Teachers can evaluate how their students perform compared with students in other schools, and districts overall as well as 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. In the Smarter Balanced assessments, the scale scores across grades are on the same scale because the scores are vertically linked across grades. Therefore, scale scores from one grade can be compared with the next grade, i.e., measuring the growth.

While assessment results provide valuable information to understand student performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and therefore do not represent a precise measure of 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 about 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 uses a series of quality control steps to ensure the error-free production of score reports in both online and paper-pencil formats. The quality of the information produced in the test delivery system (TDS) is tested thoroughly before, during, and after the testing window opens.

8.1 ADAPTIVE TEST CONFIGURATION

For the 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 independently checked and confirmed numerous times by multiple staff members before the testing window opens.

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 as well as 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 (computer-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 such as Windows, Linux, and iOS to ensure that the item looks consistent in all of them. Some of the layouts have the stimulus and item response options/response area displayed side by side. In each of these layouts, both stimulus and response options have independent scroll bars.

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

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

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

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 quickly and to 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 the 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 pulled 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 to monitor the scorer status. VSC is capable of dynamically embedding calibration responses in scoring sets as individual items or in sets of whichever 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-reads, 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 scorer quality and to take any corrective actions immediately. Updated real-time reports that show both daily and cumulative (project-to-date) are available. 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) and has 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 taker. MI 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 of 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—such as data about how long it takes to load, view, or respond to an item—are captured for each assessed student. All of this information is logged as well, enabling 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 of a specified range are flagged for reporting or to generate reports based on all items in the pool.

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

Table 43 presents an overview of the QA reports.

Table 43. 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 assessments, 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 the 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. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called 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 are 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 AIR 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. Additionally, 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.

Drasgow, F., Levine, M.V. & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Guo, F. (2006). Expected Classification Accuracy using the Latent Distribution. *Practical, Assessment, Research & Evaluation, 11*(6).

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement, 13*(4), 253–264.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16*(4), 247–260.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Philippine Statistician, 52*(1–4), 81–92.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced. *Journal of Educational Measurement, 13*(4), 265–276.

APPENDICES

Appendix A: Summary of the 2017–2018 Interim Assessments

The Interim Comprehensive Assessments (ICAs) are 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, and Table A–2 presents the ICA results for all students, including the average and the standard deviation of scale scores, the percentage of students in each achievement level, and the percentage of proficient students.

Table A–1. Number of Students Who Took ICAs

Grade	ELA/L				Mathematics			
	One	Two	Three	Total	One	Two	Three	Total
3	93	7	44	144	76	2	47	125
4	30	0	0	30	111	0	0	111
5	73	0	0	73	84	1	0	85
6	159	0	0	159	423	2	0	425
7	179	0	0	179	373	0	0	373
8	171	0	0	171	336	0	0	336

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

Subject	Grade	Number of Tests	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/L	3	239	2455	79	14	24	27	35	62
	4	30	2470	65	23	23	33	20	53
	5	73	2628	50	0	0	22	78	100
	6	159	2455	65	46	42	12	0	12
	7	179	2473	78	53	31	16	0	16
	8	171	2495	79	44	37	19	0	19
Math	3	221	2446	70	17	31	29	23	52
	4	111	2465	67	21	41	30	9	39
	5	86	2551	106	23	17	14	45	59
	6	427	2496	108	38	32	14	15	30
	7	373	2538	114	37	25	15	23	38
	8	336	2500	117	57	17	14	12	25

Note: Number of Tests is based on the total tests, adding multiple times for the students who took the same test more than once. The percentage of each achievement level may not add up to 100% or %Proficient due to rounding.

For the Interim Assessment Block assessments (IABs), there were 7–9 IABs for ELA/L and 6–10 IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–3 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 15,257 students took the IABs, and among 15,257 students, 6,542 students took one IAB, 3,528 students took two IABs, and so on.

Tables A–4 to A–6 disaggregated the number of students in Table A–2 by each individual block. For example, 6,542 students in grade 3 ELA/L took one IAB only. Among 6,542 students, 10 of the students took the Brief Writes IAB. Tables A–7 to A–9 show the percentage of students in each performance category for all students for each IAB.

Table A–3. Number of Students Who Took IABs

Grade	Total	Number of IABs Taken								
		1	2	3	4	5	6	7	8	9
ELA/L										
3	15,257	6,542	3,528	2,554	1,141	697	364	358	37	36
4	15,979	5,721	4,402	2,931	1,460	683	523	141	93	25
5	16,698	6,625	4,904	2,473	1,087	726	675	145	63	
6	17,135	5,873	4,837	3,557	1,614	601	538	114	1	
7	17,596	6,235	4,523	3,214	1,836	1,055	594	138	1	
8	17,165	8,199	4,463	2,941	1,045	351	166			
Mathematics										
3	24,891	9,320	5,921	5,797	2,540	1,262	51			
4	24,770	9,336	6,487	5,598	1,778	1,550	21			
5	24,657	9,877	6,919	4,978	1,695	1,173	15			
6	25,310	11,302	7,130	5,131	956	772	19			
7	24,779	10,536	6,785	5,831	794	826	7			
8	24,914	11,981	6,287	4,918	1,306	415	7			

Table A–4: ELA/L Number of Students Who Took IABs by Block Labels (Grades 3–5)

Grade	Block	Number of IABs Taken								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
3	Brief Writes	10	93	107	113	137	69	31	37	36
	Editing	1,265	988	1,252	646	563	338	356	37	36
	Language and Vocabulary Use	1,314	1,293	1,403	527	528	302	357	37	36
	Listening and Interpretation	962	945	1,088	990	663	351	358	36	36
	Reading Informational Text	1,162	1,703	1,422	638	423	321	353	36	36
	Reading Literary Text	1,464	1,313	1,172	614	423	315	357	36	36
	Research	107	212	512	302	360	235	338	36	36
	Revision	252	391	704	734	385	253	356	37	36
	Performance Task	6	118	2		3			4	36
4	Brief Writes	2	44	306	96	65	20	48	91	25
	Editing	904	1,043	1,270	854	637	519	141	93	25
	Language and Vocabulary Use	709	1,785	1,539	815	422	297	140	93	25
	Listening and Interpretation	959	975	1,409	1,079	562	406	135	93	25
	Reading Informational Text	1,457	2,473	1,505	797	525	521	139	93	25
	Reading Literary Text	1,029	1,375	1,474	812	385	520	138	93	25
	Research	530	424	603	320	292	334	105	93	25
	Revision	119	581	650	1,039	527	516	141	93	25
	Performance Task	12	104	37	28		5		2	25
5	Brief Writes	6	20	182	80	72	15	2	63	
	Editing	1,455	956	912	616	601	663	145	63	
	Language and Vocabulary Use	979	1,955	1,223	551	552	613	144	63	
	Listening and Interpretation	792	1,647	1,248	886	576	482	144	63	
	Reading Informational Text	1,675	2,381	1,207	665	478	621	145	63	
	Reading Literary Text	1,377	1,886	1,295	646	452	665	145	63	
	Research	235	310	624	199	296	332	145	63	
	Revision	102	643	681	690	565	659	145	63	
	Performance Task	4	10	47	15	38				

Table A–5: ELA/L Number of Students Who Took IABs by Block Labels (Grades 6–8)

Grade	Block	Number of IABs Taken							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
6	Brief Writes	41	133	89	29	36	178	61	1
	Editing	1,539	1,523	2,200	1,450	496	296	114	1
	Language and Vocabulary Use	449	1,866	1,048	845	452	376	114	1
	Listening and Interpretation	1,148	963	1,342	1,030	417	522	114	1
	Reading Informational Text	1,199	2,514	1,799	920	250	526	114	1
	Reading Literary Text	719	1,139	1,552	981	498	426	114	1
	Research	432	501	983	374	303	536	114	1
	Revision	247	990	1,504	826	551	368	53	1
	Performance Task	99	45	154	1	2			
7	Brief Writes	15	167	5	14	46	226	1	1
	Editing	1,240	1,656	2,319	1,178	859	411	138	1
	Language and Vocabulary Use	382	1,406	773	863	427	369	137	1
	Listening and Interpretation	518	886	1,125	1,043	932	510	138	1
	Reading Informational Text	1,274	1,717	1,106	1,011	657	503	138	1
	Reading Literary Text	2,105	1,541	1,844	1,212	894	594	138	1
	Research	381	893	601	778	1,027	593	138	1
	Revision	252	780	1,869	1,245	433	358	137	1
	Performance Task	68						1	
8	Brief Writes	4	185	27	77	54	164		
	Editing and Revising	3,688	3,068	2,488	898	343	166		
	Listening and Interpretation	1,217	986	1,917	689	331	166		
	Reading Informational Text	895	2,225	1,273	831	342	166		
	Reading Literary Text	1,883	1,410	1,900	994	334	166		
	Research	504	883	1,213	691	350	166		
	Performance Task	8	169	5		1	2		

Table A–6: Mathematics Number of Students Who Took IABs by Block Labels

Grade	Block	Number of IABs Taken					
		1	**2**	**3**	**4**	**5**	**6**
3	Geometry	429	699	1,161	1,122	1,150	51
	Measurement and Data	303	850	1,843	1,846	1,199	51
	Number and Operations in Base Ten	3,220	3,363	4,789	2,413	1,254	51
	Number and Operations – Fractions	2,081	3,008	4,607	2,410	1,242	51
	Operational and Algebraic Thinking	3,130	3,672	4,757	2,253	1,243	51
	Performance Task	157	250	234	116	222	51
4	Geometry	357	845	960	1,047	1,494	21
	Measurement and Data	243	956	702	1,041	1,548	21
	Number and Operations in Base Ten	4,393	4,357	5,144	1,710	1,549	21
	Number and Operations – Fractions	2,561	3,911	4,986	1,596	1,550	21
	Operational and Algebraic Thinking	1,649	2,696	4,696	1,482	1,529	21
	Performance Task	133	209	306	236	80	21
5	Geometry	311	572	777	1,037	1,172	15
	Measurement and Data	245	1,009	2,685	1,324	1,165	15
	Number and Operations in Base Ten	3,857	5,144	4,727	1,568	1,173	15
	Number and Operations – Fractions	4,090	4,647	4,615	1,630	1,173	15
	Operations and Algebraic Thinking	1,134	2,189	2,091	1,114	1,151	15
	Performance Task	240	277	39	107	31	15
6	Expressions and Equations	2,888	3,596	4,892	923	772	19
	Geometry	1,153	1,555	674	664	695	19
	Number System	4,400	4,339	4,599	854	767	19
	Ratios and Proportional Relationships	2,618	4,113	4,864	860	771	19
	Statistics and Probability	133	355	208	400	747	19
	Performance Task	110	302	156	123	108	19
7	Expressions and Equations	2,606	4,369	5,267	717	826	7
	Geometry	566	754	972	552	825	7
	Number System	3,427	4,856	5,285	738	823	7
	Ratios and Proportional Relationships	3,895	3,354	5,374	724	825	7
	Statistics and Probability	36	219	562	339	826	7
	Performance Task	6	18	33	106	5	7
8	Expressions and Equations I	3,597	2,729	4,412	1,189	415	7
	Expressions and Equations II	372	954	3,296	1,013	357	7
	Functions	4,115	3,697	2,929	1,218	414	7
	Geometry	1,713	2,783	2,825	812	414	7
	Number System	2,106	2,248	1,177	661	399	7
	Performance Task	78	163	115	331	76	7

Table A-7: ELA/L Percentage of Students in Achievement Levels by IAB Block Labels (Grades 3–5)

Grade	Block	Number Tested	% Below	% At/Near	% Above
3	Brief Writes	633	30	45	25
	Editing	5,481	25	52	23
	Language and Vocabulary Use	5,797	27	46	27
	Listening and Interpretation	5,429	16	53	31
	Reading Informational Text	6,094	20	51	29
	Reading Literary Text	5,730	27	38	35
	Research	2,138	15	40	45
	Revision	3,148	24	44	32
	Performance Task	169	11	53	36
4	Brief Writes	697	20	45	35
	Editing	5,486	25	54	22
	Language and Vocabulary Use	5,825	27	43	30
	Listening and Interpretation	5,643	12	56	33
	Reading Informational Text	7,535	14	52	33
	Reading Literary Text	5,851	22	48	30
	Research	2,726	23	41	36
	Revision	3,691	20	50	30
	Performance Task	213	16	56	27
5	Brief Writes	440	45	37	18
	Editing	5,411	18	48	33
	Language and Vocabulary Use	6,080	26	48	26
	Listening and Interpretation	5,838	13	48	39
	Reading Informational Text	7,235	9	50	41
	Reading Literary Text	6,529	14	45	41
	Research	2,204	20	39	40
	Revision	3,548	25	43	32
	Performance Task	114	14	58	28

Table A-8: ELA/L Percentage of Students in Achievement Levels by IAB Block Labels (Grades 6–8)

Grade	Block	Number Tested	% Below	% At/Near	% Above
6	Brief Writes	568	29	53	18
	Editing	7,619	22	60	18
	Language and Vocabulary Use	5,151	28	47	25
	Listening and Interpretation	5,537	19	44	37
	Reading Informational Text	7,323	18	49	33
	Reading Literary Text	5,430	16	51	33
	Research	3,244	22	40	38
	Revision	4,540	28	53	19
	Performance Task	301	16	48	37
7	Brief Writes	475	41	39	20
	Editing	7,802	17	68	15
	Language and Vocabulary Use	4,358	29	48	23
	Listening and Interpretation	5,153	19	54	27
	Reading Informational Text	6,407	23	48	30
	Reading Literary Text	8,329	24	48	29
	Research	4,412	17	52	32
	Revision	5,075	32	50	17
	Performance Task	69	6	29	65
8	Brief Writes	511	29	45	26
	Editing and Revising	10,651	22	54	24
	Listening and Interpretation	5,306	15	60	25
	Reading Informational Text	5,732	15	45	40
	Reading Literary Text	6,687	19	43	38
	Research	3,807	22	43	35
	Performance Task	185	11	39	50

Table A-9: Mathematics Percentage of Students in Achievement Levels by IAB Block Labels

Grade	Block	Number Tested	% Below	% At/Near	% Above
3	Geometry	4,612	20	51	29
	Measurement and Data	6,092	20	41	39
	Number and Operations in Base Ten	15,090	32	38	30
	Number and Operations – Fractions	13,399	14	43	43
	Operational and Algebraic Thinking	15,106	31	48	21
	Performance Task	1,030	14	50	36
4	Geometry	4,724	6	60	33
	Measurement and Data	4,511	12	46	42
	Number and Operations in Base Ten	17,174	27	47	26
	Number and Operations – Fractions	14,625	26	41	33
	Operational and Algebraic Thinking	12,073	31	49	20
	Performance Task	985	12	52	36
5	Geometry	3,884	21	53	26
	Measurement and Data	6,443	24	43	34
	Number and Operations in Base Ten	16,484	31	46	23
	Number and Operations – Fractions	16,170	32	43	25
	Operations and Algebraic Thinking	7,694	19	47	34
	Performance Task	709	16	47	37
6	Expressions and Equations	13,090	29	41	30
	Geometry	4,760	25	37	38
	Number System	14,978	35	43	22
	Ratios and Proportional Relationships	13,245	37	35	28
	Statistics and Probability	1,862	10	56	35
	Performance Task	818	15	73	12
7	Expressions and Equations	13,792	24	44	32
	Geometry	3,676	15	58	27
	Number System	15,136	25	49	26
	Ratios and Proportional Relationships	14,179	24	52	24
	Statistics and Probability	1,989	27	49	25
	Performance Task	175	28	51	21
8	Expressions and Equations I	12,349	32	50	18
	Expressions and Equations II	5,999	33	41	26
	Functions	12,380	36	42	22
	Geometry	8,554	19	46	35
	Number System	6,598	26	37	36
	Performance Task	770	17	59	24

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

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

Group	2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3												
All Students	38,942	54	2438	89	38,097	52	2432	91	37,525	53	2435	90
Female	19,139	58	2447	88	18,506	56	2442	89	18,417	57	2443	88
Male	19,803	50	2430	90	19,591	48	2423	91	19,108	49	2427	91
African American	4,874	31	2392	81	4,841	30	2388	83	4,764	33	2395	84
AmerIndian/Alaskan	90	48	2422	78	97	37	2399	83	110	50	2422	87
Asian	2,151	74	2480	84	2,049	71	2472	84	2,022	73	2479	85
Hispanic/Latino	9,854	33	2395	82	9,847	31	2390	85	10,287	32	2392	84
Pacific Islander	47	38	2420	92	33	61	2444	84	46	46	2438	85
White	20,601	67	2465	82	19,903	65	2459	83	18,889	67	2464	80
Two or More Races	1,325	57	2450	87	1,327	55	2443	91	1,407	58	2445	90
LEP	3,554	16	2361	70	4,011	18	2361	76	4,153	18	2360	76
Special Education	4,332	17	2357	78	4,490	16	2349	78	4,871	16	2355	78
Grade 4												
All Students	38,450	56	2480	96	39,228	54	2477	96	38,376	55	2479	97
Female	18,805	59	2490	94	19,281	58	2487	93	18,646	59	2488	95
Male	19,645	52	2471	97	19,947	50	2468	97	19,730	52	2470	99
African American	4,955	31	2427	87	4,939	32	2428	88	4,854	34	2431	90
AmerIndian/Alaskan	102	42	2446	98	86	47	2465	84	105	41	2451	85
Asian	1,996	74	2526	91	2,109	76	2530	88	2,010	75	2525	89
Hispanic/Latino	9,383	33	2430	89	10,078	33	2430	90	10,195	35	2432	93
Pacific Islander	29	55	2486	89	42	43	2457	92	37	65	2502	93
White	20,825	70	2511	85	20,623	67	2506	86	19,781	68	2509	87
Two or More Races	1,160	59	2493	95	1,351	58	2489	92	1,394	59	2490	100
LEP	2,962	14	2384	78	3,372	15	2386	80	3,776	18	2392	83
Special Education	4,934	17	2390	84	5,006	17	2389	85	5,174	17	2388	86

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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

Group	2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5												
All Students	39,010	59	2517	97	38,748	56	2512	100	39,594	58	2517	98
Female	19,273	64	2531	94	19,028	61	2524	97	19,454	63	2528	95
Male	19,737	53	2504	98	19,720	52	2501	102	20,140	54	2506	100
African American	4,840	33	2461	90	5,019	31	2454	91	5,034	36	2467	90
AmerIndian/Alaskan	112	54	2501	95	104	38	2480	97	82	55	2489	100
Asian	2,003	77	2563	89	1,992	75	2564	96	2,109	79	2571	90
Hispanic/Latino	9,201	37	2467	92	9,580	34	2461	93	10,458	38	2470	94
Pacific Islander	43	63	2525	109	29	69	2526	74	49	43	2495	101
White	21,826	72	2547	86	20,830	71	2544	89	20,476	72	2547	87
Two or More Races	985	62	2528	96	1,194	62	2526	96	1,386	63	2529	95
LEP	2,694	13	2411	75	2,779	9	2400	76	3,186	13	2410	79
Special Education	5,070	17	2420	84	5,464	16	2416	86	5,520	18	2423	86
Grade 6												
All Students	39,071	55	2536	98	39,180	54	2534	98	39,019	54	2534	101
Female	18,963	60	2548	95	19,355	59	2547	95	19,152	59	2546	97
Male	20,108	50	2525	100	19,825	49	2522	99	19,866	50	2522	103
African American	4,881	31	2482	91	4,889	31	2483	89	5,034	32	2484	92
AmerIndian/Alaskan	95	47	2527	94	105	47	2521	94	119	36	2498	99
Asian	1,990	73	2580	90	1,980	74	2585	91	1,931	77	2591	93
Hispanic/Latino	8,794	31	2481	94	9,438	31	2481	94	9,938	32	2482	95
Pacific Islander	32	50	2541	105	44	45	2523	107	32	56	2533	91
White	22,299	68	2565	87	21,699	67	2564	87	20,706	68	2565	89
Two or More Races	980	56	2542	95	1,025	57	2547	95	1,259	58	2542	99
LEP	2,112	6	2411	75	2,315	5	2406	73	2,502	6	2406	73
Special Education	5,193	15	2438	87	5,415	14	2438	84	5,839	15	2436	89

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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

Group	2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7												
All Students	40,085	55	2559	100	39,212	55	2556	102	39,391	55	2556	104
Female	19,410	61	2573	96	19,056	60	2568	99	19,421	61	2572	100
Male	20,675	50	2546	101	20,156	50	2544	104	19,970	49	2541	107
African American	4,917	29	2502	89	4,933	30	2499	96	4,895	31	2501	97
AmerIndian/Alaskan	113	43	2537	95	100	46	2539	96	95	52	2544	107
Asian	1,994	77	2613	91	1,982	74	2607	95	1,942	76	2612	94
Hispanic/Latino	8,836	32	2505	95	8,956	32	2501	99	9,757	33	2502	101
Pacific Islander	43	56	2555	117	34	59	2574	111	46	59	2560	122
White	23,119	67	2587	89	22,182	68	2586	90	21,546	68	2588	92
Two or More Races	1,063	59	2566	101	1,025	56	2561	99	1,110	57	2564	104
LEP	2,074	5	2430	71	2,110	5	2421	77	2,410	5	2421	79
Special Education	5,232	15	2460	86	5,368	15	2455	91	5,632	15	2454	92
Grade 8												
All Students	39,351	55	2574	100	40,139	54	2569	103	39,427	56	2575	103
Female	19,157	62	2589	96	19,440	60	2585	98	19,178	62	2591	99
Male	20,194	49	2559	102	20,699	48	2554	104	20,245	50	2560	104
African American	5,068	32	2520	92	4,978	30	2513	94	4,932	33	2522	95
AmerIndian/Alaskan	94	44	2556	93	108	44	2544	92	98	38	2546	96
Asian	1,925	76	2626	93	1,973	76	2627	94	1,975	76	2629	95
Hispanic/Latino	8,546	33	2519	95	9,068	32	2516	99	9,258	34	2522	98
Pacific Islander	26	58	2585	106	41	61	2590	100	37	62	2595	109
White	22,770	67	2601	90	22,921	65	2597	93	22,056	69	2605	92
Two or More Races	922	59	2582	100	1,050	57	2578	102	1,071	56	2581	102
LEP	1,791	4	2436	68	1,857	3	2428	71	2,112	5	2437	72
Special Education	5,171	15	2473	85	5,358	14	2470	89	5,557	16	2476	89

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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

Group	2014–2015				2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	38,249	48	2427	80	38,870	53	2438	81	38,016	53	2439	83	37,472	54	2440	84
Female	18,701	47	2426	77	19,109	52	2438	78	18,464	53	2439	79	18,393	53	2439	81
Male	19,548	49	2428	83	19,761	53	2439	84	19,552	54	2440	86	19,079	55	2442	87
African American	4,943	21	2379	71	4,860	27	2391	75	4,826	29	2393	77	4,751	30	2395	79
AmerIndian/Alaskan	111	36	2406	85	90	51	2431	77	96	42	2417	67	110	45	2427	77
Asian	1,961	71	2477	80	2,147	78	2491	76	2,042	76	2490	78	2,024	79	2496	77
Hispanic/Latino	9,176	24	2385	73	9,833	31	2398	75	9,817	33	2401	77	10,270	33	2400	78
Pacific Islander	32	34	2416	70	46	46	2421	77	33	52	2441	77	46	50	2441	72
White	20,829	62	2453	71	20,569	67	2463	72	19,881	66	2464	74	18,866	68	2467	74
Two or More Races	1,197	49	2433	79	1,325	56	2446	77	1,321	58	2448	83	1,405	56	2448	84
LEP	3,117	11	2358	68	3,546	20	2377	70	4,005	24	2385	75	4,158	24	2380	77
Special Education	4,384	15	2350	80	4,324	18	2360	82	4,484	18	2361	81	4,865	19	2361	83
Grade 4																
All Students	38,829	44	2470	80	38,387	48	2478	82	39,162	50	2482	85	38,307	51	2484	85
Female	19,180	43	2469	76	18,773	47	2476	78	19,254	49	2480	81	18,618	50	2482	80
Male	19,649	45	2471	84	19,614	49	2480	86	19,908	51	2483	89	19,689	52	2485	90
African American	4,783	17	2419	70	4,938	21	2427	72	4,927	25	2432	78	4,839	26	2434	79
AmerIndian/Alaskan	115	34	2452	74	102	36	2450	87	86	43	2474	74	104	42	2462	80
Asian	2,002	70	2523	79	1,992	73	2533	82	2,106	77	2543	78	2,007	78	2541	78
Hispanic/Latino	8,929	21	2426	72	9,372	24	2434	74	10,055	29	2439	79	10,178	30	2443	79
Pacific Islander	41	46	2468	96	29	55	2488	77	41	46	2465	85	37	49	2491	88
White	21,971	57	2494	71	20,794	62	2504	72	20,598	64	2508	75	19,747	65	2511	75
Two or More Races	988	46	2480	83	1,160	51	2488	81	1,349	53	2491	82	1,395	53	2491	87
LEP	2,942	11	2400	70	2,954	12	2405	69	3,370	15	2411	73	3,773	19	2418	76
Special Education	4,695	11	2392	76	4,916	13	2401	75	4,998	15	2402	80	5,169	16	2402	82

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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

Group	2014–2015				2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	39,044	37	2493	87	38,941	41	2501	89	38,656	43	2505	93	39,540	45	2510	92
Female	18,980	35	2492	83	19,242	40	2500	86	18,990	42	2504	89	19,439	44	2510	89
Male	20,064	38	2495	91	19,699	42	2502	93	19,666	44	2506	96	20,101	46	2510	96
African American	4,889	11	2434	75	4,830	14	2440	77	4,994	16	2445	81	5,031	19	2453	82
AmerIndian/Alaskan	96	20	2468	69	112	32	2488	84	101	32	2480	89	82	29	2488	78
Asian	2,019	60	2547	87	1,999	68	2562	87	1,987	70	2570	90	2,107	74	2577	85
Hispanic/Latino	8,550	15	2444	78	9,173	18	2452	80	9,545	21	2458	83	10,442	24	2466	85
Pacific Islander	30	33	2499	85	43	37	2511	103	29	48	2506	83	49	33	2475	99
White	22,499	49	2520	77	21,798	54	2530	79	20,805	57	2535	82	20,449	59	2539	82
Two or More Races	961	35	2498	86	986	43	2512	91	1,195	46	2515	93	1,380	48	2520	90
LEP	2,586	5	2410	70	2,688	6	2415	69	2,770	7	2417	72	3,188	9	2425	77
Special Education	4,958	7	2409	77	5,055	9	2416	78	5,453	10	2418	82	5,511	12	2422	82
Grade 6																
All Students	39,870	37	2513	100	38,965	41	2521	104	39,031	44	2526	106	38,946	44	2527	107
Female	19,372	37	2516	94	18,921	41	2523	99	19,287	44	2530	101	19,115	45	2531	102
Male	20,498	37	2511	105	20,044	41	2519	108	19,744	43	2523	111	19,830	43	2523	112
African American	4,841	12	2449	88	4,860	14	2452	95	4,864	18	2461	97	5,020	19	2464	100
AmerIndian/Alaskan	121	21	2483	92	95	31	2499	94	103	37	2511	102	118	31	2495	107
Asian	1,979	65	2584	95	1,988	66	2588	99	1,976	71	2602	99	1,929	73	2608	100
Hispanic/Latino	8,577	15	2456	95	8,769	17	2461	97	9,397	20	2467	100	9,918	22	2472	101
Pacific Islander	40	53	2537	111	32	41	2530	117	44	39	2524	126	32	47	2532	93
White	23,299	48	2542	86	22,243	53	2553	89	21,627	57	2559	92	20,674	58	2561	92
Two or More Races	1,013	39	2520	100	978	40	2525	101	1,020	45	2538	102	1,255	46	2536	108
LEP	2,230	4	2402	88	2,107	4	2402	86	2,307	5	2405	88	2,495	5	2407	88
Special Education	5,042	7	2408	95	5,158	7	2412	96	5,391	8	2413	97	5,832	9	2415	100

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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

Group	2014–2015				2015–2016				2016–2017				2017–2018			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 7																
All Students	39,001	39	2530	106	39,961	42	2538	108	39,033	43	2541	111	39,265	44	2542	113
Female	18,952	38	2532	101	19,352	42	2540	102	18,969	42	2542	106	19,382	45	2546	110
Male	20,049	39	2528	111	20,609	42	2536	112	20,064	43	2541	115	19,883	44	2539	117
African American	5,026	14	2466	94	4,895	14	2467	95	4,906	16	2469	97	4,873	18	2473	100
AmerIndian/Alaskan	88	18	2491	92	113	29	2509	89	100	27	2508	102	95	38	2521	112
Asian	1,901	68	2605	101	1,988	71	2617	103	1,983	70	2618	106	1,939	73	2628	106
Hispanic/Latino	8,270	16	2468	98	8,798	19	2477	101	8,883	20	2479	102	9,719	21	2481	104
Pacific Islander	25	32	2525	101	43	44	2546	119	33	48	2569	122	46	39	2550	141
White	22,816	50	2560	93	23,063	54	2569	93	22,106	56	2575	97	21,486	58	2578	99
Two or More Races	875	40	2537	103	1,061	44	2544	108	1,022	40	2540	109	1,107	44	2550	113
LEP	2,053	4	2412	87	2,057	5	2415	89	2,091	5	2416	88	2,405	5	2417	91
Special Education	4,957	7	2421	93	5,189	9	2427	99	5,334	9	2430	97	5,607	9	2427	99
Grade 8																
All Students	39,764	37	2541	114	39,181	40	2551	116	39,955	42	2554	120	39,294	43	2558	120
Female	19,282	38	2546	108	19,069	42	2557	110	19,350	43	2560	114	19,100	44	2564	115
Male	20,482	36	2536	120	20,112	39	2546	121	20,605	40	2549	125	20,190	42	2553	125
African American	5,073	12	2468	94	5,043	15	2479	100	4,950	15	2475	103	4,909	18	2483	105
AmerIndian/Alaskan	106	23	2504	102	94	20	2509	107	109	28	2520	98	98	23	2518	107
Asian	1,791	64	2621	113	1,922	69	2635	113	1,970	72	2645	114	1,975	72	2646	114
Hispanic/Latino	8,203	15	2476	102	8,504	17	2485	103	9,008	19	2489	108	9,209	20	2493	108
Pacific Islander	37	32	2521	112	26	31	2551	127	41	59	2593	116	37	57	2589	127
White	23,706	48	2573	104	22,679	52	2585	104	22,831	54	2589	107	21,997	56	2595	107
Two or More Races	848	35	2543	112	913	43	2559	115	1,046	43	2561	123	1,069	43	2563	118
LEP	1,935	4	2416	89	1,779	3	2419	85	1,845	4	2418	90	2,101	4	2426	89
Special Education	4,921	6	2429	94	5,131	7	2437	95	5,297	8	2438	101	5,527	8	2438	100

Note: AmerIndian or Alaskan = American Indian or Alaskan Native, Pacific Islander = Native Hawaiian/Pacific Islander

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
Grade 3											
All Students	37,525	79	89	70	66	88	71	83	59	55	82
Female	18,417	78	88	70	66	88	70	81	59	55	83
Male	19,108	79	89	70	66	88	71	84	59	55	82
African American	4,764	79	90	70	66	85	71	85	59	54	77
AmerIndian/Alaskan	110	77	88	69	68	83	69	83	56	58	75
Asian	2,022	81	88	69	66	91	73	79	56	56	87
Hispanic/Latino	10,287	79	90	70	65	84	71	85	59	54	75
Pacific Islander	46	77	82*	72	67	90	68	69*	64	55	85
White	18,889	78	86	70	66	88	70	77	58	55	84
Two or More Races	1,407	79	89	70	66	90	72	83	58	56	84
LEP	4,153	81	91	70	66	81	74	87	59	54	65
Special Education	4,871	83	92	69	65	84	77	89	58	53	73
Grade 4											
All Students	38,376	77	89	60	62	88	69	83	47	52	82
Female	18,646	77	88	60	62	88	69	81	48	51	83
Male	19,730	77	90	60	62	87	70	85	47	52	81
African American	4,854	78	90	60	62	83	70	86	48	51	74
AmerIndian/Alaskan	105	74	90	60	63	87	65	79	50	54	72
Asian	2,010	79	86	59	62	91	72	77	45	53	86
Hispanic/Latino	10,195	78	91	60	62	84	70	86	48	51	74
Pacific Islander	37	78	87*	60*	62	89	71	83*	41*	55	84
White	19,781	77	86	60	62	88	69	77	47	52	83
Two or More Races	1,394	78	88	60	62	89	71	82	47	51	85
LEP	3,776	81	92	60	62	80	75	89	48	50	62
Special Education	5,174	83	93	60	62	83	78	91	47	50	71
Grade 5											
All Students	39,594	78	89	64	72	86	70	83	52	63	79
Female	19,454	78	89	64	72	86	70	81	52	62	80
Male	20,140	79	90	64	72	86	71	84	52	63	78
African American	5,034	78	91	64	72	80	70	86	53	63	69
AmerIndian/Alaskan	82	79	95	67	70	79	72	90	49	62	74
Asian	2,109	81	88	64	72	89	74	78	51	62	85
Hispanic/Latino	10,458	79	90	64	73	82	71	86	53	63	72
Pacific Islander	49	79	91	65	74	82	70	86	58	57	74
White	20,476	78	87	64	72	86	70	77	51	63	80
Two or More Races	1,386	78	88	64	72	67	70	81	52	63	81
LEP	3,186	84	92	64	72	72	78	90	53	59	53
Special Education	5,520	83	92	63	72	82	77	90	52	60	69

*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
Grade 6											
All Students	39,019	78	89	68	74	85	70	83	57	65	77
Female	19,152	78	88	68	74	85	70	80	57	65	78
Male	19,866	79	90	68	74	84	71	84	57	65	75
African American	5,034	79	90	68	73	81	71	85	58	64	68
AmerIndian/Alaskan	119	81	91	70	73	91	73	86	61	63	79
Asian	1,931	81	87	70	74	88	73	79	56	65	83
Hispanic/Latino	9,938	80	91	68	74	81	72	86	58	65	67
Pacific Islander	32	77	83*	73*	73	80*	68	79*	60*	63	73*
White	20,706	78	87	68	74	85	69	77	56	66	78
Two or More Races	1,259	79	87	69	74	85	70	81	56	65	78
LEP	2,502	88	94	68	72	80	83	92	57	56	54
Special Education	5,839	85	93	68	73	81	79	90	57	61	68
Grade 7											
All Students	39,391	78	89	67	75	84	70	83	56	67	75
Female	19,421	78	88	67	75	85	69	80	55	67	76
Male	19,970	79	90	67	75	83	71	84	56	67	74
African American	4,895	79	90	67	75	79	71	86	57	65	63
AmerIndian/Alaskan	95	77	92	64	70	90	70	85	53	64	77
Asian	1,942	80	88	67	75	87	72	78	55	66	82
Hispanic/Latino	9,757	80	90	67	75	81	72	86	57	66	67
Pacific Islander	46	83	89	67*	79	87	76	89	49*	71	84
White	21,546	77	87	66	75	84	69	77	54	68	76
Two or More Races	1,110	79	89	67	76	86	70	81	56	67	78
LEP	2,410	88	93	66	73	76*	83	92	54	56	52*
Special Education	5,632	84	92	67	74	81	77	89	56	62	63
Grade 8											
All Students	39,427	79	88	70	77	84	71	81	59	70	75
Female	19,178	79	87	70	77	84	71	79	59	70	76
Male	20,245	79	89	70	77	82	71	83	60	70	73
African American	4,932	80	89	70	78	79	72	84	60	69	66
AmerIndian/Alaskan	98	77	90	69	78	73	69	79	64	64	66
Asian	1,975	81	88	69	77	87	73	79	58	69	81
Hispanic/Latino	9,258	80	89	71	77	80	72	85	60	69	67
Pacific Islander	37	80	81*	67*	78	88	73	77*	51*	69	86
White	22,056	79	86	70	77	84	70	76	59	70	76
Two or More Races	1,071	79	87	71	77	84	71	78	61	69	78
LEP	2,112	88	93	71	74	82*	83	92	58	58	62*
Special Education	5,557	84	91	70	77	81	77	88	59	66	67

*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
Grade 3											
All Students	37,472	83	90	74	80	90	77	85	64	72	85
Female	18,393	83	90	74	80	89	76	84	64	72	84
Male	19,079	84	91	74	79	90	77	86	64	72	86
African American	4,751	84	92	74	78	87	77	88	64	70	79
AmerIndian/Alaskan	110	82	86	76	82	86	75	83	65	72	80
Asian	2,024	86	86	73	80	93	81	78	61	73	91
Hispanic/Latino	10,270	83	91	74	79	87	77	87	64	71	79
Pacific Islander	46	81	79*	76	79	91	74	72*	66	72	90
White	18,866	83	87	74	80	90	76	80	64	72	86
Two or More Races	1,405	83	89	74	79	90	77	84	64	71	86
LEP	4,158	85	92	73	80	85	79	89	63	72	74
Special Education	4,865	88	94	73	79	87	83	92	63	70	80
Grade 4											
All Students	38,307	84	90	80	79	90	77	84	73	71	85
Female	18,618	84	89	80	79	89	77	83	73	71	84
Male	19,689	84	91	80	79	90	78	85	72	71	86
African American	4,839	84	91	80	78	87	78	86	73	69	79
AmerIndian/Alaskan	104	83	90	78	83	80	76	85	72	75	69
Asian	2,007	86	89	79	79	93	81	80	70	71	90
Hispanic/Latino	10,178	84	90	80	79	86	77	86	73	70	79
Pacific Islander	37	85	94*	81	78*	94*	79	76*	79	70*	88*
White	19,747	84	88	80	79	90	77	78	72	72	86
Two or More Races	1,395	84	90	79	78	91	77	82	73	69	87
LEP	3,773	86	92	80	78	86	79	88	73	68	75
Special Education	5,169	88	93	80	77	89	82	91	72	68	80
Grade 5											
All Students	39,540	83	90	77	71	90	76	85	68	61	86
Female	19,439	82	90	77	71	90	76	84	69	61	85
Male	20,101	83	91	77	71	90	77	87	68	61	86
African American	5,031	85	92	78	71	86	79	89	68	60	77
AmerIndian/Alaskan	82	83	87	82	68	92	76	81	74	60	83
Asian	2,107	86	88	78	72	93	80	80	68	61	92
Hispanic/Latino	10,442	84	91	77	71	86	77	88	68	60	79
Pacific Islander	49	83	95	78	70	84*	75	91	70	58	69*
White	20,449	82	88	77	71	90	75	80	68	62	86
Two or More Races	1,380	83	89	78	71	92	76	84	69	61	87
LEP	3,188	87	93	76	71	83	82	91	67	57	74
Special Education	5,511	88	94	76	71	87	84	92	65	60	79

*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	Al	L1	L2	L3	L4
Grade 6											
All Students	38,946	83	92	78	72	90	77	87	70	62	85
Female	19,115	83	91	78	72	89	76	86	70	62	84
Male	19,830	84	92	78	72	90	77	88	70	62	85
African American	5,020	85	93	77	72	85	79	90	70	60	76
AmerIndian/Alaskan	118	87	92	81	74	92	80	91	68	64	89
Asian	1,929	86	88	78	72	95	81	80	69	62	92
Hispanic/Latino	9,918	85	93	78	72	87	79	90	70	61	78
Pacific Islander	32	79	89*	77	70	91*	71	82*	67	65	76*
White	20,674	82	90	78	72	89	74	82	70	63	85
Two or More Races	1,255	84	92	79	73	90	77	87	71	62	86
LEP	2,495	91	95	77	72	90	87	94	66	58	75
Special Education	5,832	90	95	77	72	84	86	94	68	58	76
Grade 7											
All Students	39,265	83	91	76	74	90	77	86	67	65	86
Female	19,382	83	91	76	74	90	76	85	67	65	85
Male	19,883	84	92	76	75	91	77	88	67	65	86
African American	4,873	85	93	76	73	86	79	90	66	62	77
AmerIndian/Alaskan	95	82	89	71	73	95	76	86	59	67	84
Asian	1,939	86	89	74	75	94	81	81	65	65	92
Hispanic/Latino	9,719	85	93	76	74	87	79	89	67	64	78
Pacific Islander	46	85	89	76	74*	94	79	84	69	53*	94
White	21,486	82	89	76	75	90	75	81	67	66	86
Two or More Races	1,107	83	89	76	74	92	77	84	68	64	88
LEP	2,405	91	95	76	72	88	87	94	64	56	81
Special Education	5,607	90	95	75	73	87	86	94	64	62	77
Grade 8											
All Students	39,294	82	90	71	72	91	75	85	61	61	86
Female	19,100	81	89	72	72	90	74	84	62	62	85
Male	20,190	83	91	71	72	91	76	87	61	61	87
African American	4,909	85	92	72	71	85	78	90	60	59	77
AmerIndian/Alaskan	98	82	90	67	76	92	75	86	60	58	86
Asian	1,975	85	87	71	72	94	80	79	61	62	92
Hispanic/Latino	9,209	84	92	71	71	87	77	88	61	59	80
Pacific Islander	37	84	92*	69*	74*	92	77	84*	58*	66*	89
White	21,997	81	87	71	72	91	73	80	62	62	86
Two or More Races	1,069	82	90	73	71	92	75	85	62	61	87
LEP	2,101	91	95	70	74	85	87	94	56	55	77
Special Education	5,527	89	94	70	71	88	85	93	58	58	79

*The classification index is based on n<10.