Connecticut Smarter Balanced Summative Assessments

2018–2019 Technical Report

CSDE

CONNECTICUT STATE
DEPARTMENT OF EDUCATION

Submitted to
Connecticut State Department of Education
by the 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 alternate assessment based on alternate academic achievement standards (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.** The CAT is an online adaptive test that provides an individualized assessment for each student.
- **Performance Task.** A PT is a task that challenges students to apply their knowledge and skills to respond to real-world problems. PTs can best be described as collections of 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 PT 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 PTs in ELA/L while keeping PTs in mathematics assessment. For the paper-pencil tests, the test booklet will include both non-PT and PT components, but only the non-PT component will be scored for ELA/L.
- Reported scores for combining claim 2 (writing) and 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).** ICAs test the same content and report scores on the same scale as the summative assessments.
- **Interim Assessment Blocks (IABs).** IABs focus on smaller sets of related concepts and provide more detailed information about student learning.

This report provides a technical summary of the 2018–2019 summative assessments in ELA/L and mathematics administered in grades 3–8 under the Connecticut Smarter Balanced assessments. The report includes eight chapters: Overview; Test Administration; Summary of the 2018–2019 Operational Test Administration; Validity; Reliability; Scoring; Reporting and Interpreting Scores; and Quality Control Procedures. The data included in this report are based on Connecticut data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs and a summary of their performance are provided in Appendix A.

While this report includes information on all aspects of the technical quality of the Smarter Balanced test administration for Connecticut, it is an addendum to the 2018–2019 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], & 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* (U.S. Department of Education, 2015). 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 2018–2019 Smarter Balanced assessments testing window spanned approximately two and a half months for the summative assessments and eight months for the interim assessments. The paper-pencil fixed-form tests for summative assessments were administered concurrently during the online summative window. Table 1 shows the testing windows for both online and paper-pencil assessments.

Table 1. 2018–2019 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	3/25/2019	6/7/2019	Online Adaptive
	3–8	3/25/2019	6/7/2019	Paper-Pencil Fixed Forms
Interim Comprehensive Assessments	3–8, 11	9/26/2018	6/14/2019	Online Fixed Forms
Interim Assessment Blocks	3–8, 11	9/26/2018	6/14/2019	Online Fixed Forms

2.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The Smarter Balanced assessments are administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the Smarter Balanced assessments, several assessment options were available for the 2018–2019 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2018–2019. 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. 2018–2019 Testing Options

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

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

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

2.2.1 Administrative Roles

The key personnel involved with the test administration for the Connecticut State Department of Education (CSDE) are District Administrators (DAs), District Test Coordinators (DTCs), School Test Coordinators (STCs), Teachers (TEs), and Test Administrators (TAs). 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 DA may add users with DTC roles in the Test Information Distribution Engine (TIDE). For example, a director of special education may need DTC privileges in TIDE to access district-level data for the purposes of verifying test settings for designated supports and accommodations. DAs have the same test administration responsibilities as DTCs. Their primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

District Test Coordinator

The DTC is primarily responsible for coordinating the administration of the Smarter Balanced 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 AIR Secure Browser, are properly installed and functional
- Importing users (including STCs, TEs, and TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all STCs, TEs, TAs, and TCs
- Ensuring that all personnel are trained on how to administer the Smarter Balanced assessments properly
- Monitoring the secure administration of the tests
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE and Smarter Balanced policies

School Test Coordinator

The STC is primarily responsible for coordinating the administration of the Smarter Balanced 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 testing windows, establishing a testing schedule with DTCs, TEs, and TAs

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

Teacher

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

Test Administrator

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

TAs are responsible for the following:

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

2.2.2 Online Test Administration

Within Connecticut's testing window, schools can set testing schedules, allowing students to test in intervals (e.g., multiple sessions) rather than in one long 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), their first name, and the session ID into the Student Interface using computers provided by the school. The TE or TA then verifies that the students are taking the appropriate assessments with the appropriate accessibility features. (See Section 2.6 for a list of accommodations.) Students can begin testing only when the TA or TE confirms the settings. The TA or TE then reads the *Directions for Administration* in the *Online Smarter Balanced Test Administration Manual* aloud to the students and guides them through the login process.

Once an assessment has started, the student must answer all the test 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 (PTs). The same rules that apply to the CAT for reviews and changes to responses also apply to PTs.

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

During a test session, TEs or TAs may pause the test for a student or group of students to take a break. It is up to the TEs or TAs to determine an appropriate stopping point; however, to ensure the integrity of test scores or testing, the CAT cannot be paused for more than 30 minutes for 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 always remain in the room during a test session to monitor student testing. Once the test session ends, the TAs or TEs must ensure that each student has successfully logged out of the system. Then the TAs or TEs must collect and send for secure shredding any handouts or scratch paper that students used during the assessment.

2.2.3 Paper-Pencil Test Administration

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

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

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

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

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

Table 3. Number of Students Who Took Paper-Pencil Tests in the 2018–2019 Summative Test Administration

Subject	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
ELA/L	8	6	*	*	*	6	31
Mathematics	8	6	*	*	*	*	30

*This amount is suppressed to protect student confidentiality.

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 2018–2019 test administration, Smarter Balanced added the Braille Hybrid Adaptive Test (Braille HAT) for mathematics. The Braille HAT consists of a fixed-form segment, a CAT segment, and a fixed-form PT. The fixed-form segment includes items with tactile graphics which can be embossed at the testing location or received as a package of pre-embossed materials through the CSDE. All items on the Braille HAT can be presented to the students using a Refreshable Braille Display (RBD).

The braille interface is described as follows:

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

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

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

2.3 Training and Information for Test Coordinators and Administrators

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

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

Office Hour Webinars

During the testing window, the CSDE and 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 2018–2019 school year. In the fall of 2018, Burmese and Illustration glossaries 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. 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 an opportunity to practice the PT.

The training tests are designed to provide students and teachers with opportunities to quickly familiarize themselves with the software and navigational tools they will use for the upcoming Smarter Balanced assessments for ELA/L and mathematics. Training tests are available for both ELA/L and mathematics, and the tests are organized by grade bands (grades 3–5, grades 6–8, and grade 11), with each test containing 5–10 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 2019 Smarter Balanced assessments in ELA/L and mathematics.

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

The *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 AIR 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 Online Reporting System (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 test delivery system (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 *AIRWays User Guide* provides instructions and support for users viewing student interim assessment performance reports in AIRWays and scoring interim items.

All manuals and user guides pertaining to the 2018–2019 online testing are available on the portal, and DAs, DTCs, and STCs used the manuals and user guides to train TAs and TEs in test administration policies and procedures.

Brochures and Quick Guides

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

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

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

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

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

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

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

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

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

User Role Permissions for Online Systems Brochure: This brochure outlines the user roles and permissions for each secure online testing system used to administer the online assessments for the Connecticut

Comprehensive Assessment Program. These systems include TIDE, ORS, TA Interface, DEI, Teacher Hand Scoring System (THSS), AVA, and the AIRWays Reporting System.

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

2.3.2 District Test Coordinator Training Workshops

DTC training workshops were held on January 23–25, 2019, at the Institute of Technology and Business Development (ITBD) in New Britain, Connecticut. 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 secure file transfer site to the online testing system during the testing period.

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

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

2.4.2 System Security

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

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

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

AIR Secure Browser: A key role of the TC is to ensure that the AIR Secure Browser is properly installed on the computers used for the administration of the online assessments. Developed by the testing contractor, the AIR Secure Browser prevents students from accessing other computers or Internet applications and from copying test information. The AIR 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 AIR 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 PTs, different forms are distributed throughout a classroom so that students receive different forms of the PTs.

After the Test

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

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

2.4.4 Test Security Violations

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

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

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

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

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

2.5 STUDENT PARTICIPATION

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

2.5.1 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

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

2.6 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

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

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

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

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

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

For additional information about the availability of designated supports and accommodations, refer to the Connecticut's Assessment Guidelines for complete information at this URL:

https://ct.portal.airast.org/core/fileparse.php/51/urlt/CSDE-1819-Assessment-Guidelines.pdf

2.6.1 Online Universal Tools for All Students

Universal tools are access features of an assessment or exam that are embedded or non-embedded components of the test administration system. Universal tools are available to all students based on their preference and selection and have been pre-set in TIDE. In the 2018–2019 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 Dictionary: An English dictionary is available for the full-write portion of an ELA/L PT. A full write is the second part of a PT.

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

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

*Global Notes***.** Global notes is a notepad that is available for ELA/L PT in which students complete a full write. The student clicks the notepad icon for the notepad to appear. During the ELA/L PT, the notes are retained from segment to segment so that the student may go back to the notes even though he or she cannot go back 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 mathematics items. They are available only with the specific items for which the Smarter Balanced item specifications indicate that one or more of these tools would be appropriate.

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

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

2.6.2 Designated Supports and Accommodations

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

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

Embedded Designated Supports

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

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 TDS to be pre-set for the entire test. This support is generally most beneficial for students with visual disabilities. Selections are entered in the TIDE system prior to testing.

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

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

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

Translations (glossaries) (for mathematics): Translated glossaries are a language support provided for selected construct-irrelevant terms for mathematics. Translations for these terms appear on the computer screen when students click on them. The following language glossaries were offered: Arabic, 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.

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 stimuli items): This is printed text that appears on the computer screen as audio materials are presented.

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.

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

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 who need a paper copy of a passage or stimulus, permission for the students to request printing must first be set in TIDE.

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

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

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 2018–2019 administration. Tables 5–10 provide the number of students who were offered the accommodations and designated supports.

Table 4. 2018–2019 Universal Tools, Designated Supports, and Accommodations

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

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

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

[2] For ELA/L PT 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 grade 4 and above mathematics tests

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

[14] For ELA/L reading passages, all grades

[15] For ELA/L PT 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	*	*	*	10	*	*
Braille	*			*		
Closed Captioning	16	19	25	36	29	29
Text-to-Speech: Passages and Items	1,098	1,090	1,116	933	962	797
Non-Embedded Accommodations						
Alternate Response Options	7	*	7	*	*	*
Speech-to-Text	144	190	164	149	113	65

*This amount is suppressed to protect student confidentiality.

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	9	17	15	15	21	19
	LEP		*	*	*		
	Special Ed	6	9	9	8	13	14
Masking	Overall	140	126	135	91	93	89
	LEP	25	26	20	11	6	11
	Special Ed	100	101	108	83	80	80
Mouse Pointer	Overall		*	*			
	LEP			*			
	Special Ed		*	*			
Print Size	Overall	24	24	39	31	18	16
	LEP	*	*	6	*	*	*
	Special Ed	15	14	31	22	7	8
Streamline	Overall	210	176	178	111	128	98
	LEP	36	31	22	9	23	17
	Special Ed	132	122	120	98	125	92
Text-to-Speech: Items	Overall	6,231	6,228	5,886	4,233	3,735	3,330
	LEP	2,663	2,482	2,148	1,393	1,192	1,063
	Special Ed	2,105	2,484	2,501	2,267	2,012	1,668

*This amount is suppressed to protect student confidentiality.

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	*	7	7	*	*	*
	LEP						
	Special Ed	*	*	*		*	*
Color Overlay	Overall	*	8	*	7	8	8
	LEP		*				
	Special Ed	*	7	*	6	*	7
Magnification	Overall	*	6	*	9	7	6
	LEP	*		*	*		
	Special Ed	*	*	*	*	*	*
Noise Buffers	Overall	16	9	8	9	*	*
	LEP	*	*		*		
	Special Ed	11	7	*	6	*	*
Read-Aloud Items	Overall	134	118	142	58	47	57
	LEP	34	31	70	23	26	24
	Special Ed	81	83	70	40	26	37
Separate Setting	Overall	3,633	4,006	4,137	3,607	3,484	3,118
	LEP	745	770	732	579	497	428
	Special Ed	2,587	2,982	3,028	2,842	2,785	2,471
Simplified Test Directions	Overall	1,164	695	729	532	501	481
	LEP	327	268	291	227	206	224
	Special Ed	503	451	483	388	370	340
Translated Test Directions	Overall	97	113	139	139	130	125
	LEP	94	112	137	135	129	123
	Special Ed	14	13	19	18	9	14

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	*	*	*	11	*	*
Braille	*	*		*		
Non-Embedded Accommodations						
100s Number Table	113	752	609	286	178	137
Abacus	*	*	*	*	6	*
Alternate Response Options	7	*	7	*	*	*
Calculator	*	*	40	209	335	380
Multiplication Table		1,999	2,610	2,539	2,428	1,900
Speech-to-Text	140	184	142	133	96	54

Table 9. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	9	14	15	16	21	19
	LEP			*	*		
	Special Ed	6	6	9	9	13	14
Masking	Overall	140	126	135	86	93	73
	LEP	25	26	20	9	6	8
	Special Ed	99	100	108	78	80	66
Mouse Pointer	Overall		*	*			
	LEP			*			
	Special Ed		*	*			
Print Size	Overall	32	29	45	40	26	24
	LEP	*	*	*	*	*	*
	Special Ed	21	13	34	24	12	13
Streamline	Overall	207	158	176	111	129	98
	LEP	35	24	22	9	24	17
	Special Ed	128	120	122	97	126	92
Text-to-Speech: Items	Overall	100	88	66	125	107	90
	LEP	31	17	20	26	19	21
	Special Ed	42	55	39	73	80	52
Text-to-Speech: Stimuli and Items	Overall	7,808	7,748	7,429	5,663	5,158	4,587
	LEP	2,977	2,783	2,416	1,586	1,426	1,255
	Special Ed	3,305	3,667	3,733	3,292	3,073	2,559
Translation (Glossary): Spanish	Overall	606	643	522	647	647	640
	LEP	602	634	511	637	638	625
	Special Ed	64	86	71	79	85	87
Translation (Glossary): Other Languages	Overall	40	53	52	47	42	41
	LEP	40	52	52	44	42	40
	Special Ed		*	*	*		

*This amount is suppressed to protect student confidentiality.

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	*	6	7	*	*	*
	LEP						
	Special Ed	*	*	*		*	*
Color Overlay	Overall	*	8	*	7	8	8
	LEP		1				
	Special Ed	*	7	*	6	*	7
Magnification	Overall	*	6	6	8	7	7
	LEP	*		*			*
	Special Ed	*	*	*	*	*	*
Noise Buffers	Overall	14	10	9	9	*	*
	LEP	*	*		*		*
	Special Ed	11	8	6	6	*	*
Read Aloud Stimuli and Items	Overall	144	123	132	55	59	88
	LEP	35	32	52	21	31	31
	Special Ed	85	86	78	39	33	64
Read Aloud Stimuli and Items (Spanish)	Overall	65	59	82	28	40	30
	LEP	62	57	81	26	39	29
	Special Ed	12	11	13	*	7	8
Separate Setting	Overall	3,630	4,027	4,144	3,626	3,510	3,143
	LEP	751	785	729	586	508	430
	Special Ed	2,591	3,003	3,035	2,853	2,810	2,489
Simplified Test Directions	Overall	1,155	706	736	545	510	491
	LEP	318	275	283	218	192	217
	Special Ed	507	479	506	401	391	350
Translated Test Directions	Overall	93	96	120	127	128	110
	LEP	91	95	118	123	127	108
	Special Ed	15	13	20	16	14	12
Translation (Glossary): Spanish	Overall	29	26	76	26	45	32
	LEP	26	25	74	26	44	31
	Special Ed	*	6	8	8	13	*
Translation (Glossary): Other Languages	Overall	*	11	8	7	*	*
	LEP	*	9	8	6	*	*
	Special Ed						*

*This amount is suppressed to protect student confidentiality.

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, and test starting and ending times. AIR's TDS captures all 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

Changes in student scores between administration years are examined using a regression model to check for outliers. For these between-year comparisons, students' current-year scores are regressed on their test scores from the previous year and on the number of days between the two years' test-end dates (to control for the instruction time between the two test scores). Between-year comparisons are performed between the current school year (e.g., 2018–2019) and the year before the current school year (e.g., 2017–2018).

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

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

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

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

The variance of average 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, p. 456),

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

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

The QA report includes a list of the flagged aggregate units and 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 analysis 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) and Sotaridona, Pornel, and Vallejo (2003), aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the "asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05" (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using l_z for systematic flagging of aberrant response patterns. Students with $l_z$values 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 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 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 in the next subsection), with backup copies remaining on the drive arrays of the disabled satellite.

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

Hub

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

Demographic and History Servers

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

Quality Assurance System

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

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

- The system’s hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from the system’s data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level are redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching in all server cabinets.
- Data are protected by nightly backups. A full weekly backup and incremental nightly backups protect data. Should a catastrophic event occur, 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 need to rerun it.

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

3. SUMMARY OF 2018–2019 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

All Connecticut students enrolled in grades 3–8 in all public schools are required to participate in the Smarter Balanced ELA/L and mathematics assessments. Tables 11–12 present the demographic composition of Connecticut students who meet the 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	36,516	37,727	38,605	39,588	39,165	39,372
Female	17,890	18,486	18,733	19,412	19,200	19,362
Male	18,626	19,239	19,871	20,175	19,961	20,006
African American	4,603	4,820	4,955	5,069	5,068	4,917
AmerIndian/Alaskan	101	104	111	80	117	100
Asian	1,945	2,015	2,003	2,059	1,922	1,917
Hispanic/Latino	10,122	10,477	10,371	10,575	10,134	9,883
Pacific Islander	29	42	36	45	29	48
White	18,236	18,857	19,683	20,320	20,584	21,345
Two or More Races	1,480	1,412	1,446	1,440	1,311	1,162
LEP	4,287	3,999	3,387	2,710	2,429	2,225
Special Education	5,018	5,443	5,647	5,759	6,086	5,790

Note. African American= Black or African American; AmerIndian/Alaskan= American Indian or Alaska Native; Pacific Islander= Native Hawaiian or Other Pacific Islander

Table 12. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	36,460	37,675	38,514	39,488	39,002	39,216
Female	17,877	18,467	18,690	19,374	19,125	19,290
Male	18,583	19,206	19,823	20,113	19,873	19,922
African American	4,597	4,805	4,940	5,051	5,038	4,890
AmerIndian/Alaskan	101	104	110	81	116	98
Asian	1,944	2,013	1,997	2,055	1,917	1,914
Hispanic/Latino	10,107	10,454	10,344	10,537	10,072	9,811
Pacific Islander	29	42	36	44	29	47
White	18,202	18,848	19,644	20,286	20,525	21,295
Two or More Races	1,480	1,409	1,443	1,434	1,305	1,161
LEP	4,286	3,992	3,375	2,697	2,406	2,202
Special Education	5,028	5,448	5,632	5,725	6,042	5,712

3.2 SUMMARY OF STUDENT PERFORMANCE

Tables 13–16 summarize overall student performance in the 2018–2019 summative test for all students and by subgroups, including the average and the standard deviation of overall scale scores, the percentage of students in each achievement level, and the percentage of proficient students. Figures 1 and 2 show the percentage of proficient students in five years for all students (cohort comparisons). Figures 3 and 4 show the average scale scores in five years for all students. In ELA/L, student performance is compared for four years because ELA/L scores in 2014–2015 were based on both computer-adaptive test (CAT) and performance task (PT) components while ELA/L scores from 2015–2016 were based on the CAT component only. 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.

Table 13. Scale Score Mean, Standard Deviations, & Percent Proficient for Overall and by Subgroup: ELA/L Grades 3–5

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	36,516	2437	91	23	22	23	31	54
Female	17,890	2445	89	21	21	25	34	58
Male	18,626	2429	92	26	23	22	28	51
African American	4,603	2395	86	39	27	19	15	34
AmerIndian/Alaskan	101	2416	83	30	23	30	18	48
Asian	1,945	2481	87	11	16	22	51	73
Hispanic/Latino	10,122	2397	87	38	26	20	16	35
Pacific Islander	29	2413	70	14	41	34	10	45
White	18,236	2464	82	13	20	27	41	67
Two or More Races	1,480	2447	93	22	20	22	36	58
LEP	4,287	2369	79	50	27	15	7	22
Special Education	5,018	2358	80	59	23	12	7	18
Grade 4								
All Students	37,727	2478	99	28	18	23	32	55
Female	18,486	2487	96	24	18	24	34	58
Male	19,239	2470	101	31	18	22	29	51
African American	4,820	2432	91	46	21	19	14	34
AmerIndian/Alaskan	104	2463	87	28	23	34	15	49
Asian	2,015	2530	91	12	12	23	53	76
Hispanic/Latino	10,477	2432	93	45	21	20	15	35
Pacific Islander	42	2476	78	21	24	31	24	55
White	18,857	2510	89	16	16	25	43	68
Two or More Races	1,412	2489	97	23	18	23	35	58
LEP	3,999	2391	84	62	20	13	5	18
Special Education	5,443	2389	87	67	16	11	7	18
Grade 5								
All Students	38,605	2516	100	24	18	30	28	58
Female	18,733	2528	97	20	18	31	31	63
Male	19,871	2506	102	28	18	29	25	54
African American	4,955	2466	94	41	23	24	12	36
AmerIndian/Alaskan	111	2479	96	33	23	30	14	43
Asian	2,003	2568	90	10	12	30	48	78
Hispanic/Latino	10,371	2470	95	40	22	25	13	38
Pacific Islander	36	2526	101	19	14	33	33	67
White	19,683	2547	89	13	15	34	38	72
Two or More Races	1,446	2526	103	23	16	29	32	61
LEP	3,387	2415	80	63	23	12	2	14
Special Education	5,647	2420	88	62	20	13	5	18

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

Table 14. Scale Score Mean, Standard Deviations, & Percent Proficient for Overall and by Subgroup: ELA/L Grades 6–8

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	39,588	2538	99	21	24	33	22	55
Female	19,412	2550	96	17	23	35	25	60
Male	20,175	2526	101	25	24	31	19	51
African American	5,069	2493	91	35	31	26	8	34
AmerIndian/Alaskan	80	2510	86	25	34	31	10	41
Asian	2,059	2597	88	7	14	34	45	79
Hispanic/Latino	10,575	2490	95	37	28	26	9	35
Pacific Islander	45	2507	97	33	27	24	16	40
White	20,320	2567	89	11	20	38	30	68
Two or More Races	1,440	2547	97	18	23	34	25	58
LEP	2,710	2415	75	71	22	6	1	7
Special Education	5,759	2442	86	58	26	13	3	15
Grade 7								
All Students	39,165	2559	105	23	21	35	21	56
Female	19,200	2574	100	18	21	37	24	61
Male	19,961	2546	107	27	22	33	18	51
African American	5,068	2507	98	40	27	26	8	33
AmerIndian/Alaskan	117	2521	98	33	29	29	9	38
Asian	1,922	2619	97	9	14	33	44	77
Hispanic/Latino	10,134	2510	101	38	26	27	8	36
Pacific Islander	29	2573	103	17	24	31	28	59
White	20,584	2591	93	12	18	41	29	70
Two or More Races	1,311	2567	105	20	21	35	24	59
LEP	2,429	2425	78	75	18	6	0	6
Special Education	6,086	2460	93	60	23	14	3	17
Grade 8								
All Students	39,372	2574	104	22	22	36	20	56
Female	19,362	2591	101	17	21	38	24	62
Male	20,006	2558	105	27	24	34	16	50
African American	4,917	2522	96	38	29	27	7	34
AmerIndian/Alaskan	100	2563	102	26	24	33	17	50
Asian	1,917	2635	92	8	14	37	40	78
Hispanic/Latino	9,883	2521	100	38	27	27	8	34
Pacific Islander	48	2574	121	25	17	38	21	58
White	21,345	2605	94	12	19	42	27	69
Two or More Races	1,162	2581	103	18	25	36	21	57
LEP	2,225	2432	69	79	18	3	0	3
Special Education	5,790	2474	88	59	25	13	2	16

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

Table 15. Scale Score Mean, Standard Deviations, & Percent Proficient for Overall and by Subgroup: Mathematics Grades 3–5

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	36,460	2443	86	24	21	29	26	55
Female	17,877	2441	83	24	22	29	25	54
Male	18,583	2445	88	23	21	29	28	56
African American	4,597	2397	79	43	26	21	11	31
AmerIndian/Alaskan	101	2429	76	30	24	28	19	47
Asian	1,944	2497	81	9	11	28	52	79
Hispanic/Latino	10,107	2404	81	39	26	23	12	35
Pacific Islander	29	2432	71	10	31	45	14	59
White	18,202	2470	76	12	19	34	35	69
Two or More Races	1,480	2448	87	22	21	29	28	57
LEP	4,286	2388	79	46	26	20	8	28
Special Education	5,028	2364	82	60	21	13	6	19
Grade 4								
All Students	37,675	2486	87	20	28	27	25	52
Female	18,467	2484	83	20	29	28	23	51
Male	19,206	2489	91	20	26	26	27	54
African American	4,805	2437	80	38	35	20	8	28
AmerIndian/Alaskan	104	2475	72	20	31	37	13	49
Asian	2,013	2547	78	5	15	27	52	80
Hispanic/Latino	10,454	2445	81	35	34	21	10	31
Pacific Islander	42	2480	69	14	38	29	19	48
White	18,848	2515	76	9	24	33	34	67
Two or More Races	1,409	2495	87	18	26	28	28	56
LEP	3,992	2420	79	46	33	15	5	21
Special Education	5,448	2404	83	56	27	12	5	17
Grade 5								
All Students	38,514	2513	94	28	26	20	27	47
Female	18,690	2512	90	27	28	20	25	45
Male	19,823	2513	98	28	24	20	28	48
African American	4,940	2456	85	51	28	12	9	21
AmerIndian/Alaskan	110	2477	84	41	31	14	15	28
Asian	1,997	2578	86	10	16	20	55	75
Hispanic/Latino	10,344	2469	88	45	28	15	11	27
Pacific Islander	36	2509	97	33	19	28	19	47
White	19,644	2543	83	15	24	25	36	61
Two or More Races	1,443	2519	98	27	25	19	30	49
LEP	3,375	2434	79	62	26	9	4	13
Special Education	5,632	2422	85	68	20	7	5	12

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

Table 16. Scale Score Mean, Standard Deviations, & Percent Proficient for Overall and by Subgroup: Mathematics Grades 6–8

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	39,488	2530	109	27	27	21	24	45
Female	19,374	2534	104	25	28	23	24	47
Male	20,113	2527	113	29	27	20	25	44
African American	5,051	2471	101	48	30	14	8	22
AmerIndian/Alaskan	81	2497	94	32	37	19	12	31
Asian	2,055	2616	95	7	16	20	57	78
Hispanic/Latino	10,537	2476	103	45	30	15	9	24
Pacific Islander	44	2511	93	32	34	23	11	34
White	20,286	2564	94	15	26	26	33	59
Two or More Races	1,434	2537	108	26	28	21	25	46
LEP	2,697	2409	92	74	20	4	2	6
Special Education	5,725	2419	102	70	20	6	4	10
Grade 7								
All Students	39,002	2547	115	29	25	22	24	46
Female	19,125	2550	111	28	26	23	24	46
Male	19,873	2544	119	31	24	21	25	46
African American	5,038	2478	104	53	27	13	7	20
AmerIndian/Alaskan	116	2513	109	40	29	21	10	31
Asian	1,917	2636	110	10	14	19	57	76
Hispanic/Latino	10,072	2487	106	49	28	15	9	24
Pacific Islander	29	2567	113	28	17	21	34	55
White	20,525	2584	100	16	24	28	33	61
Two or More Races	1,305	2560	117	27	24	21	29	50
LEP	2,406	2419	88	78	17	4	1	5
Special Education	6,042	2435	101	71	18	7	4	11
Grade 8								
All Students	39,216	2558	123	34	23	19	24	44
Female	19,290	2565	118	31	24	21	25	45
Male	19,922	2552	128	36	22	18	24	42
African American	4,890	2483	106	59	22	13	6	19
AmerIndian/Alaskan	98	2532	120	38	26	19	17	37
Asian	1,914	2653	116	12	15	19	54	74
Hispanic/Latino	9,811	2493	109	56	23	13	8	21
Pacific Islander	47	2575	130	28	34	9	30	38
White	21,295	2597	110	20	23	24	33	57
Two or More Races	1,161	2564	127	33	24	17	26	43
LEP	2,202	2418	83	87	10	3	1	3
Special Education	5,712	2438	101	76	16	5	3	8

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

Figure 1. ELA/L Percent Proficient Across Years

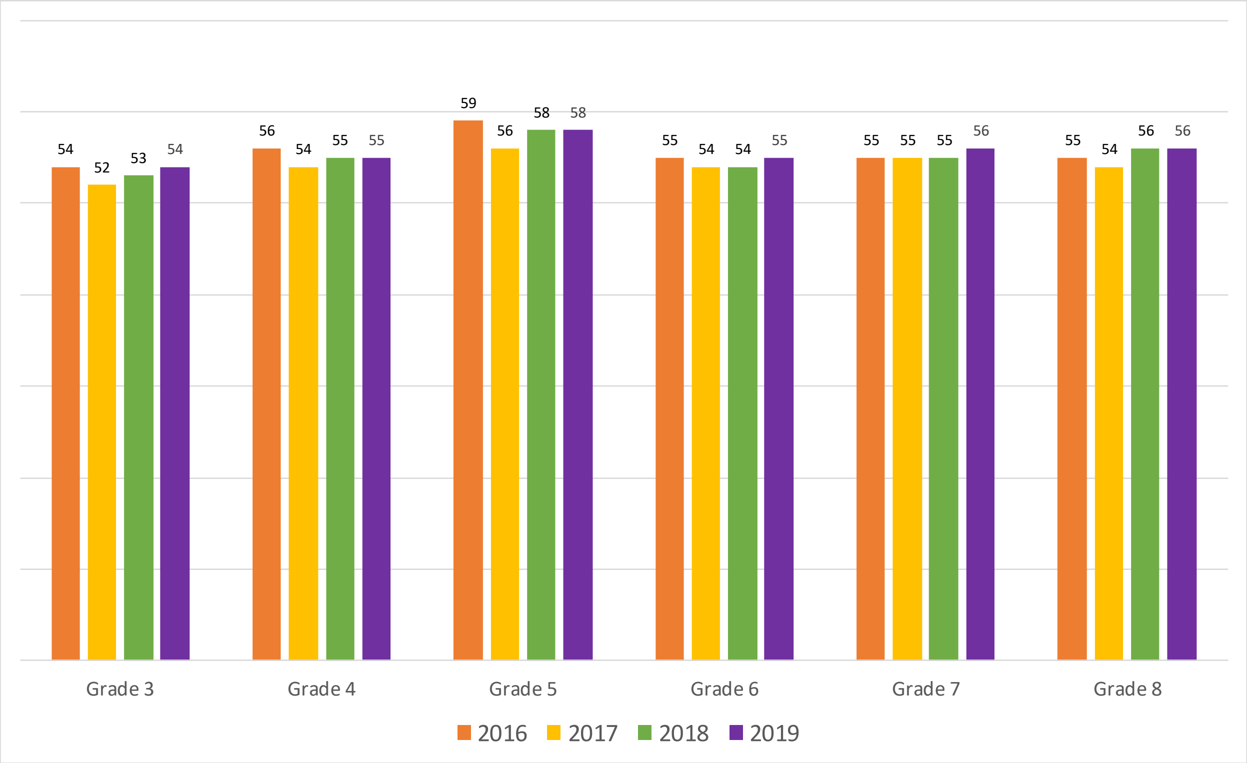


Figure 2. Mathematics Percent Proficient Across Years

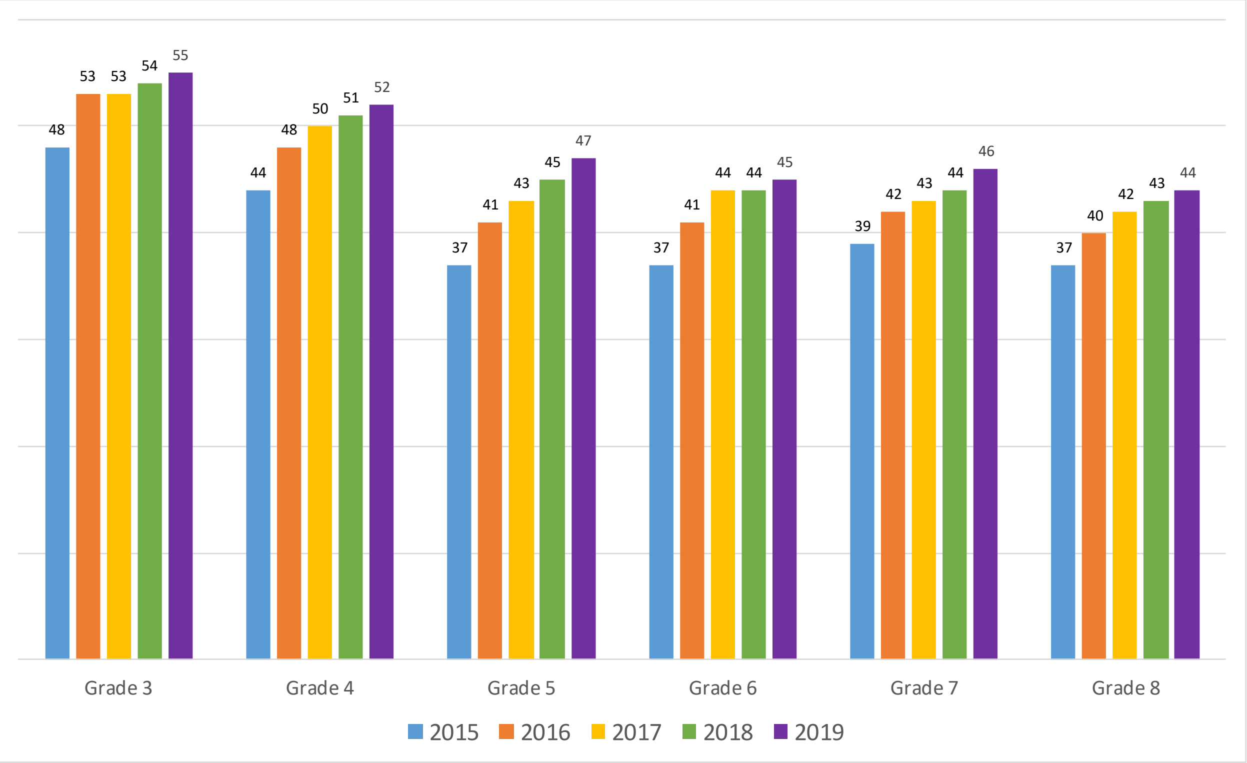


Figure 3. ELA/L Average Scale Score Across Years

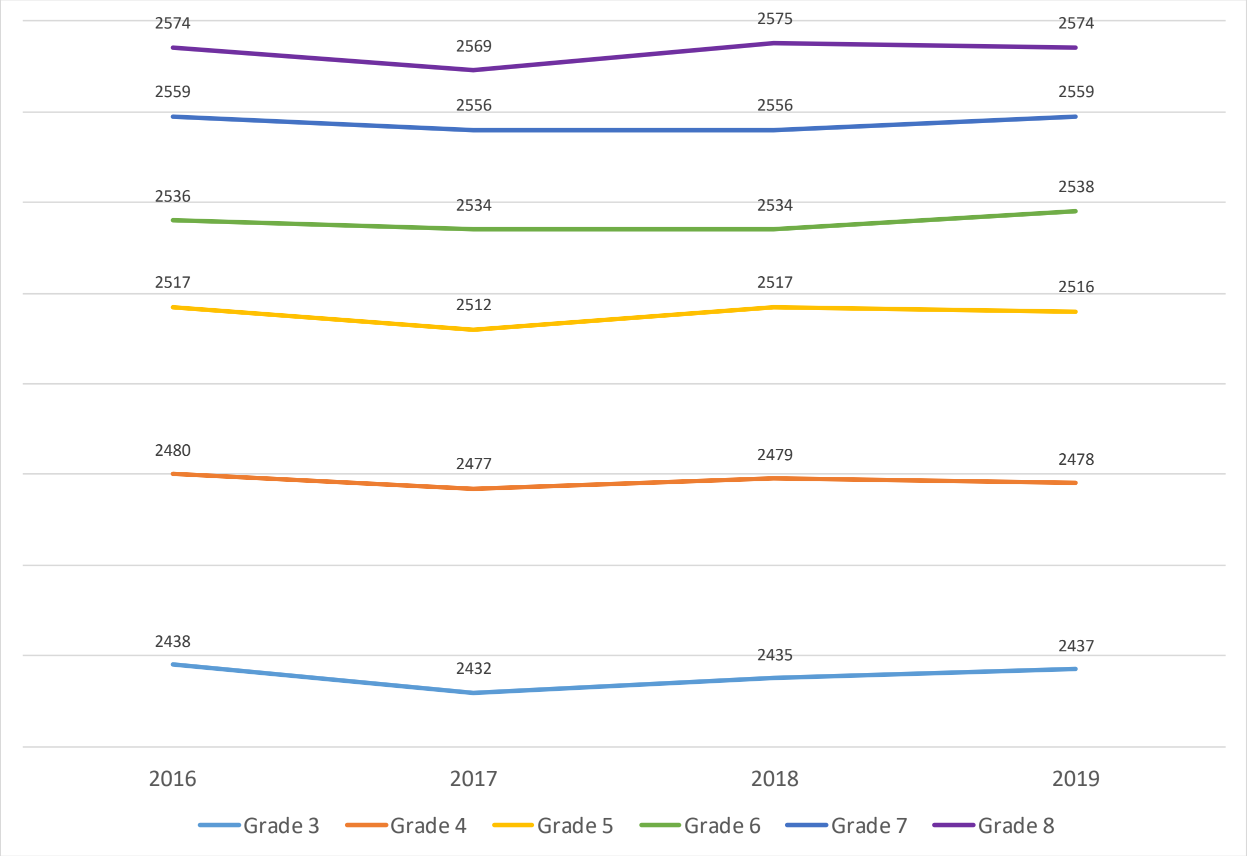
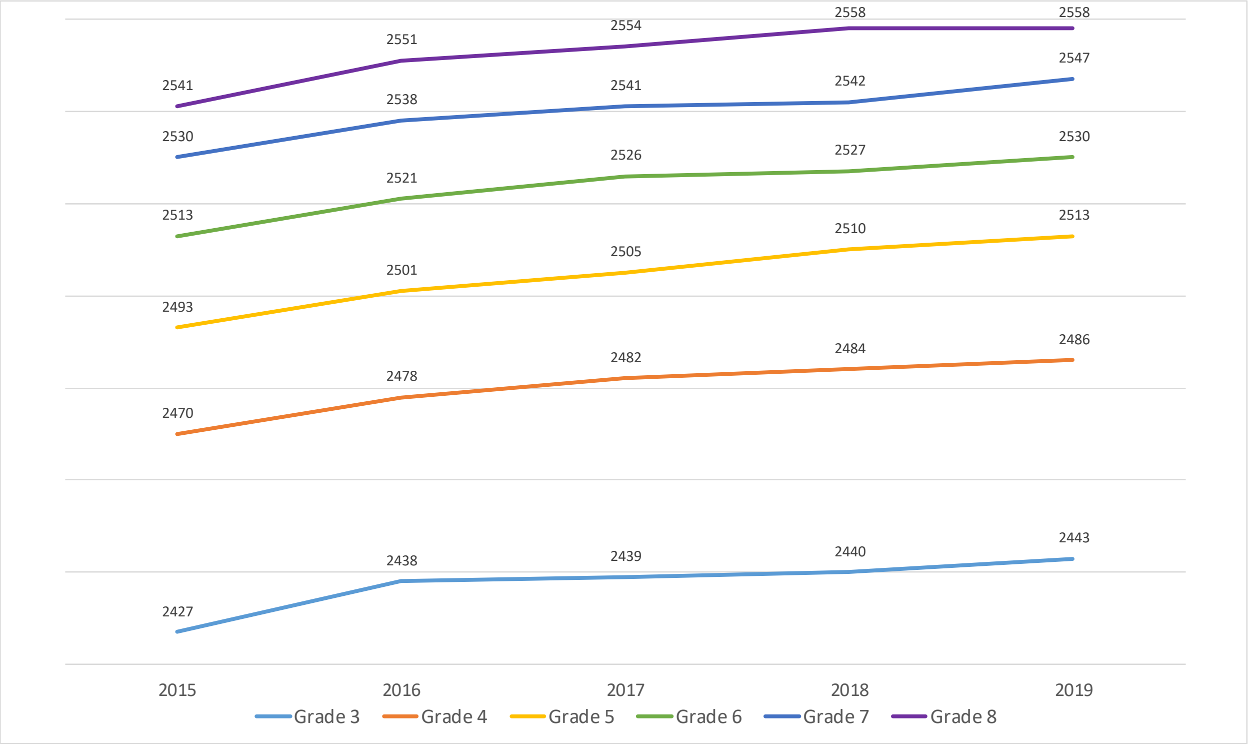


Figure 4. Mathematics Average Scale Score Across Years



Because the precision of scores in each claim is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three performance categories, taking into account the standard error of measurement (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 and 4: Writing and Research**	**Claim 3 Listening**
3	Below	24	29	15
	At/Near	45	43	61
	Above	31	29	24
4	Below	22	29	14
	At/Near	47	44	60
	Above	30	27	26
5	Below	22	26	16
	At/Near	43	41	61
	Above	35	34	24
6	Below	26	25	12
	At/Near	45	46	66
	Above	29	29	23
7	Below	25	24	15
	At/Near	44	47	66
	Above	31	29	19
8	Below	25	27	14
	At/Near	43	44	62
	Above	33	30	24

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

Grade	Performance Category	Claim 1	Claims 2 and 4	Claim 3
3	Below	29	24	21
	At/Near	32	44	45
	Above	39	32	33
4	Below	31	27	25
	At/Near	32	44	43
	Above	37	28	31
5	Below	36	30	29
	At/Near	31	44	46
	Above	33	26	25
6	Below	36	32	30
	At/Near	34	44	45
	Above	31	24	24
7	Below	38	30	23
	At/Near	30	44	53
	Above	32	26	25
8	Below	38	31	28
	At/Near	32	42	48
	Above	30	27	23

Legend:
Claim 1: Concepts and Procedures;
Claims 2 and 4: Problem Solving and 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 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	1:47	0:53	2:07	2:15	2:26	2:42	3:11
4	1:52	1:00	2:11	2:21	2:33	2:50	3:22
5	1:51	0:49	2:11	2:20	2:32	2:48	3:19
6	1:47	0:47	2:07	2:16	2:27	2:43	3:11
7	1:40	0:46	1:58	2:06	2:17	2:33	3:04
8	1:34	0:42	1:51	1:58	2:08	2:22	2:50

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	2:41	2:53	3:08	3:29	4:05
4	2:16	1:06	2:47	2:59	3:16	3:40	4:20
5	2:30	1:11	3:03	3:17	3:34	3:58	4:40
6	2:25	1:04	2:54	3:06	3:22	3:44	4:23
7	1:58	0:55	2:22	2:33	2:46	3:06	3:40
8	2:03	0:57	2:29	2:39	2:52	3:11	3:46
CAT Component							
3	1:28	0:45	1:49	1:58	2:08	2:23	2:49
4	1:36	0:49	1:58	2:07	2:19	2:37	3:08
5	1:35	0:45	1:56	2:05	2:15	2:31	2:57
6	1:37	0:43	1:57	2:05	2:15	2:30	2:56
7	1:27	0:40	1:46	1:53	2:03	2:17	2:43
8	1:30	0:42	1:49	1:57	2:06	2:20	2:47
PT Component							
3	0:42	0:25	0:54	0:59	1:05	1:13	1:29
4	0:41	0:24	0:52	0:56	1:02	1:10	1:26
5	0:55	0:35	1:09	1:16	1:24	1:36	1:58
6	0:48	0:30	1:00	1:05	1:12	1:22	1:40
7	0:31	0:21	0:39	0:43	0:49	0:56	1:10
8	0:33	0:21	0:42	0:46	0:51	0:58	1:11

3.4 Distribution of Student Ability and Item Difficulty

Figures 5–10 display the empirical distribution of the Connecticut student scale scores in the 2018–2019 administration and the distribution of the administered summative item difficulty parameters for overall and by reporting category. For overall, the student ability distribution is shifted to the left in all grades and subjects, a pattern more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to accurately measure high-performing students but needs additional easy items to better measure low-performing students. At the reporting category level, the student ability distribution is shifted to the left in claim 3 (Listening) in ELA/L. In mathematics, the student ability distribution is shifted to the left for all claims except for claim 1 in lower grades. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool and to augment the pool in proportion to the test blueprint constraints (e.g., content, Depth-of-Knowledge [DOK], item type, and item difficulties) to better measure low performing students.

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

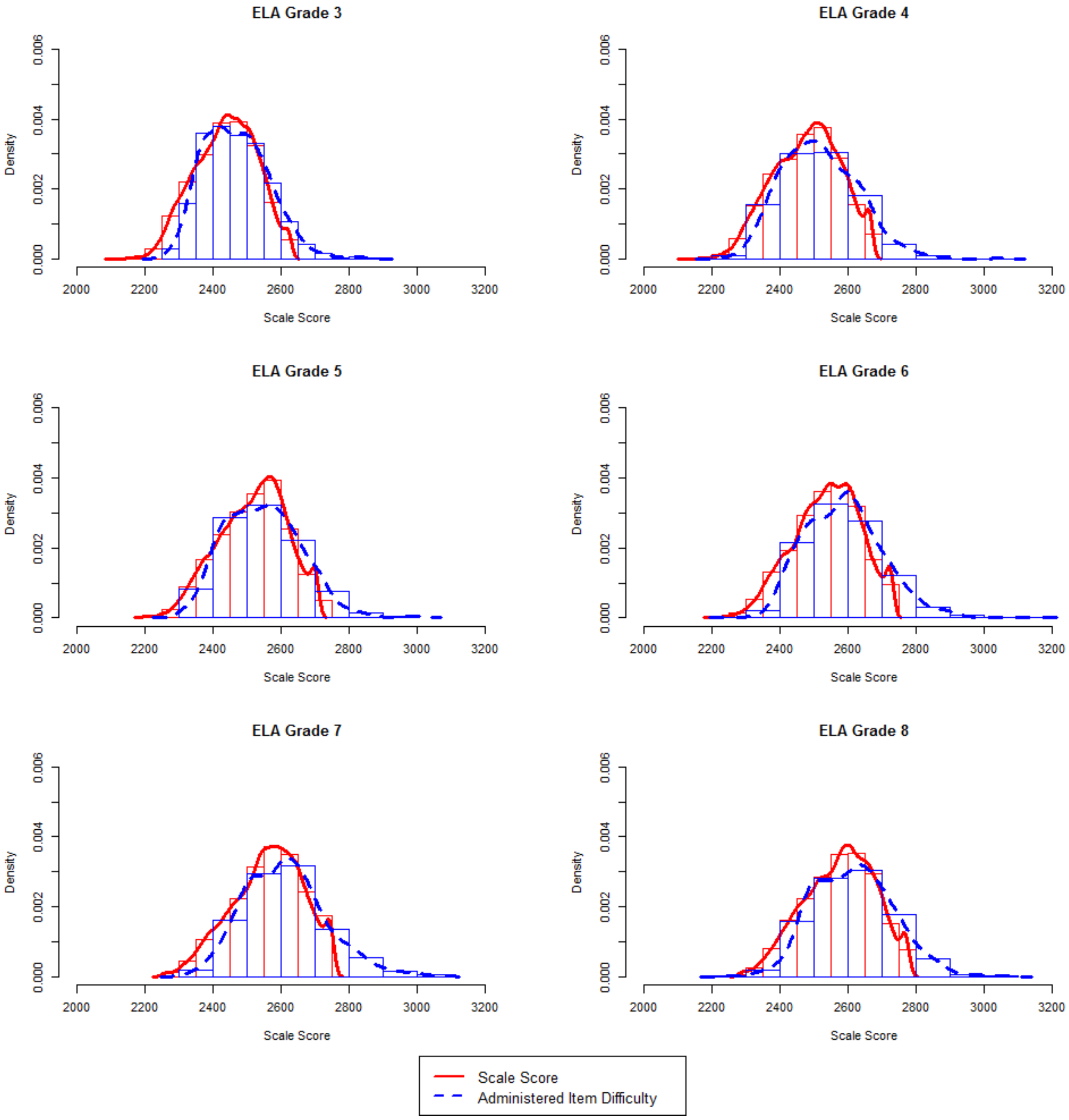
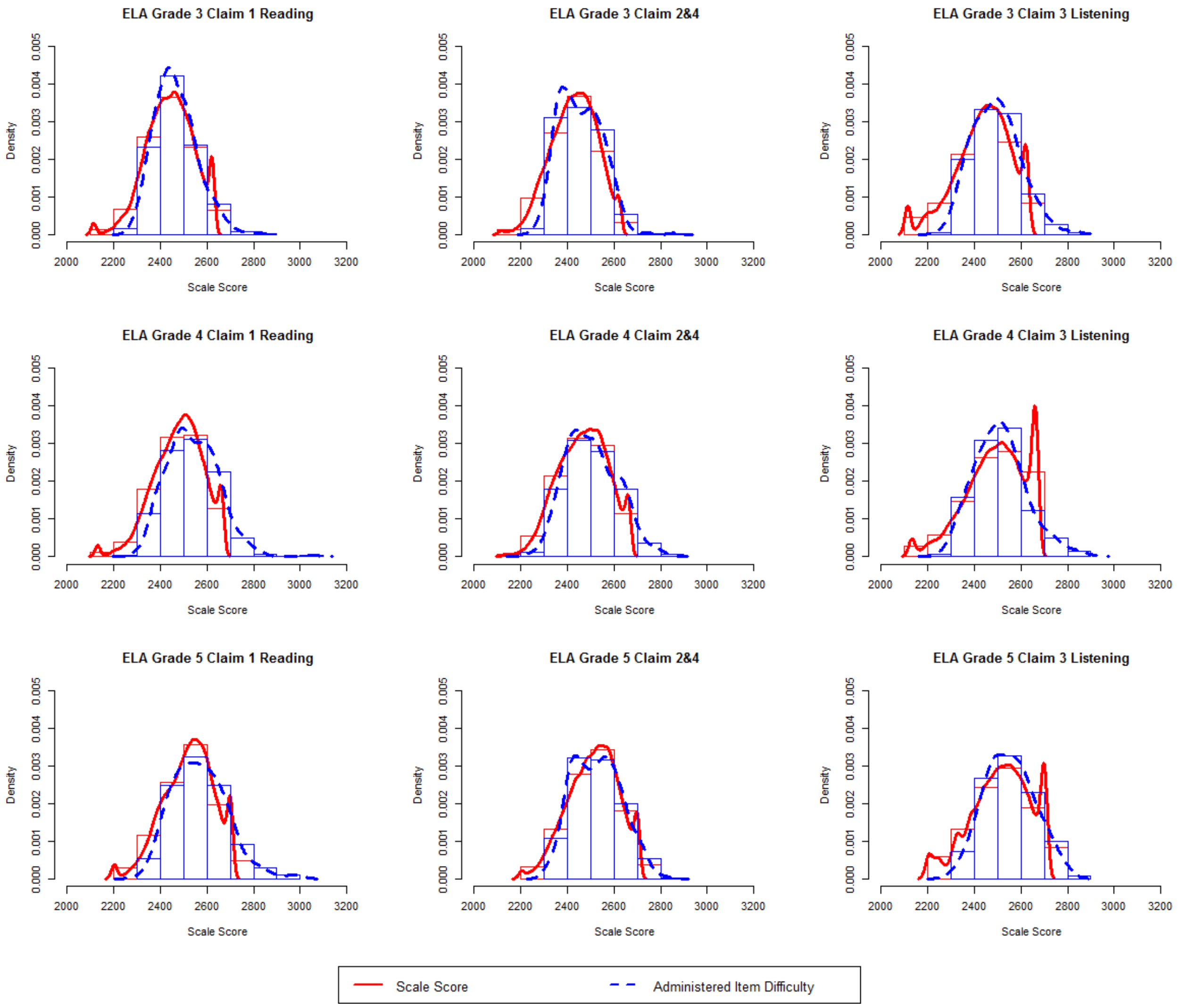


Figure 6. Student Ability—Item Difficulty Distribution by Claim: ELA/L (Grades 3–5)



ELA Grade 3 Claim 1 Reading

ELA Grade 3 Claim 2&4

ELA Grade 3 Claim 3 Listening

ELA Grade 4 Claim 1 Reading

ELA Grade 4 Claim 2&4

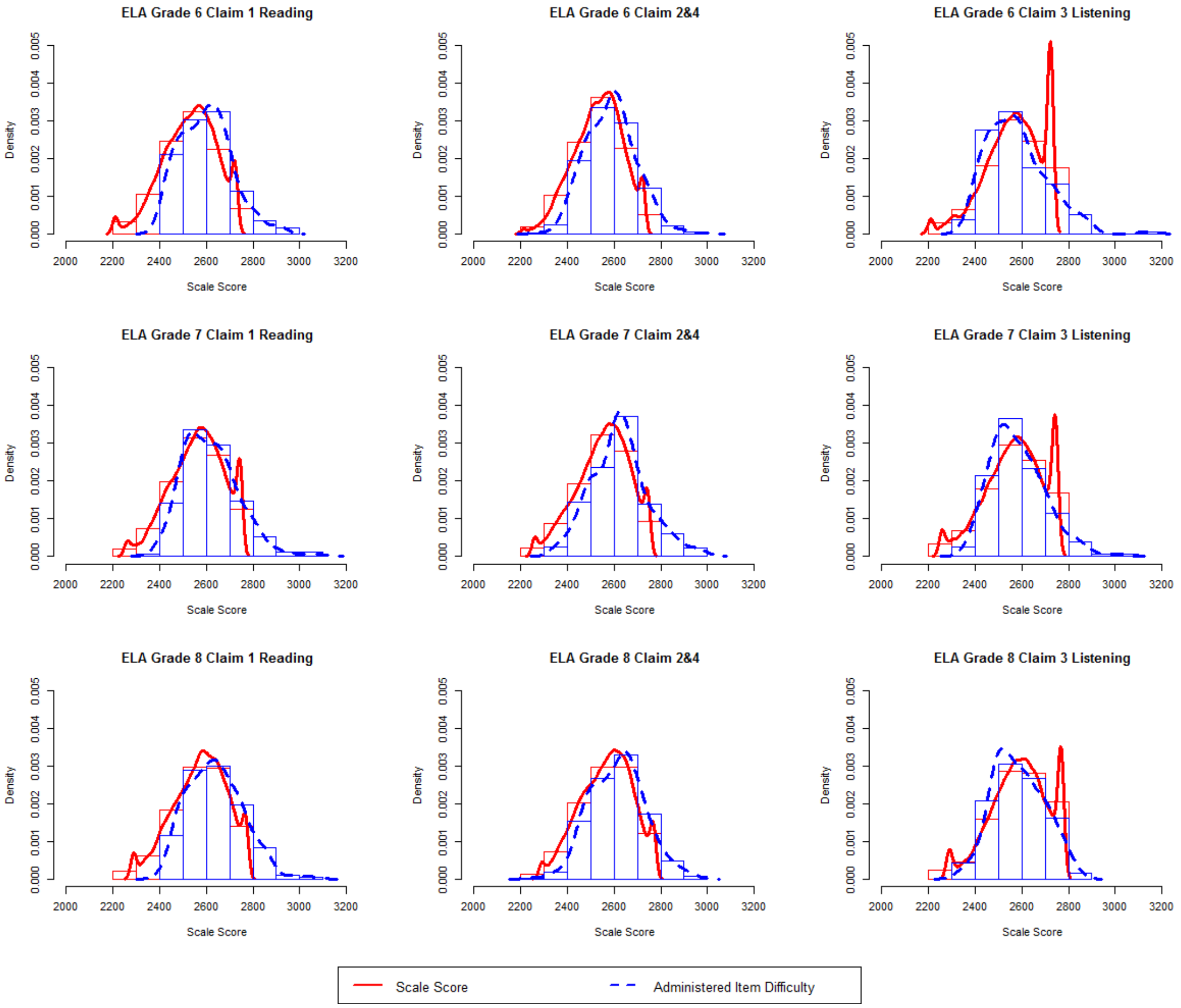
ELA Grade 4 Claim 3 Listening

ELA Grade 5 Claim 1 Reading

ELA Grade 5 Claim 2&4

ELA Grade 5 Claim 3 Listening

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



ELA Grade 6 Claim 1 Reading

ELA Grade 6 Claim 2&4

ELA Grade 6 Claim 3 Listening

ELA Grade 7 Claim 1 Reading

ELA Grade 7 Claim 2&4

ELA Grade 7 Claim 3 Listening

ELA Grade 8 Claim 1 Reading

ELA Grade 8 Claim 2&4

ELA Grade 8 Claim 3 Listening

Figure 8. Student Ability—Item Difficulty Distribution for Mathematics

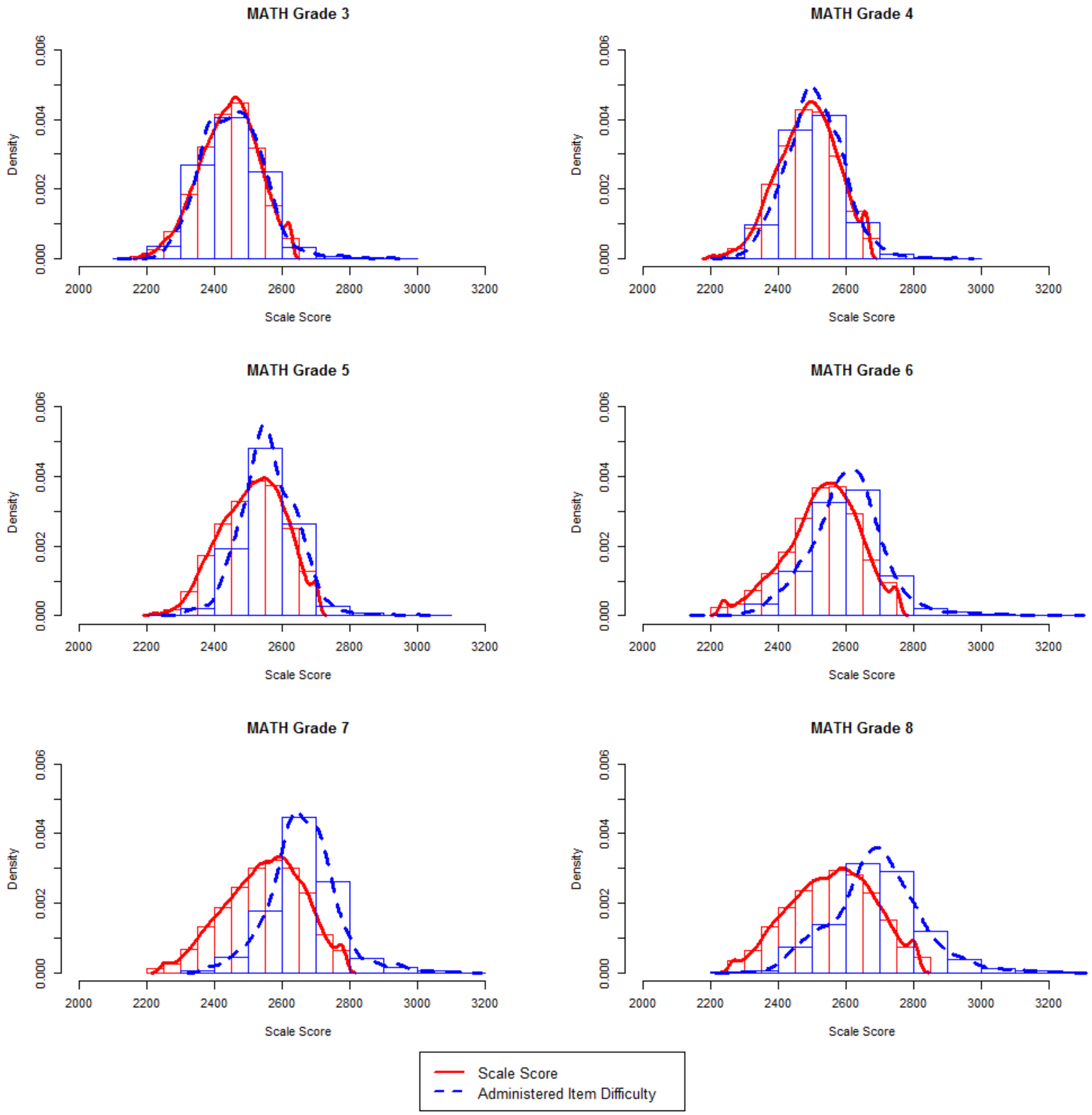


Figure 9. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 3–5)

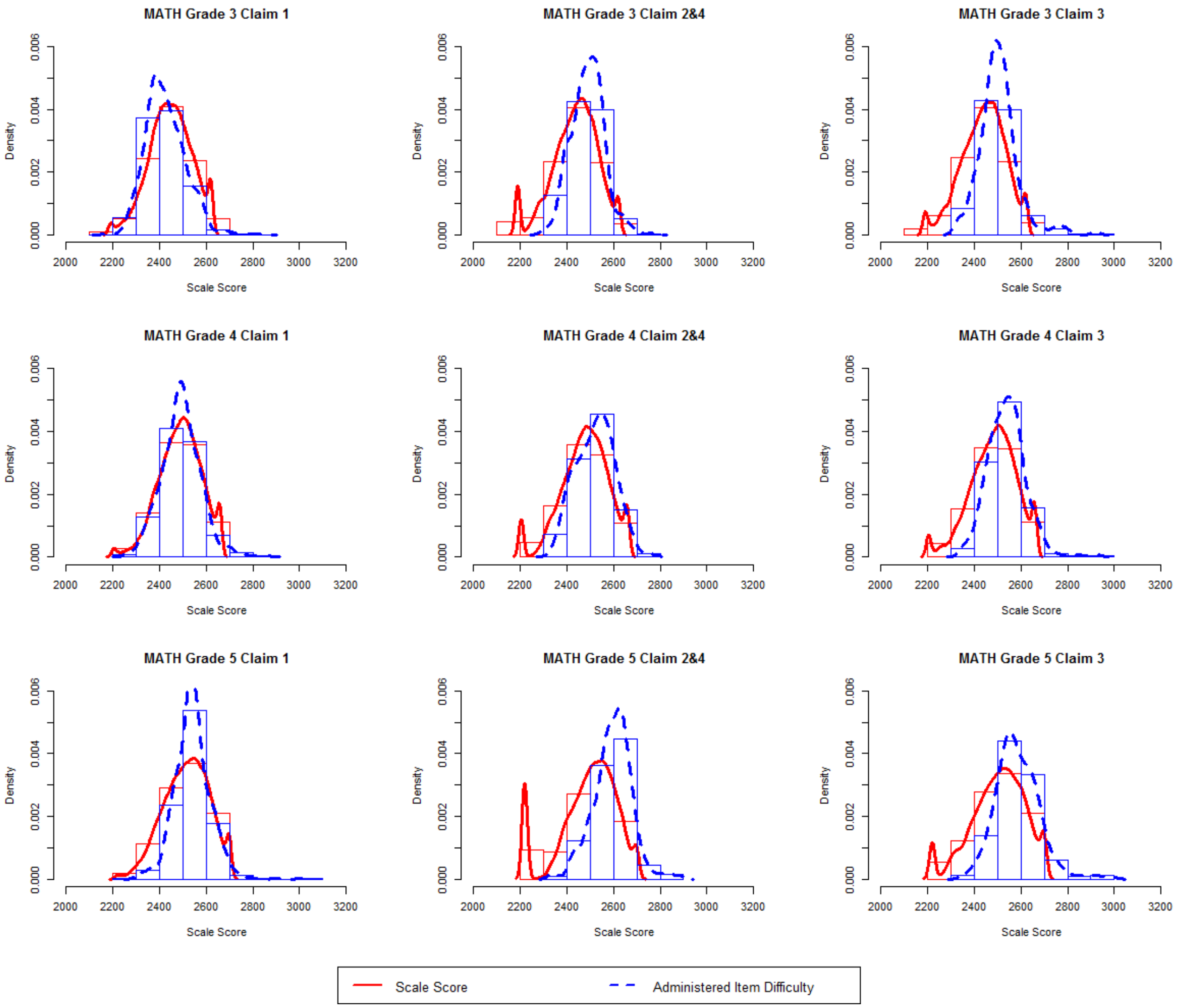
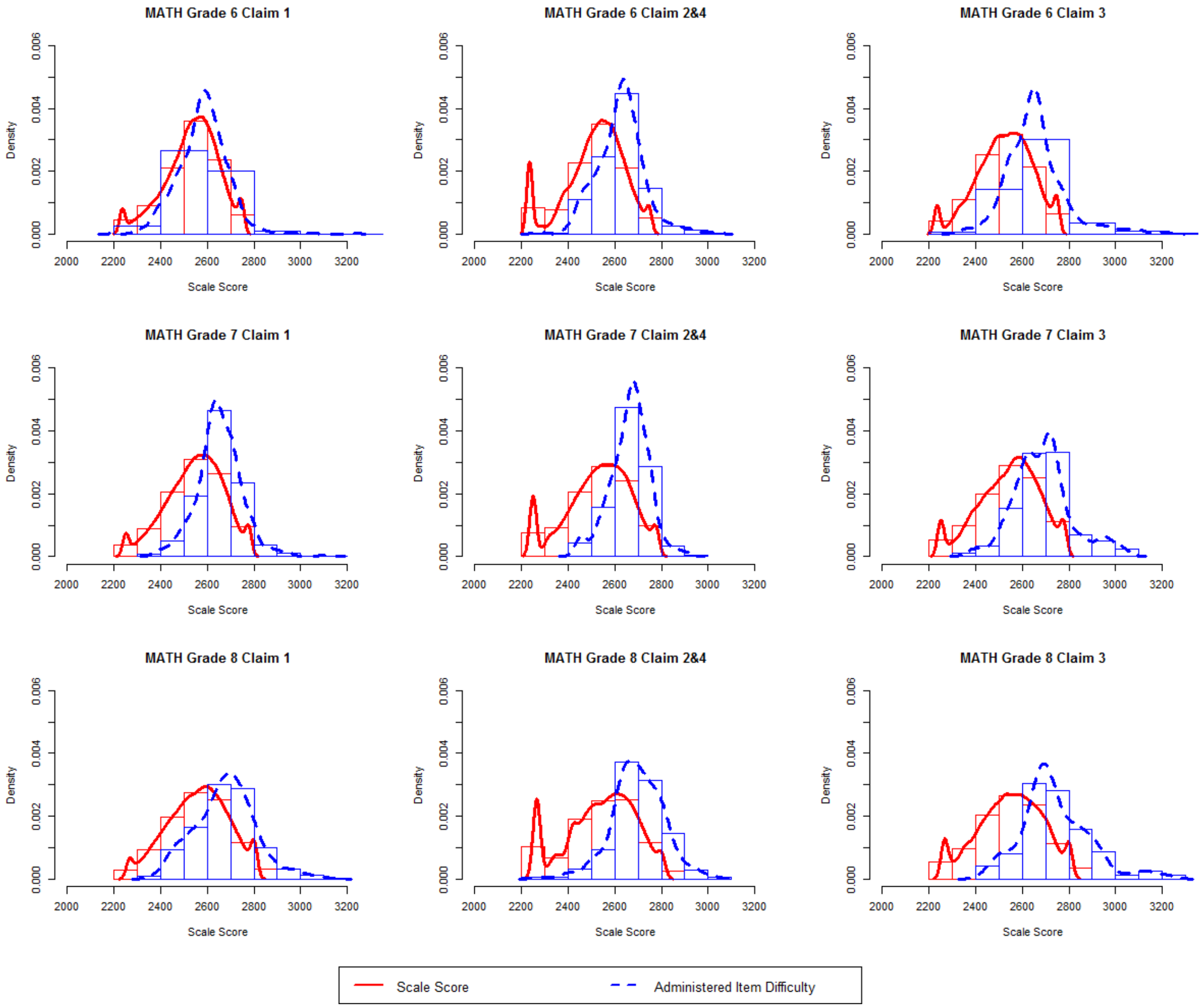


Figure 10. Student Ability—Item Difficulty Distribution by Claim: Mathematics (Grades 6–8)



MATH Grade 6 Claim 1

MATH Grade 6 Claim 2&4

MATH Grade 6 Claim 3

MATH Grade 7 Claim 1

MATH Grade 7 Claim 2&4

MATH Grade 7 Claim 3

MATH Grade 8 Claim 1

MATH Grade 8 Claim 2&4

MATH Grade 8 Claim 3

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 specify a range of items to be administered in each claim, content domain/standards, and/or targets. Moreover, blueprints constrain the Depth of Knowledge (DOK) and item and passage types. For DOK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In 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. 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 grade 6. 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, each administered fewer or more items 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	9	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	9	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 (G3–5)**	**Required Items (G6–8)**	**%Blueprint Match**					
			G3	**G4**	**G5**	**G6**	**G7**	**G8**
Claim 1 DOK1		≤ 5				100	100	100
Claim 1 DOK2	≥ 7		100	100	100			
Claim 1 DOK3 or higher	≥ 2	≥ 2	100	100	100	100	100	100
Claim 1 Short Answer in Target 2 or 4	0–1	0–1	100	100	100	100	100	100
Claim 1 Short Answer in Target 9 or 11	0–1	0–1	100	100	100	100	100	100
Claim 2 DOK2	≥ 4	≥ 4	100	100	100	100	100	100
Claim 2 DOK3 or higher	≥ 1	≥ 1	100	100	100	100	100	100
Claim 2 Brief Write	1	1	100	100	100	100	100	100
Claim 3 DOK2 or higher	≥ 3	≥ 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 and 4	Overall	6	100	6	100	6	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3	Overall	8	100	8	100	8	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	3	100	3	100	3	100
	Targets B, E	3	100	3	100	3	100
	Targets C, F	2	100	2	100	2	100

Table 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 B, G	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	100
	Targets B, E, G					5–6	100
	Targets F, H					2–3	100
	Supporting Cluster					4–5	100
	Targets A, I, J					4–5	100
2 and 4	Overall	6	100	6	100	6	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	2. Target A	2	100	2	100	2	100
	2. Targets B, C, D	1	100	1	100	1	100
	4. Targets A, D	1	100	1	100	1	100
	4. Targets B, E	1	100	1	100	1	100
	4. Targets C, F	1	100	1	100	1	100
3-Calc	Overall	7	100	8	100	8	100
	DOK 3 or higher	≥ 2	100	≥ 2	100	≥ 2	100
	Targets A, D	2–3	100	3	100	3	100
	Targets B, E	2–3	100	3	100	3	100
	Targets C, F, G	1–2	100	2	100	2	100
3-No Calc	Overall	1	100				

Table 25 summarizes the target coverage by claim that includes the number of unique targets administered in each delivered test. 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	7–14	3–5	1–1	3–3
6	14	5	1	3	10	5	1	3	9–11	5–5	1–1	3–3
7	14	5	1	3	11	5	1	3	8–11	3–5	1–1	3–3
8	14	5	1	3	11	5	1	3	8–11	3–5	1–1	3–3
Mathematics												
3	11	4	6	6	11	2	6	3	9–11	2–2	4–6	2–3
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	8–9	2–2	3–6	3–4
6	10	4	7	6	10	2	5	3	8–10	2–2	3–7	3–3
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 $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$, where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y.

When corrected for attenuation (above diagonal), the correlations among claim scores are higher than observed correlations. The disattenuated correlations are quite high. The correction for attenuation is large because the marginal reliabilities of claim 3 scores in ELA/L and the marginal reliabilities of claims 2 and 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 is low, the performance of all the claim scores is reported in three performance categories. The distribution of performance categories for each claim is provided in Tables 17 and 18, Section 3.2. Scale scores are not reported for claims.

Table 26. Correlations among Claims for ELA/L

Grade	**Claim**	**Observed & Disattenuated Correlation**		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1: Reading		0.97	0.98
	Claims 2 & 4: Writing & Research	0.76		0.97
	Claim 3: Listening	0.66	0.68	
4	Claim 1: Reading		0.98	0.99
	Claims 2 & 4: Writing & Research	0.76		0.98
	Claim 3: Listening	0.65	0.67	
5	Claim 1: Reading		0.99	1
	Claims 2 & 4: Writing & Research	0.79		0.99
	Claim 3: Listening	0.68	0.69	
6	Claim 1: Reading		0.98	1
	Claims 2 & 4: Writing & Research	0.77		1
	Claim 3: Listening	0.64	0.67	
7	Claim 1: Reading		0.99	1
	Claims 2 & 4: Writing & Research	0.78		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	0.69	0.70	

Table 27. Correlations among Claims for Mathematics

Grade	**Claim**	**Observed & Disattenuated Correlation**		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		0.99	0.96
	Claims 2 & 4	0.80		1
	Claim 3	0.80	0.75	
4	Claim 1		0.99	0.99
	Claims 2 & 4	0.82		1
	Claim 3	0.82	0.77	
5	Claim 1		1	0.99
	Claims 2 & 4	0.78		1
	Claim 3	0.79	0.74	
6	Claim 1		1	1
	Claims 2 & 4	0.83		1
	Claim 3	0.81	0.77	
7	Claim 1		1	1
	Claims 2 & 4	0.82		1
	Claim 3	0.79	0.73	
8	Claim 1		1	1
	Claims 2 & 4	0.79		1
	Claim 3	0.80	0.72	

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 (SEMs). 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 item response theory (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 (CAT), because selected items vary across students, the measurement error can vary for the same ability depending on the selected items for each student.

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

5.1 MARGINAL RELIABILITY

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

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

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

where N is the number of students; $CSEM_i$is the conditional SEM of the scale score for student i, and $\sigma^2$is the variance of the scale score. The higher 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

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

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	36,516	38	41	0.91	2437	91	27
4	37,727	38	41	0.91	2478	99	30
5	38,605	38	41	0.91	2516	100	29
6	39,588	38	41	0.90	2538	99	31
7	39,165	38	41	0.90	2559	105	32
8	39,372	40	41	0.91	2574	104	31
Mathematics							
3	36,460	39	40	0.95	2443	85	19
4	37,675	37	40	0.95	2486	87	20
5	38,514	38	40	0.94	2512	94	23
6	39,488	38	39	0.94	2530	109	26
7	39,002	38	40	0.94	2547	115	29
8	39,216	38	40	0.94	2558	123	31

5.2 STANDARD ERROR CURVES

Figures 11 and 12 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 11. Conditional Standard Error of Measurement for ELA/L

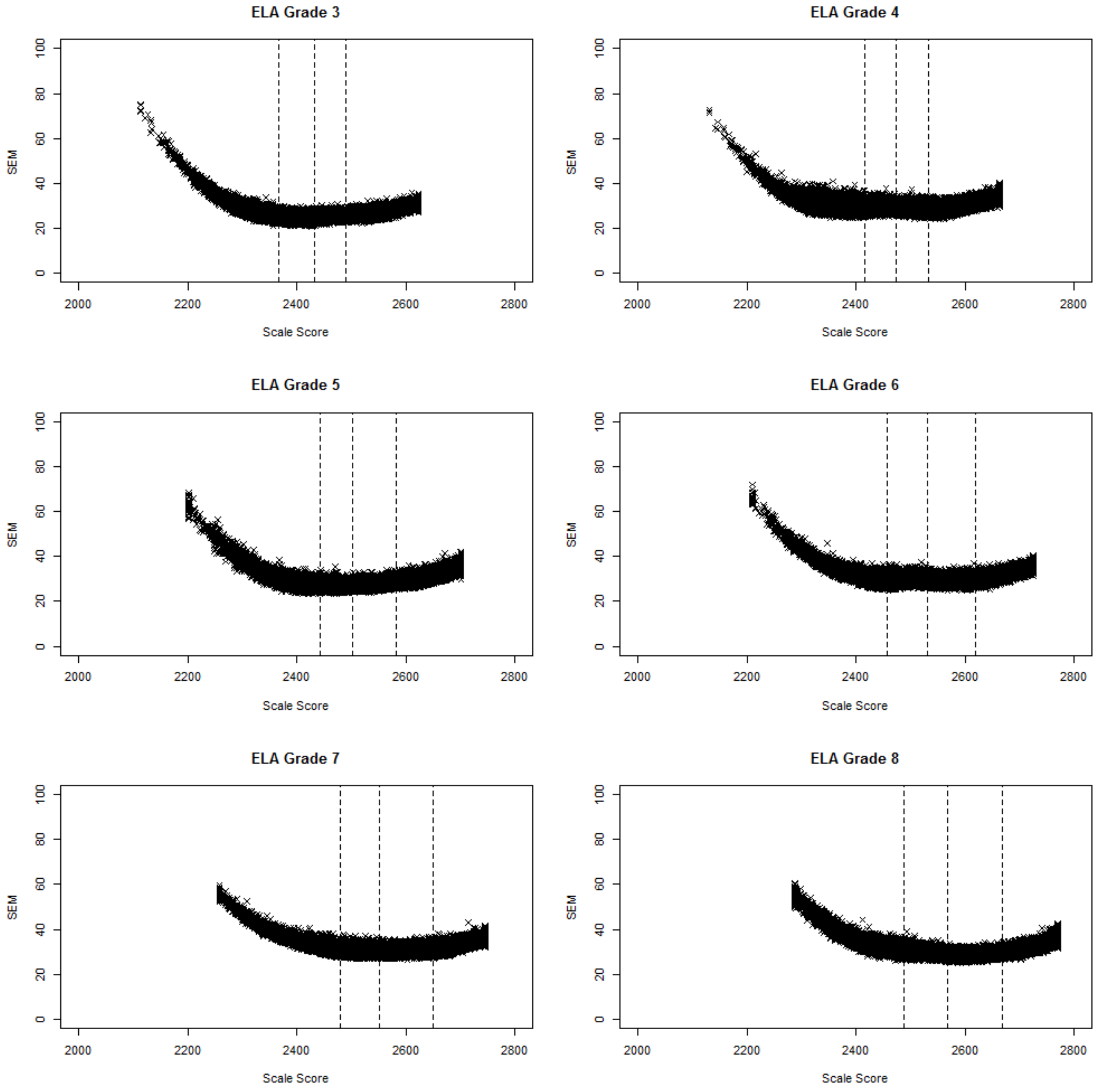
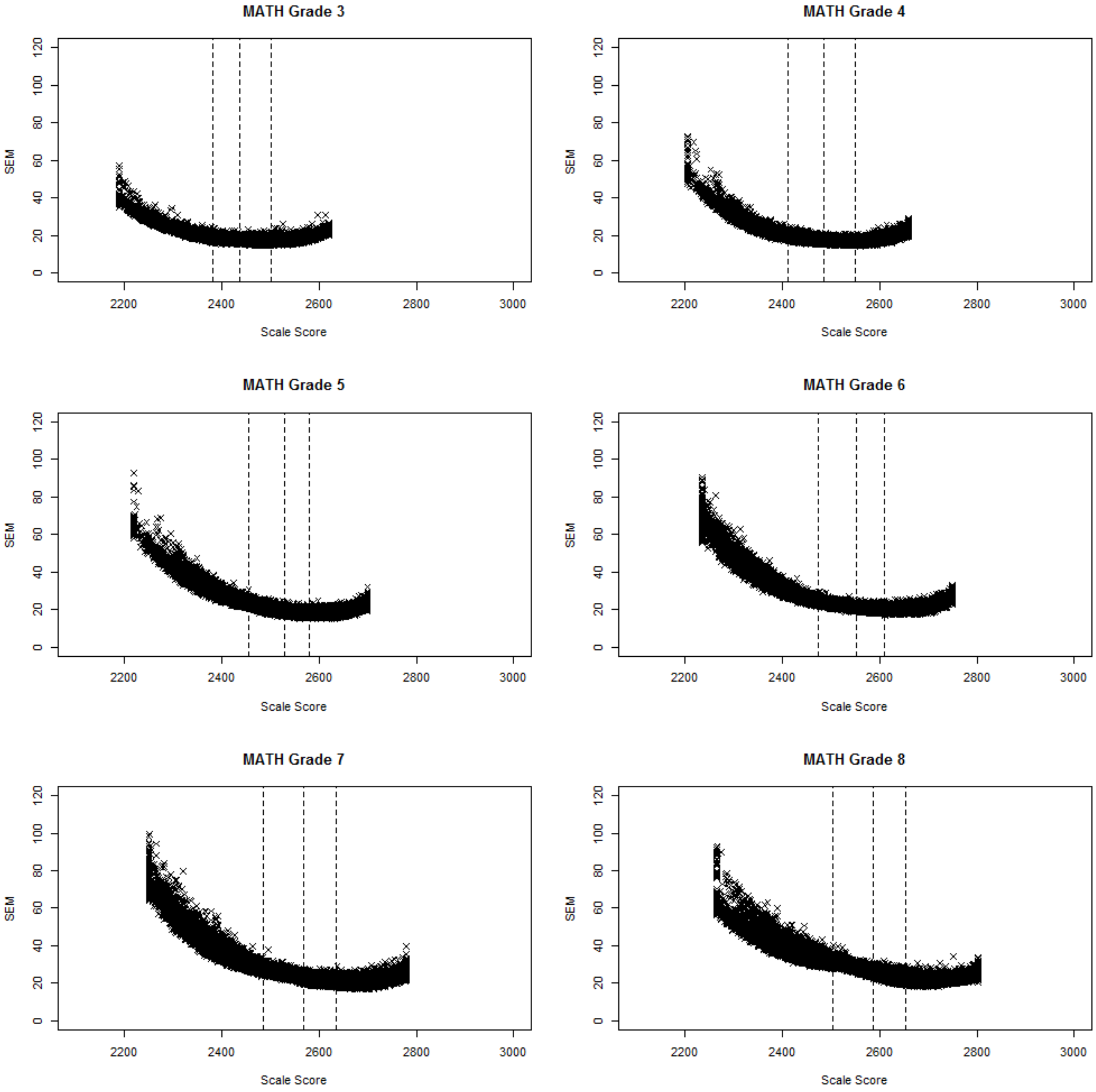


Figure 12. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the Figures 11 and 12 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 each cut score and the difference in average conditional SEMs between two cut scores. As shown in Figures 11 and 12, 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	31	25	25	27	27
4	32	29	29	30	30
5	31	27	28	31	29
6	34	30	29	32	31
7	36	30	31	33	32
8	36	30	29	32	31
Mathematics					
3	23	18	17	18	19
4	25	18	17	19	20
5	30	21	18	19	23
6	36	22	20	21	26
7	39	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 Standard Errors of Measurement between Two Cuts

Grade	L2 Cut	L3 Cut	L4 Cut	|L2-L3|	|L3-L4|	|L2-L4|
ELA/L						
3	26	25	26	1	1	0
4	29	29	29	0	1	0
5	27	27	29	0	1	2
6	29	29	29	0	0	0
7	31	30	31	0	1	1
8	31	29	30	2	1	1
Mathematics						
3	19	18	17	1	1	2
4	19	17	17	2	0	2
5	23	19	18	5	1	6
6	24	21	20	3	1	4
7	28	23	21	4	2	7
8	31	26	22	5	4	9

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

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

For the *i*th student, the student's estimated ability is $\hat{\theta}_i$ with 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

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

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 being at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and one minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, 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

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

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

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

where the likelihood function based on general IRT models is

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = \left(a_j, b_j, c_j\right)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \dots, b_{jK_i})$ 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 78% 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 narrower intervals.

Accuracy of classifications is higher than the consistency of classifications in all achievement levels. The accuracy is higher than the consistency because the accuracy is based on one test with a measurement error and the true score while the consistency is based on two tests with measurement errors. The classification indexes by subgroup 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	84	77
	L1	89	83	90	85
	L2	69	58	73	63
	L3	65	54	79	72
	L4	88	82	90	86
4	Overall	77	70	84	78
	L1	89	83	90	84
	L2	60	47	80	72
	L3	62	51	79	71
	L4	88	82	90	86
5	Overall	79	71	83	77
	L1	90	84	91	86
	L2	64	52	77	68
	L3	72	63	71	61
	L4	86	80	91	86
6	Overall	78	69	83	77
	L1	89	81	92	87
	L2	68	57	78	69
	L3	73	64	72	62
	L4	85	77	90	85
7	Overall	78	70	84	77
	L1	89	83	91	87
	L2	67	55	76	67
	L3	75	67	75	65
	L4	84	76	91	86
8	Overall	79	71	83	76
	L1	88	82	91	87
	L2	70	59	71	61
	L3	77	70	71	61
	L4	84	76	91	87

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 students in these subgroups 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	0.91	0.91	0.90	0.90	0.91
Female	0.91	0.90	0.91	0.90	0.90	0.91
Male	0.91	0.91	0.92	0.91	0.91	0.91
African American	0.90	0.89	0.90	0.88	0.89	0.89
AmerIndian/Alaskan	0.90	0.88	0.91	0.87	0.89	0.91
Asian	0.91	0.89	0.89	0.87	0.89	0.88
Hispanic/Latino	0.90	0.89	0.90	0.89	0.89	0.90
Pacific Islander	0.86	0.86	0.92	0.89	0.91	0.93
White	0.89	0.89	0.89	0.88	0.88	0.89
Two or More Races	0.92	0.90	0.92	0.90	0.90	0.91
LEP	0.87	0.86	0.85	0.80	0.79	0.73
Special Education	0.86	0.87	0.88	0.85	0.86	0.85

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	0.94	0.94
Female	0.95	0.94	0.94	0.94	0.94	0.93
Male	0.95	0.95	0.94	0.94	0.94	0.94
African American	0.93	0.93	0.91	0.91	0.90	0.89
AmerIndian/Alaskan	0.94	0.93	0.92	0.92	0.92	0.93
Asian	0.95	0.94	0.94	0.94	0.95	0.95
Hispanic/Latino	0.94	0.93	0.92	0.92	0.90	0.90
Pacific Islander	0.93	0.93	0.94	0.92	0.94	0.95
White	0.94	0.94	0.94	0.94	0.94	0.94
Two or More Races	0.95	0.95	0.95	0.94	0.94	0.94
LEP	0.93	0.91	0.87	0.84	0.78	0.73
Special Education	0.93	0.92	0.88	0.88	0.85	0.85

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	2441	102	50
	Claims 2 & 4: Writing & Research	16	16	0.83	2429	99	41
	Claim 3: Listening	8	9	0.60	2438	120	76
4	Claim 1: Reading	14	16	0.75	2479	107	54
	Claims 2 & 4: Writing & Research	16	16	0.81	2468	107	47
	Claim 3: Listening	8	9	0.57	2489	128	83
5	Claim 1: Reading	14	16	0.77	2520	108	52
	Claims 2 & 4: Writing & Research	16	16	0.83	2511	108	44
	Claim 3: Listening	8	9	0.59	2513	126	81
6	Claim 1: Reading	14	16	0.76	2529	115	56
	Claims 2 & 4: Writing & Research	16	16	0.80	2532	105	47
	Claim 3: Listening	8	9	0.46	2566	120	88
7	Claim 1: Reading	14	16	0.78	2561	114	53
	Claims 2 & 4: Writing & Research	16	16	0.79	2552	114	52
	Claim 3: Listening	8	9	0.51	2567	123	87
8	Claim 1: Reading	16	16	0.78	2572	116	54
	Claims 2 & 4: Writing & Research	16	16	0.81	2566	112	49
	Claim 3: Listening	8	9	0.56	2589	121	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	2446	92	28
	Claims 2 & 4	8	11	0.71	2435	98	53
	Claim 3	9	11	0.76	2439	95	46
4	Claim 1	20	20	0.91	2489	91	27
	Claims 2 & 4	8	10	0.75	2477	101	51
	Claim 3	9	11	0.76	2481	99	49
5	Claim 1	20	20	0.90	2516	98	32
	Claims 2 & 4	8	10	0.64	2495	122	73
	Claim 3	9	10	0.71	2503	114	61
6	Claim 1	19	19	0.89	2533	116	38
	Claims 2 & 4	9	10	0.72	2517	125	67
	Claim 3	9	11	0.74	2525	119	61
7	Claim 1	20	20	0.89	2547	122	40
	Claims 2 & 4	9	10	0.68	2533	135	76
	Claim 3	8	10	0.67	2543	130	75
8	Claim 1	20	20	0.89	2560	130	43
	Claims 2 & 4	8	10	0.65	2542	146	86
	Claim 3	9	10	0.70	2549	137	75

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

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

where the vector $\boldsymbol{b}_{\boldsymbol{i}}^{'} = (b_{i,1}, \dots, b_{i,m_i})$ 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 $p_{ij}(z_{ij}|\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i})$ 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 $m_i = 1$,

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

in the case of items with two or more points,

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

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

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

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

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

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

6.2 Rules for Transforming Theta to Vertical Scale Scores

The student's performance in each subject is summarized in an overall test score referred to as a *scale score*. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced Assessment Consortium. Table 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
Mathematics	3–8	79.3	2514.9

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

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

where SE_{ss} is the standard error of the ability estimate on the reporting scale, 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 errors 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
Mathematics	3	-4.1132	1.3335	2189	2621
Mathematics	4	-3.9204	1.8191	2204	2659
Mathematics	5	-3.7276	2.3290	2219	2700
Mathematics	6	-3.5348	2.9455	2235	2748
Mathematics	7	-3.3420	3.3238	2250	2778
Mathematics	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. The difference between the proficiency cut score and the claim score plus or minus 1.5 times standard error of the claim is used to determine the relative strengths and weaknesses.

For summative tests, the specific rules are as follows:

- Below Standard (Code = 1): if $round(SS_{rc} + 1.5 * SE(SS_{rc}),0) < SS_p$
- 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 of the four claims in ELA/L and claim 1 for mathematics.

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

6.6.1 Target Scores Relative to Student's Overall Estimated Ability

By defining $p_{ij} = p(z_{ij} = 1)$, representing the probability that student *j* responds correctly to item I, z_{ij} represents the *j*th student's score on the *i*th item. For items with one score point, 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:

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

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:

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

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:

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

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

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

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,

$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}$, 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 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 $\bar{\delta}_{Tg} - se\left(\bar{\delta}_{Tg}\right) \geq 0.07$ then performance is *above* the Proficiency Standard.
- If $\bar{\delta}_{Tg} + se\left(\bar{\delta}_{Tg}\right) \leq -0.07$, then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.
- If $se\left(\bar{\delta}_{Tg}\right)$ > 0.2, data are insufficient.

6.7 HANDSCORING

AIR provides the automated electronic scoring and Measurement Incorporated (MI) provides all handscoring for the Smarter Balanced summative assessments. Short-answer (SA) items and full-write items in ELA/L and SA items in mathematics are scored by human raters; this is also referred to as *handscoring*. The procedures for scoring these items are specified by Smarter Balanced.

Outlined in the following paragraphs is the scoring process MI follows. This procedure is used to score responses to all constructed-response short answer and essay items.

6.7.1 Rater Selection

MI maintains a large pool of raters at each scoring center, as well as distributive raters who work remotely. MI's recruiting team first recruits qualified raters who have experience scoring the Smarter Balanced assessment. Rater accuracy parameters are used to focus recruitment efforts for experienced Smarter Balanced raters in order to recruit the most objectively accurate raters. Once recruited, experienced raters are assigned to the content area and grade bands in which they are most experienced. These experienced, demonstrably accurate raters make up the majority of the total rater pool. To supplement this core pool, MI contacts other raters in their database who have experience successfully scoring other large-scale assessments. These raters are assigned to the grade level, subject area, and item type for which they are most qualified based on their performance on similar projects. Returning staff are selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. Finally, MI targets recruitment of new raters for site-based and remote scoring as needed, in order to continue to identify talent across the country that will best fulfill the handscoring requirements. For new raters, MI's recruiting team reviews applications, including prospective raters' resumes, references, proof of degree, and recognition of rater requirements, before offering employment.

In selecting team leaders, MI scoring leadership review the files of all returning staff. They look for people who are experienced team leaders with a record of good performance on previous projects and also consider raters 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, raters, 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 Rater Training

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

Once hired, raters 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). Raters are trained on a specific item type (i.e., brief writes, reading, research, full-writes, or mathematics). Within each group, raters are divided into teams consisting of one team leader and 10–15 raters. Each team leader and rater are assigned a unique number for easy identification of their scoring work throughout the scoring session. The number of items an individual rater scores is minimized so that the rater 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. Raters 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 raters.

After the contracts and nondisclosure forms are signed and the scoring director completes his or her introductory remarks, training begins. Rater 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 raters 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 rater monitoring.

After completing the first training set, raters 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, raters 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 raters unable to meet the qualifying standards are not permitted to score that item. Raters who reach the qualifying standard on some items but not others will only score the items on which they have successfully qualified. All raters understand this stipulation when they are hired.

Training is carefully orchestrated so that raters 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 score all responses accurately. 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.** Raters 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.** Raters train and qualify on a baseline set within a specific grade band and target.
- **Mathematics.** Raters train on baseline items, which qualify the raters for that item as well as any items associated with it; for items with no associated items, training is for the specific item.

Rater 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. Raters generally work 6.5 hours per day, excluding breaks. Evening shift raters work 3.75 hours, excluding breaks.

Multiple strategies are used to minimize rater bias. First, raters do not have access to any student identifiers. Unless the students sign their names, write about their home towns, or in some way provide other identifying information as part of their response, the raters have no knowledge of student characteristics. Second, all raters 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 raters' 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, raters are monitored and any instances of raters making scoring decisions based on anything except the criteria are discussed. Raters are further monitored, and if any continue to exhibit bias after receiving a reasonable amount of feedback, they are dismissed.

MI also implements a series of automated score verifications to ensure the accuracy of scores. For example, MI conducts a blank check that resets scores when a condition code of "blank" is assigned to a response that has one or more characters in the response string (e.g., a response made up of spaces or tabs). In this case, the score is recorded only after three independent raters have assigned a condition code of "blank" to a response that appears blank but includes characters in the response string. A similar check is run when a score or condition code other than "blank" is assigned to a response that includes no characters in the response string. Automatic resetting of double-scored responses when two raters assign non-adjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score provides an additional score verification. In addition to automatically resetting and rescoring these responses, the rater information is captured in a report and reviewed by scoring directors, as one of many tools used to determine re-training needs.

6.7.3 Rater 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 rater.

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 rater's work throughout every project. Rater status reports are used to monitor raters' scoring habits during the Smarter Balanced handscoring project.

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

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

- Rater 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 rater
- Percentage of responses scored that day within one point of agreement with a second rater
- 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 used 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 re-training of individual raters or the group as a whole. If a rater is consistently scoring high or low, this can be easily determined along with the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all re-training efforts.

6.7.4 Rater Monitoring and Re-Training

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

Re-training is an ongoing process once scoring is underway. Daily analysis of the rater status reports enables management personnel to identify individual or group re-training 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 re-training procedures include room-wide discussions led by the scoring director, team discussions conducted by team leaders, and one-on-one discussions with individual raters. It is standard practice to conduct morning room-wide re-training at MI each day, with a more extensive re-training on Monday mornings in order to re-anchor the raters after a weekend away from scoring.

Each student response is scored holistically by a trained and qualified rater 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 raters who are not aware of the score assigned by the first rater or even that the response has been read before. MI's QA/reliability procedures allow the handscoring staff to identify struggling raters very early and begin re-training at once. While re-training these raters, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a re-training method. MI shows raters' responses that the raters have scored incorrectly, explains the correct scores, and has the raters change the scores.

During scoring, raters 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 Validity Checks

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

MI staff and all clients have access to real-time validity reports that include the response identification number, the scores assigned by the raters, 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. Re-training may be conducted with the raters using the validity data as a guide for how to focus the re-training. 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 re-training and dismissing raters.

MI has amassed a large, longitudinal dataset of rater performance data from years of Smarter Balanced handscoring. In spring 2019, MI launched an enhanced accuracy monitoring system drawing on these data. This system used validity responses, calibrated to fit a unidimensional item response theory (IRT) model for each content area/item type. Calibrating validity responses allows us to prioritize them (using correlations and fit statistics) so that those responses that provide the greatest information about rater accuracy are distributed to raters first. MI runs nightly analyses to evaluate performance nightly during scoring. Empirically-determined cutpoints are used to classify raters into performance tiers based on recent validity and inter-rater reliability (IRR). A rater with unacceptable performance initially receives feedback and additional monitoring in the form of increased read-behinds. If performance does not improve quickly, the rater is assigned an assessment composed of validity responses, the results of which determine whether the rater may continue to score.

6.7.6 Rater Dismissal

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

6.7.7 Rater Agreement

The inter-rater reliability (IRR) is computed-based on scorable responses (numeric scores) and scored by two independent raters 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 raters. The IRR is computed based on the raters who scored student responses in Connecticut.

In ELA/L, the short answer items are scored in 0–2. Mathematics SA items are scored using 0–1, 0–2, or 0–3 rubrics.

Tables 39–40 summarize the inter-rater reliability based on items with a sample size greater than 50. The inter-rater reliability is presented with average of percent exact agreement, minimum and maximum percent exact agreements, combined percent exact and percent adjacent agreement, and quadratic weighted Kappa (QWK).

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

Grade	# of Items	% Exact			% (Exact+ Adjacent)	QWK
		Average	Min	Max		
3	18	84	76	95	100	0.72
4	30	82	71	95	100	0.73
5	22	79	64	91	100	0.71
6	18	75	70	88	100	0.67
7	23	73	64	87	100	0.64
8	24	74	61	91	100	0.66

Table 40. Mathematics Rater Agreements

Grade	Score Points	# of Items	% Exact			% (Exact+ Adjacent)	QWK
			Average	Min	Max		
3	1	10	92	88	94	100	0.81
3	2	31	90	79	100	100	0.92
3	3	6	93	89	99	100	0.97
4	1	11	85	76	94	100	0.67
4	2	42	89	73	98	100	0.90
4	3	4	88	86	89	100	0.95
5	1	5	93	90	98	100	0.64
5	2	51	89	78	98	100	0.88
5	3	8	85	78	99	100	0.87
6	1	12	98	96	100	100	0.87
6	2	41	89	76	97	100	0.89
7	1	8	96	90	100	100	0.77
7	2	25	89	82	94	100	0.86
7	3	1	82	82	82	100	0.89
8	1	15	92	82	99	100	0.81
8	2	26	91	83	99	100	0.89

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 handscored, 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 ORS 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.

This section describes the types of scores reported in the ORS and 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 that are 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 (SEM) • 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 Alaskan Native Asian Black or African American Hispanic or Latino Native Hawaiian or Other Pacific Islander White Demographic Race Two or More Races

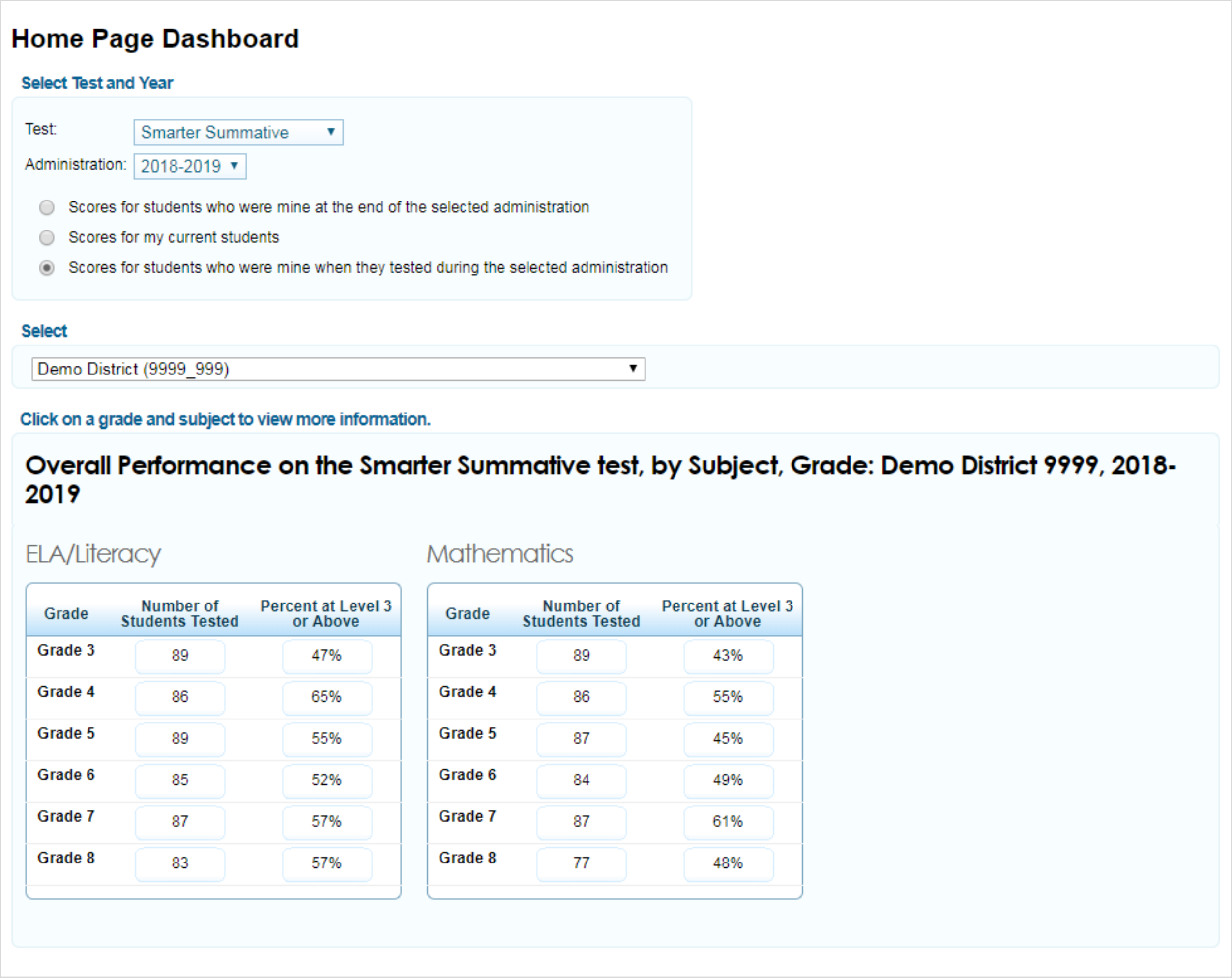
7.1.2 The Online Reporting System

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



Home Page Dashboard

Select Test and Year

Test: Smarter Summative ▼

Administration: 2018-2019 ▼

○ Scores for students who were mine at the end of the selected administration

○ Scores for my current students

● Scores for students who were mine when they tested during the selected administration

Select

Demo District (9999_999) ▼

Click on a grade and subject to view more information.

Overall Performance on the Smarter Summative test, by Subject, Grade: Demo District 9999, 2018-2019

ELA/Literacy

Grade	Number of Students Tested	Percent at Level 3 or Above
Grade 3	89	47%
Grade 4	86	65%
Grade 5	89	55%
Grade 6	85	52%
Grade 7	87	57%
Grade 8	83	57%

Mathematics

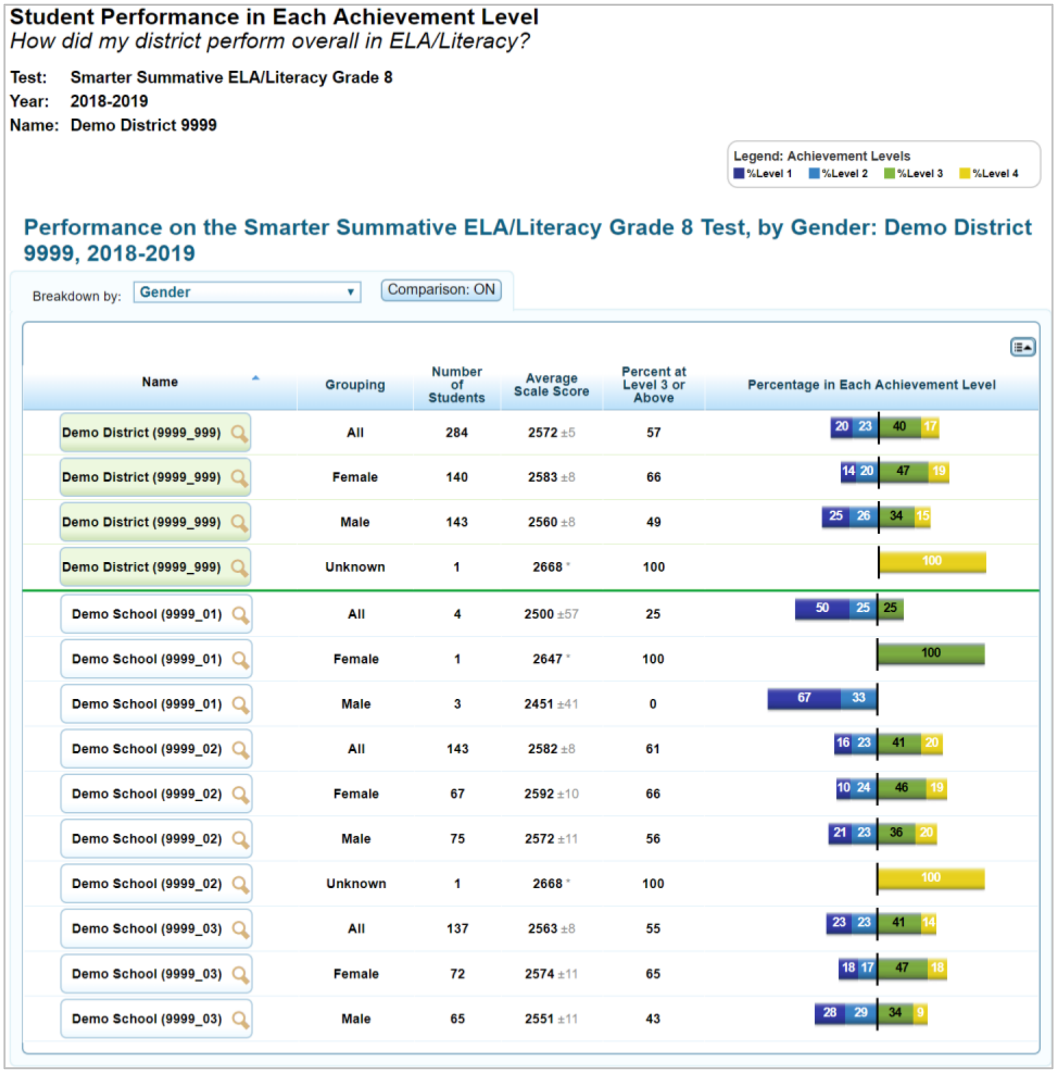
Grade	Number of Students Tested	Percent at Level 3 or Above
Grade 3	89	43%
Grade 4	86	55%
Grade 5	87	45%
Grade 6	84	49%
Grade 7	87	61%
Grade 8	77	48%

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 tested, (2) average scale score and standard error associated with 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 8
Year: 2018-2019
Name: Demo District 9999

Legend: Achievement Levels
🟦 %Level 1 🟦 %Level 2 🟩 %Level 3 🟨 %Level 4

Performance on the Smarter Summative ELA/Literacy Grade 8 Test, by Gender: Demo District 9999, 2018-2019

Breakdown by: Gender
Comparison: ON

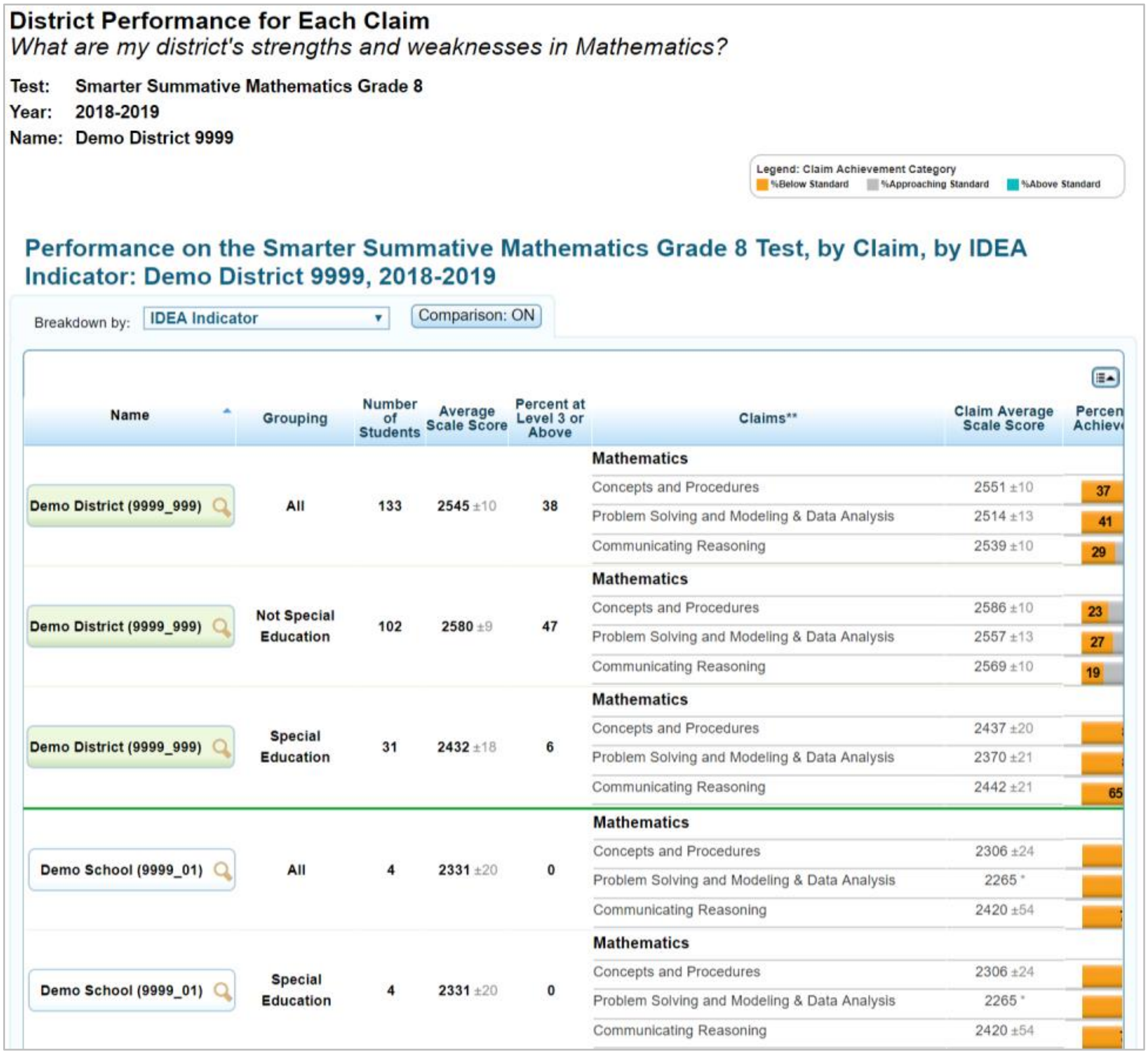
Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Percentage in Each Achievement Level
Demo District (9999_999)	All	284	2572 ±5	57	20 23 40 17
Demo District (9999_999)	Female	140	2583 ±8	66	14 20 47 19
Demo District (9999_999)	Male	143	2560 ±8	49	25 26 34 15
Demo District (9999_999)	Unknown	1	2668 *	100	100
Demo School (9999_01)	All	4	2500 ±57	25	50 25 25
Demo School (9999_01)	Female	1	2647 *	100	100
Demo School (9999_01)	Male	3	2451 ±41	0	67 33
Demo School (9999_02)	All	143	2582 ±8	61	16 23 41 20
Demo School (9999_02)	Female	67	2592 ±10	66	10 24 46 19
Demo School (9999_02)	Male	75	2572 ±11	56	21 23 36 20
Demo School (9999_02)	Unknown	1	2668 *	100	100
Demo School (9999_03)	All	137	2563 ±8	55	23 23 41 14
Demo School (9999_03)	Female	72	2574 ±11	65	18 17 47 18
Demo School (9999_03)	Male	65	2551 ±11	43	28 29 34 9

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 tested, (2) average scale score and standard error associated with the average scale score, (3) percentage of students at Level 3 or above, and (4) percentage of students in each claim performance category.

As with 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 IDEA Indicator.

Exhibit 3. Claim Detail Page for Mathematics by IDEA Indicator: District Level



District Performance for Each Claim

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

Test: Smarter Summative Mathematics Grade 8
Year: 2018-2019
Name: Demo District 9999

Legend: Claim Achievement Category 🟧 %Below Standard ⬜ %Approaching Standard 🟦 %Above Standard

Performance on the Smarter Summative Mathematics Grade 8 Test, by Claim, by IDEA Indicator: Demo District 9999, 2018-2019

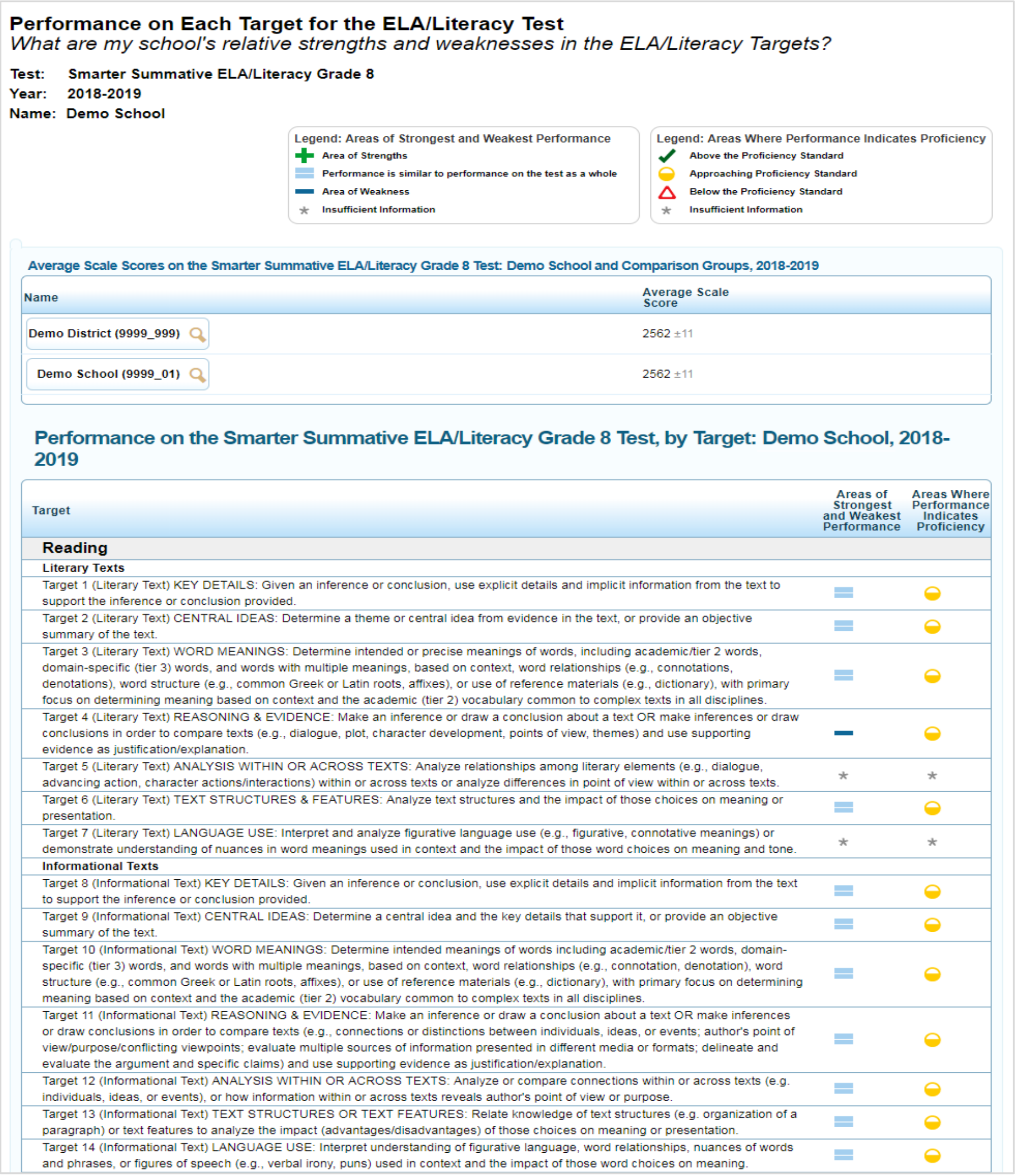
Breakdown by: IDEA Indicator ▾ Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Claims**	Claim Average Scale Score	Percen Achiev
Demo District (9999_999)	All	133	2545 ±10	38	**Mathematics**		
					Concepts and Procedures	2551 ±10	37
					Problem Solving and Modeling & Data Analysis	2514 ±13	41
					Communicating Reasoning	2539 ±10	29
Demo District (9999_999)	Not Special Education	102	2580 ±9	47	**Mathematics**		
					Concepts and Procedures	2586 ±10	23
					Problem Solving and Modeling & Data Analysis	2557 ±13	27
					Communicating Reasoning	2569 ±10	19
Demo District (9999_999)	Special Education	31	2432 ±18	6	**Mathematics**		
					Concepts and Procedures	2437 ±20	
					Problem Solving and Modeling & Data Analysis	2370 ±21	
					Communicating Reasoning	2442 ±21	65
Demo School (9999_01)	All	4	2331 ±20	0	**Mathematics**		
					Concepts and Procedures	2306 ±24	
					Problem Solving and Modeling & Data Analysis	2265 *	
					Communicating Reasoning	2420 ±54	
Demo School (9999_01)	Special Education	4	2331 ±20	0	**Mathematics**		
					Concepts and Procedures	2306 ±24	
					Problem Solving and Modeling & Data Analysis	2265 *	
					Communicating Reasoning	2420 ±54	

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 8**
Year: **2018-2019**
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

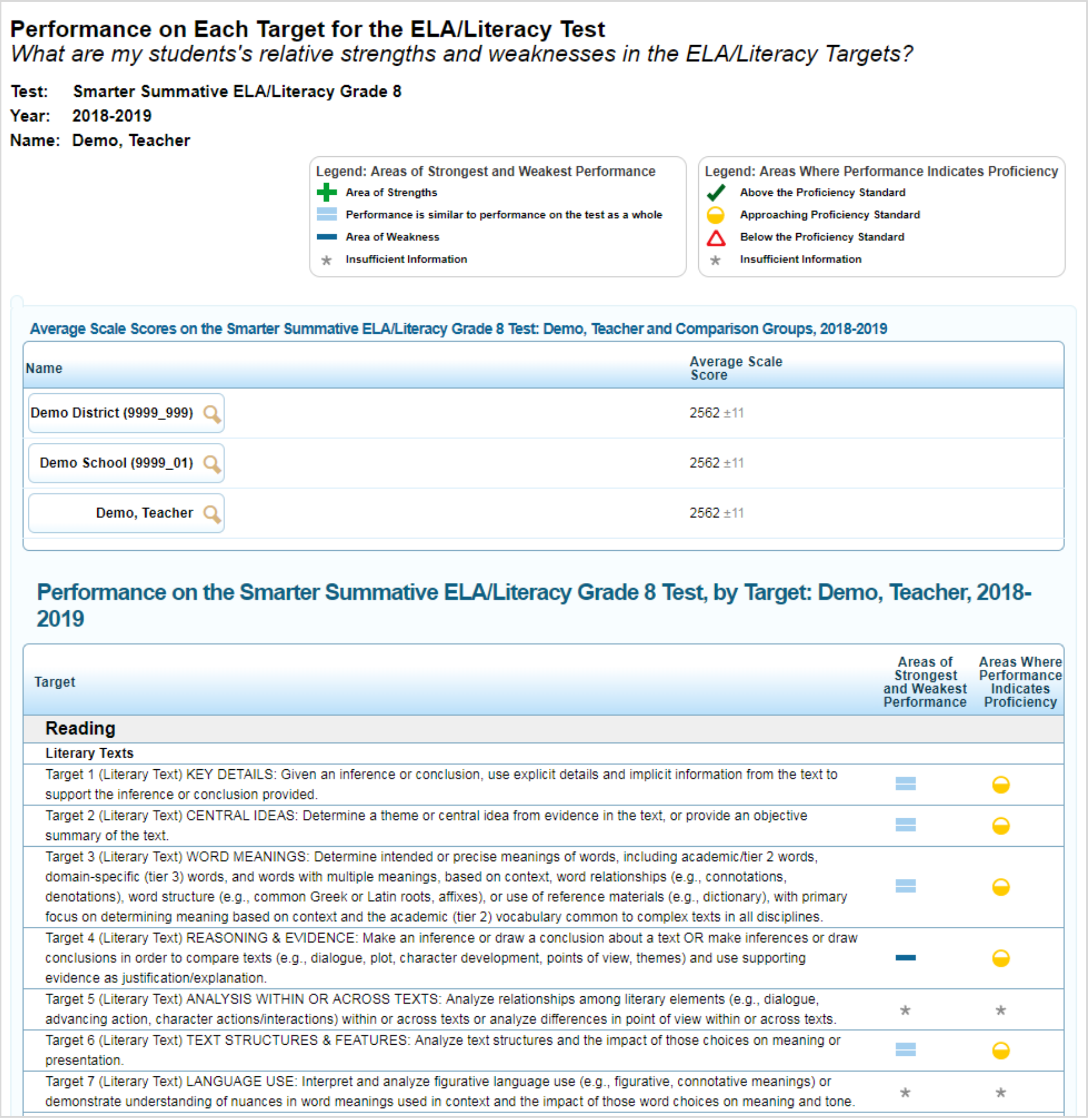
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 8 Test: Demo School and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2562 ±11
Demo School (9999_01)	2562 ±11

Performance on the Smarter Summative ELA/Literacy Grade 8 Test, by Target: Demo School, 2018-2019

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 evidence in the text, or provide an objective summary of the text.	═	◒
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., dialogue, plot, character development, points of view, themes) and use supporting evidence as justification/explanation.	▬	◒
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Analyze relationships among literary elements (e.g., dialogue, advancing action, character actions/interactions) within or across texts or analyze differences in 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 an objective summary of the text.	═	◒
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., connotation, denotation), 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., connections or distinctions between individuals, ideas, or events; author's point of view/purpose/conflicting viewpoints; evaluate multiple sources of information presented in different media or formats; delineate and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	═	◒
Target 12 (Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Analyze or compare connections within or across texts (e.g. individuals, ideas, or events), or how information within or across texts reveals author's point of view or purpose.	═	◒
Target 13 (Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. organization of a paragraph) or text features to analyze the impact (advantages/disadvantages) 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., verbal irony, puns) used in context and the impact of those word choices on meaning.	═	◒

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



Performance on Each Target for the ELA/Literacy Test
What are my students's relative strengths and weaknesses in the ELA/Literacy Targets?

Test: Smarter Summative ELA/Literacy Grade 8
Year: 2018-2019
Name: Demo, Teacher

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

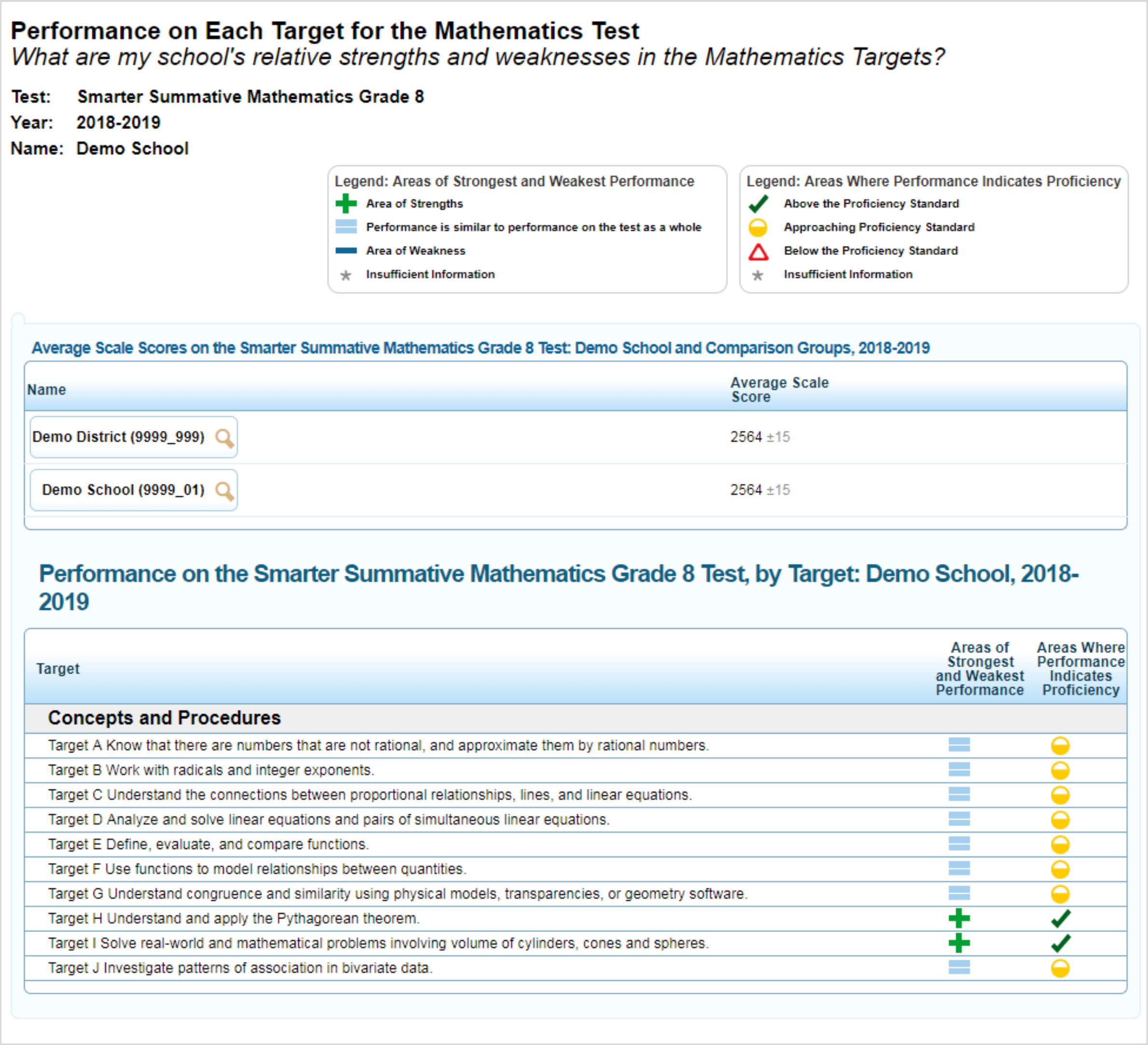
Average Scale Scores on the Smarter Summative ELA/Literacy Grade 8 Test: Demo, Teacher and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2562 ±11
Demo School (9999_01)	2562 ±11
Demo, Teacher	2562 ±11

Performance on the Smarter Summative ELA/Literacy Grade 8 Test, by Target: Demo, Teacher, 2018-2019

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 evidence in the text, or provide an objective summary of the text.	═	◒
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., dialogue, plot, character development, points of view, themes) and use supporting evidence as justification/explanation.	▬	◒
Target 5 (Literary Text) ANALYSIS WITHIN OR ACROSS TEXTS: Analyze relationships among literary elements (e.g., dialogue, advancing action, character actions/interactions) within or across texts or analyze differences in 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.	✱	✱

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 8
Year: 2018-2019
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

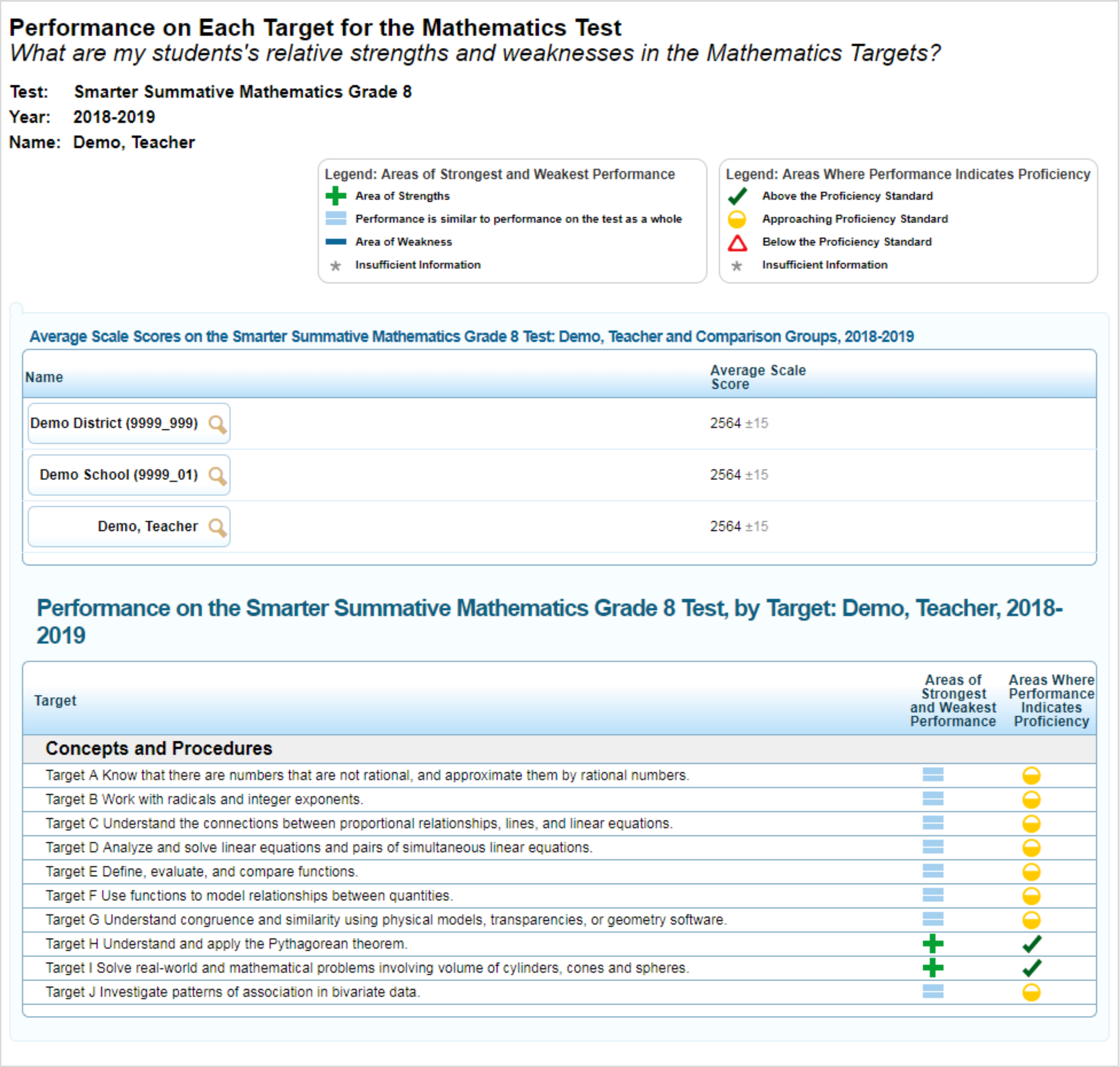
Average Scale Scores on the Smarter Summative Mathematics Grade 8 Test: Demo School and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2564 ±15
Demo School (9999_01)	2564 ±15

Performance on the Smarter Summative Mathematics Grade 8 Test, by Target: Demo School, 2018-2019

Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Concepts and Procedures		
Target A Know that there are numbers that are not rational, and approximate them by rational numbers.	☰	◒
Target B Work with radicals and integer exponents.	☰	◒
Target C Understand the connections between proportional relationships, lines, and linear equations.	☰	◒
Target D Analyze and solve linear equations and pairs of simultaneous linear equations.	☰	◒
Target E Define, evaluate, and compare functions.	☰	◒
Target F Use functions to model relationships between quantities.	☰	◒
Target G Understand congruence and similarity using physical models, transparencies, or geometry software.	☰	◒
Target H Understand and apply the Pythagorean theorem.	✚	✓
Target I Solve real-world and mathematical problems involving volume of cylinders, cones and spheres.	✚	✓
Target J Investigate patterns of association in bivariate data.	☰	◒

Exhibit 7. Target Detail Page for Mathematics: Teacher Level



Performance on Each Target for the Mathematics Test
What are my students's relative strengths and weaknesses in the Mathematics Targets?

Test: Smarter Summative Mathematics Grade 8
Year: 2018-2019
Name: Demo, Teacher

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 8 Test: Demo, Teacher and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2564 ±15
Demo School (9999_01)	2564 ±15
Demo, Teacher	2564 ±15

Performance on the Smarter Summative Mathematics Grade 8 Test, by Target: Demo, Teacher, 2018-2019

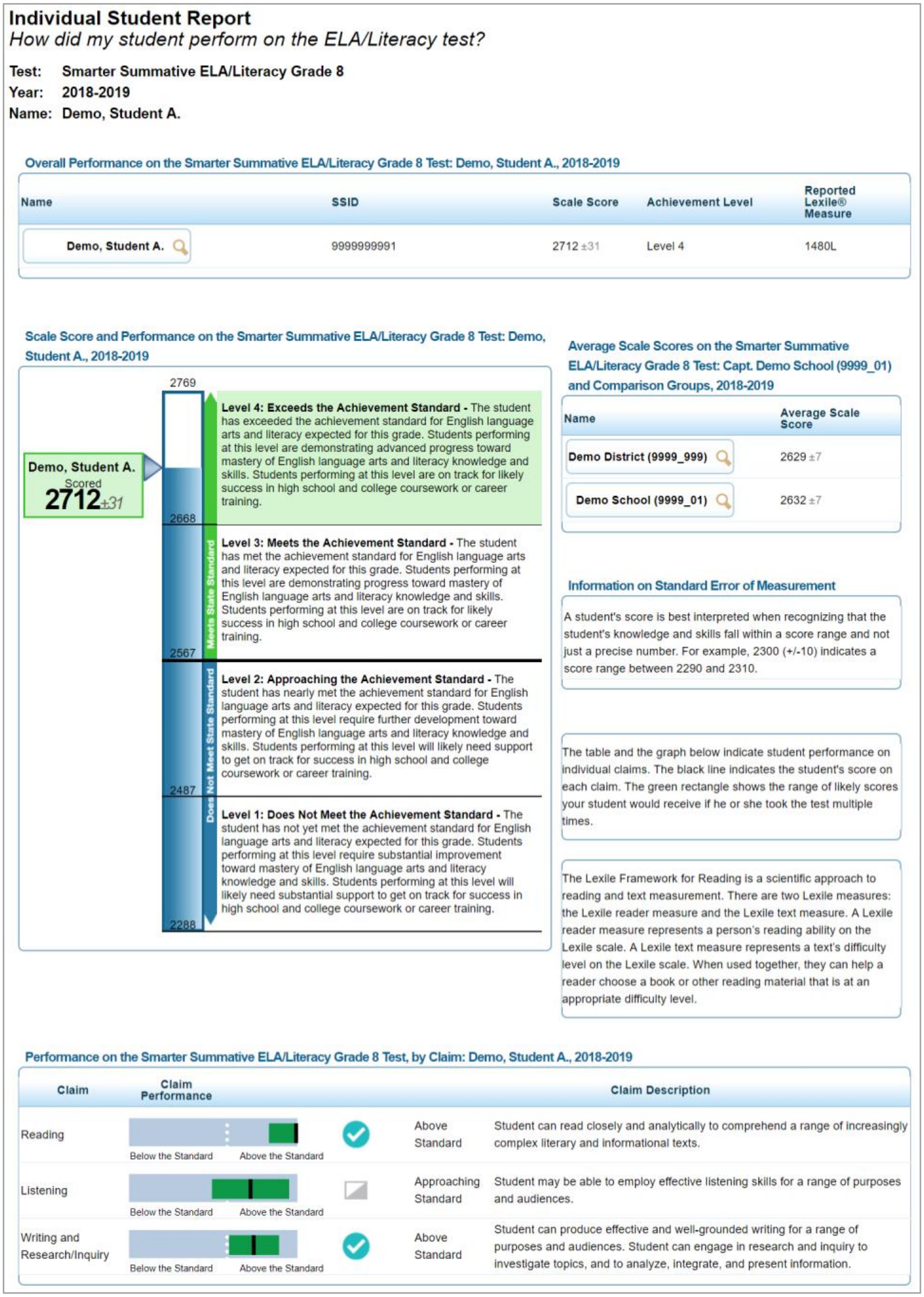
Target	Areas of Strongest and Weakest Performance	Areas Where Performance Indicates Proficiency
Concepts and Procedures		
Target A Know that there are numbers that are not rational, and approximate them by rational numbers.	▭	◓
Target B Work with radicals and integer exponents.	▭	◓
Target C Understand the connections between proportional relationships, lines, and linear equations.	▭	◓
Target D Analyze and solve linear equations and pairs of simultaneous linear equations.	▭	◓
Target E Define, evaluate, and compare functions.	▭	◓
Target F Use functions to model relationships between quantities.	▭	◓
Target G Understand congruence and similarity using physical models, transparencies, or geometry software.	▭	◓
Target H Understand and apply the Pythagorean theorem.	✚	✓
Target I Solve real-world and mathematical problems involving volume of cylinders, cones and spheres.	✚	✓
Target J Investigate patterns of association in bivariate data.	▭	◓

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 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 that define 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 the 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 8**
Year: **2018-2019**
Name: **Demo, Student A.**

Overall Performance on the Smarter Summative ELA/Literacy Grade 8 Test: Demo, Student A., 2018-2019

Name	SSID	Scale Score	Achievement Level	Reported Lexile® Measure
Demo, Student A.	9999999991	2712 ±31	Level 4	1480L

Scale Score and Performance on the Smarter Summative ELA/Literacy Grade 8 Test: Demo, Student A., 2018-2019

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

Level 3: Meets the Achievement Standard - 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.

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

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

Average Scale Scores on the Smarter Summative ELA/Literacy Grade 8 Test: Capt. Demo School (9999_01) and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2629 ±7
Demo School (9999_01)	2632 ±7

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.

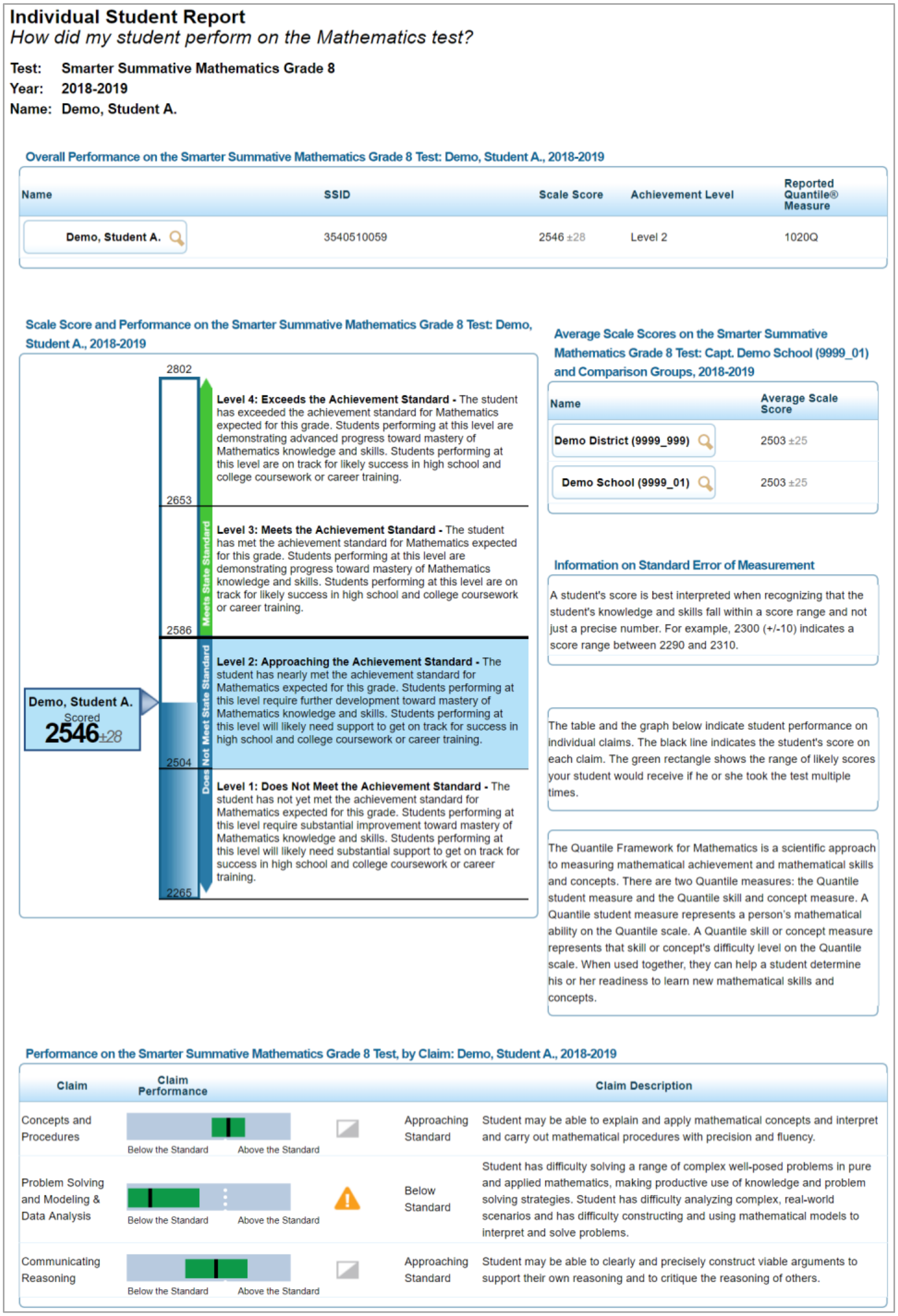
The table and the graph below indicate student performance on individual claims. The black line indicates the student's score on each claim. The green rectangle shows the range of likely scores your student would receive if he or she took the test multiple times.

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

Performance on the Smarter Summative ELA/Literacy Grade 8 Test, by Claim: Demo, Student A., 2018-2019

Claim	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	◪	Approaching Standard	Student may be able to 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 8**
Year: **2018-2019**
Name: **Demo, Student A.**

Overall Performance on the Smarter Summative Mathematics Grade 8 Test: Demo, Student A., 2018-2019

Name	SSID	Scale Score	Achievement Level	Reported Quantile® Measure
Demo, Student A.	3540510059	2546 ±28	Level 2	1020Q

Scale Score and Performance on the Smarter Summative Mathematics Grade 8 Test: Demo, Student A., 2018-2019

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.

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.

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.

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.

Average Scale Scores on the Smarter Summative Mathematics Grade 8 Test: Capt. Demo School (9999_01) and Comparison Groups, 2018-2019

Name	Average Scale Score
Demo District (9999_999)	2503 ±25
Demo School (9999_01)	2503 ±25

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.

The Quantile Framework for Mathematics is a scientific approach to measuring mathematical achievement and mathematical skills and concepts. There are two Quantile measures: the Quantile student measure and the Quantile skill and concept measure. A Quantile student measure represents a person's mathematical ability on the Quantile scale. A Quantile skill or concept measure represents that skill or concept's difficulty level on the Quantile scale. When used together, they can help a student determine his or her readiness to learn new mathematical skills and concepts.

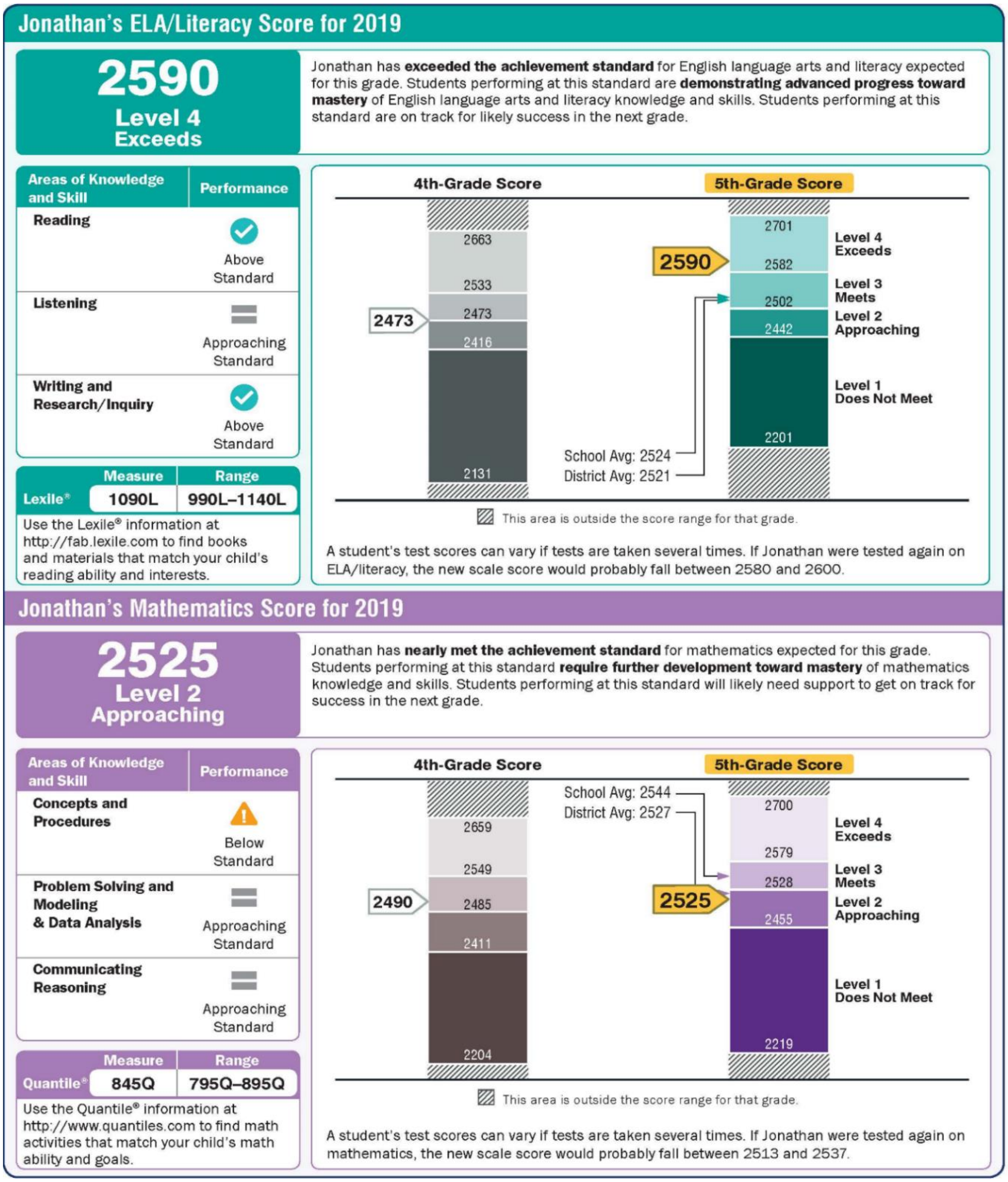
Performance on the Smarter Summative Mathematics Grade 8 Test, by Claim: Demo, Student A., 2018-2019

Claim	Claim Performance			Claim Description
Concepts and Procedures	Below the Standard Above the Standard		Approaching Standard	Student may be able to 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	⚠	Below Standard	Student has difficulty solving a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. Student has difficulty analyzing complex, real-world scenarios and has difficulty constructing and using mathematical models to interpret and solve problems.
Communicating Reasoning	Below the Standard Above the Standard		Approaching Standard	Student may be able to clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.

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 10.

Exhibit 10. Sample Paper Family Score Report



Jonathan's ELA/Literacy Score for 2019

2590
Level 4
Exceeds

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

Areas of Knowledge and Skill	Performance
Reading	✓ Above Standard
Listening	= Approaching Standard
Writing and Research/Inquiry	✓ Above Standard

	Measure	Range
Lexile®	1090L	990L–1140L

Use the Lexile® information at http://fab.lexile.com to find books and materials that match your child's reading ability and interests.

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

A student's test scores can vary if tests are taken several times. If Jonathan were tested again on ELA/literacy, the new scale score would probably fall between 2580 and 2600.

Jonathan's Mathematics Score for 2019

2525
Level 2
Approaching

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

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

	Measure	Range
Quantile®	845Q	795Q–895Q

Use the Quantile® information at http://www.quantiles.com to find math activities that match your child's math ability and goals.

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

A student's test scores can vary if tests are taken several times. If Jonathan were tested again on mathematics, the new scale score would probably fall between 2513 and 2537.

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 will 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 decide 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 [CAT] component plus a fixed-form performance task [PT] 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) is created so that all possible responses and all demographic grids are 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 re-training 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 re-training sessions when necessary. MI's QA procedures allow scoring staff to identify struggling scorers very quickly and to begin re-training 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. Re-training 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 re-training 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 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 unexpectedly low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

8.5.1 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 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 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 the QA system's validation checks. All of the above processes take milliseconds to complete; within less than one second after AIR receives handscores and they pass 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. The scripts are released for production when the output from both teams matches exactly.

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 perform the many complicated calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

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

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York, NY: John Wiley & Sons.

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.

U.S. Department of Education (2015). *Peer Review of State Assessment Systems: Non-Regulatory Guidance for States*. Washington, DC: https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf

APPENDICES

Appendix A: Summary of the 2018–2019 Interim Assessments

The Interim Comprehensive Assessments (ICA) were fixed-form tests for each grade and subject. Most students took the ICA once, but some students took it multiple times. Table A-1 presents the number of students who took the ICA by the number of attempts. Total number of tests indicate the total ICA tests taken by the total number of students, counting multiple attempts as multiple tests. For example, if a student took the ICA twice, the number of tests for this student is counted twice. Table A-2 summarizes student performance on the ICA for all tests taken, 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	Number of Students by Number of Attempts						Total Number of Tests Taken
	Once	Twice	Three Times	Four Times	Five Times	Total Number of Students	
ELA/L							
3	421	7	49	0	0	477	582
4	156	0	51	0	0	207	309
5	261	1	0	0	0	262	263
6	375	1	0	0	0	376	377
7	178	0	0	0	0	178	178
8	304	1	0	0	0	305	306
Mathematics							
3	366	13	47	0	0	426	533
4	122	1	51	0	0	174	277
5	161	1	0	0	0	162	163
6	100	1	0	0	0	101	102
7	98	0	0	0	0	98	98
8	94	0	0	0	0	94	94

* No attempted tests in grade 11

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

Subject	Grade	Total Number of Tests Taken	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
ELA/L	3	582	2391	79	40	32	14	14	28
	4	309	2468	88	29	24	19	28	47
	5	263	2498	101	31	18	29	22	51
	6	377	2506	82	25	36	31	9	40
	7	178	2552	90	21	26	37	16	53
	8	306	2565	85	15	36	38	11	49
Math	3	533	2404	72	36	34	21	8	30
	4	277	2482	86	19	35	24	22	47
	5	163	2474	89	40	31	15	13	28
	6	102	2557	97	20	26	20	34	54
	7	98	2652	97	4	20	18	57	76
	8	94	2650	85	6	12	32	50	82

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

For the Interim Assessment Block assessments (IABs), there were seven to nine IABs for ELA/L and six to ten 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 29,397 students took the IABs, and among 29,397 students, 10,904 students took one IAB, 7,207 students took two IABs, and so on.

Tables A–4 to A–7 disaggregate the number of students in Table A–2 by each individual block. For example, 10,904 students in grade 3 ELA/L took one IAB only. Among 10,904 students, 876 of the students took the Brief Writes IAB, 839 students took the Editing IAB, and so on. Tables A–8 to A–11 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	29,397	10,904	7,207	4,445	2,785	1,593	889	700	845	29
4	28,867	11,190	7,389	4,719	1,982	1,568	940	531	501	47
5	29,811	11,181	8,060	4,985	2,224	1,075	750	902	549	85
6	28,034	10,382	6,880	5,535	2,439	1,315	724	646	112	1
7	25,796	10,076	6,353	5,100	1,738	890	961	596	77	5
8	25,367	8,998	9,229	4,005	1,097	1,730	218	90		
11	126	125	1							
Mathematics										
3	31,900	12,131	7,342	6,364	3,342	2,585	136			
4	32,020	13,116	7,246	7,192	2,475	1,808	183			
5	31,146	12,335	8,286	6,246	2,157	1,974	148			
6	29,964	12,272	7,946	7,201	1,668	823	54			
7	27,966	10,109	8,761	6,272	1,890	930	4			
8	25,736	10,596	7,600	5,545	1,375	528	92			
11	1	1								

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	876	464	775	1,144	890	605	425	810	29
	Editing	839	923	1,491	1,597	1,180	765	666	837	29
	Language and Vocabulary Use	1,394	1,350	1,481	1,259	956	527	606	842	29
	Listening and Interpretation	1,159	1,277	1,936	1,806	1,237	687	664	838	29
	Reading Informational Text	3,565	5,443	3,478	2,040	1,261	849	668	839	29
	Reading Literary Text	1,644	4,382	3,108	2,003	1,249	817	671	845	29
	Research	1,346	212	570	373	483	485	561	781	29
	Revision	70	293	462	875	580	540	585	828	29
	Performance Task	11	70	34	43	129	59	54	140	29
4	Brief Writes	251	256	491	590	599	383	329	497	47
	Editing	840	1,057	1,651	1,181	1,217	806	510	501	47
	Language and Vocabulary Use	1,268	1,560	1,797	989	1,038	695	482	501	47
	Listening and Interpretation	1,287	1,411	2,004	1,157	1,043	805	502	500	47
	Reading Informational Text	4,029	5,291	3,603	1,645	1,418	847	510	497	47
	Reading Literary Text	1,560	4,566	3,439	1,397	1,213	830	511	501	47
	Research	1,713	326	756	438	532	529	397	492	47
	Revision	236	205	365	516	746	712	457	501	47
	Performance Task	6	106	51	15	34	33	19	18	47
5	Brief Writes	428	432	558	587	480	408	545	544	85
	Editing	617	1,120	1,639	1,124	893	651	849	547	85
	Language and Vocabulary Use	1,126	1,806	1,744	1,193	620	551	736	548	85
	Listening and Interpretation	1,935	1,073	1,978	1,121	765	574	877	548	85
	Reading Informational Text	3,842	5,899	3,574	1,707	972	632	879	549	85
	Reading Literary Text	1,678	4,804	3,354	1,532	762	549	797	537	85
	Research	1,357	552	954	821	320	494	770	473	85
	Revision	172	406	983	803	549	616	814	548	85
	Performance Task	26	28	171	8	14	25	47	98	85

Table A–5: ELA/L Number of Students Who Took IABs by Block Labels (Grades 6–8, 11)

Grade	**Block**	**Number of IABs Taken**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
6	Brief Writes	92	271	364	234	184	131	84	86	1
	Editing	1,051	2,074	3,087	1,804	1,201	679	642	112	1
	Language and Vocabulary	924	1,497	936	1,285	851	568	638	112	1
	Listening and Interpretation	1,465	568	1,357	1,026	823	630	641	112	1
	Reading Informational Text	3,320	4,633	4,212	1,992	1,176	616	632	112	1
	Reading Literary Text	1,624	3,313	2,336	1,663	867	522	634	111	1
	Research	1,659	616	1,163	687	531	507	575	112	1
	Revision	224	698	3,027	1,039	907	664	640	112	1
	Performance Task	23	90	123	26	35	27	36	27	1
7	Brief Writes	489	219	316	254	220	571	301	77	5
	Editing	2,092	1,359	3,245	981	817	926	591	77	5
	Language and Vocabulary	432	942	1,005	1,230	399	759	581	77	5
	Listening and Interpretation	560	698	860	510	541	576	516	77	5
	Reading Informational Text	3,137	4,173	4,102	1,488	814	751	561	77	5
	Reading Literary Text	1,534	3,413	2,419	1,320	747	732	591	77	5
	Research	1,505	1,048	794	580	566	579	465	77	5
	Revision	325	686	2,551	585	330	862	542	75	5
	Performance Task	2	168	8	4	16	10	24	2	5
8	Brief Writes	244	400	514	365	574	218	90		
	Editing and Revising	1,430	4,329	2,904	934	1,481	212	90		
	Listening and Interpretation	228	829	1,833	629	1,657	214	90		
	Reading Informational Text	3,159	6,168	2,920	964	1,721	218	90		
	Reading Literary Text	2,242	4,945	2,727	806	1,636	214	90		
	Research	1,691	1,660	1,065	676	1,563	157	90		
	Performance Task	4	127	52	14	18	75	90		
11	Brief Writes									
	Editing	125	1							
	Language and Vocabulary Use									
	Listening and Interpretation		1							
	Reading Informational Text									
	Reading Literary Text									
	Research									
	Revision									
	Performance Task									

Table A–6: Mathematics Number of Students Who Took IABs by Block Labels (Grades 3–8)

Grade	**Block**	**Number of IABs Taken**					
		1	**2**	**3**	**4**	**5**	**6**
3	Geometry	257	704	1,054	1,964	2,519	136
	Measurement and Data	678	1,284	1,857	2,566	2,543	136
	Number and Operations in Base Ten	5,386	4,341	5,269	2,916	2,578	136
	Number and Operations – Fractions	2,023	3,521	5,420	3,016	2,582	136
	Operational and Algebraic Thinking	3,599	4,520	5,397	2,832	2,554	136
	Performance Task	188	314	95	74	149	136
4	Geometry	539	770	1,930	1,704	1,802	183
	Measurement and Data	257	944	1,202	1,373	1,765	183
	Number and Operations in Base Ten	7,502	5,281	6,432	2,343	1,806	183
	Number and Operations – Fractions	3,004	4,151	6,340	2,338	1,805	183
	Operational and Algebraic Thinking	1,690	3,165	5,586	2,017	1,804	183
	Performance Task	124	181	86	125	58	183
5	Geometry	931	889	979	1,193	1,967	148
	Measurement and Data	511	1,153	3,348	1,764	1,959	148
	Number and Operations in Base Ten	6,072	6,547	5,485	1,918	1,972	148
	Number and Operations – Fractions	3,958	4,771	5,840	2,012	1,972	148
	Operations and Algebraic Thinking	812	2,981	3,019	1,630	1,962	148
	Performance Task	51	231	67	111	38	148
6	Expressions and Equations	2,720	3,783	6,037	1,343	815	54
	Geometry	752	1,475	1,575	1,108	823	54
	Number System	3,781	4,954	6,505	1,523	709	54
	Ratios and Proportional Relationships	4,544	4,977	6,686	1,579	822	54
	Statistics and Probability	366	622	654	939	788	54
	Performance Task	109	81	146	180	158	54
7	Expressions and Equations	2,333	4,646	5,870	1,762	930	4
	Geometry	479	1,108	907	1,323	895	4
	Number System	3,009	5,653	5,708	1,783	917	4
	Ratios and Proportional Relationships	4,158	5,568	5,700	1,746	924	4
	Statistics and Probability	117	515	543	679	916	4
	Performance Task	13	32	88	267	68	4
8	Expressions and Equations I	3,318	3,127	4,722	1,240	527	92
	Expressions and Equations II	1,316	1,617	3,391	1,033	526	92
	Functions	3,505	4,514	3,229	1,285	527	92
	Geometry	1,437	2,706	3,632	810	527	92
	Number System	1,010	3,199	1,583	940	519	92
	Performance Task	10	37	78	192	14	92

Table A–7: Mathematics Number of Students Who Took IABs by Block Labels (Grade 11)

Grade	Block	Number of IABs Taken			
		1	2	3	4
11	Algebra – Linear Functions				
	Algebra – Quadratic Functions				
	Geometry – Congruence	1			
	Geometry – Measurement and Modeling				
	Geometry – Right Triangles and Trigonometric Ratios				
	Interpreting Functions				
	Number and Quantity				
	Seeing Structure in Expressions and Polynomial Expressions				
	Statistics and Probability				
	Performance Task				

Table A-8: 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	6,018	21	63	16
	Editing	8,327	26	52	22
	Language and Vocabulary Use	8,444	22	47	31
	Listening and Interpretation	9,633	17	53	31
	Reading Informational Text	18,172	23	53	24
	Reading Literary Text	14,748	29	41	30
	Research	4,840	24	45	31
	Revision	4,262	22	50	28
	Performance Task	569	29	61	9
4	Brief Writes	3,443	18	59	23
	Editing	7,810	21	54	25
	Language and Vocabulary Use	8,377	22	46	32
	Listening and Interpretation	8,756	13	58	30
	Reading Informational Text	17,887	15	53	32
	Reading Literary Text	14,064	26	50	23
	Research	5,230	29	45	26
	Revision	3,785	18	57	26
	Performance Task	329	31	55	14
5	Brief Writes	4,067	22	65	13
	Editing	7,525	15	45	40
	Language and Vocabulary Use	8,409	19	50	31
	Listening and Interpretation	8,956	15	52	34
	Reading Informational Text	18,139	10	57	33
	Reading Literary Text	14,098	17	48	35
	Research	5,826	23	45	33
	Revision	4,976	19	53	28
	Performance Task	502	25	54	21

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

Table A-9: ELA/L Percentage of Students in Achievement Levels by IAB Block Labels (Grades 6–8,11)

Grade	Block	Number Tested	% Below	% At/Near	% Above
6	Brief Writes	1,447	17	72	11
	Editing	10,651	22	59	19
	Language and Vocabulary Use	6,812	20	53	27
	Listening and Interpretation	6,623	13	50	37
	Reading Informational Text	16,694	19	55	26
	Reading Literary Text	11,071	19	55	27
	Research	5,851	22	46	33
	Revision	7,312	34	52	14
	Performance Task	388	34	51	15
7	Brief Writes	2,452	22	57	22
	Editing	10,093	16	66	17
	Language and Vocabulary Use	5,430	23	51	27
	Listening and Interpretation	4,343	19	56	26
	Reading Informational Text	15,108	23	48	29
	Reading Literary Text	10,838	20	51	29
	Research	5,619	18	55	27
	Revision	5,961	33	53	14
	Performance Task	239	44	45	11
8	Brief Writes	2,405	29	56	15
	Editing and Revising	11,380	24	53	23
	Listening and Interpretation	5,480	16	60	24
	Reading Informational Text	15,240	19	47	34
	Reading Literary Text	12,660	28	44	29
	Research	6,902	24	47	28
	Performance Task	380	38	42	21
11	Brief Writes				
	Editing	126	21	56	22
	Language and Vocabulary Use				
	Listening and Interpretation	1	100	0	0
	Reading Informational Text				
	Reading Literary Text				
	Research				
	Revision				
	Performance Task				

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

Table A–10: Mathematics Percentage of Students in Performance Categories by IAB Block Labels (Grades 3–8)

Grade	Block	Number Tested	% Below	% At/Near	% Above
3	Geometry	6,634	17	53	30
	Measurement and Data	9,064	22	40	38
	Number and Operations in Base Ten	20,626	30	39	31
	Number and Operations – Fractions	16,698	13	42	45
	Operational and Algebraic Thinking	19,038	32	46	22
	Performance Task	956	11	62	28
4	Geometry	6,928	6	59	35
	Measurement and Data	5,724	10	44	46
	Number and Operations in Base Ten	23,547	29	45	25
	Number and Operations – Fractions	17,821	24	41	35
	Operational and Algebraic Thinking	14,445	30	48	23
	Performance Task	757	11	51	38
5	Geometry	6,107	18	53	29
	Measurement and Data	8,883	20	42	38
	Number and Operations in Base Ten	22,142	31	44	25
	Number and Operations – Fractions	18,701	31	43	26
	Operations and Algebraic Thinking	10,552	20	48	33
	Performance Task	646	20	56	24
6	Expressions and Equations	14,752	25	41	34
	Geometry	5,787	28	41	31
	Number System	17,526	28	44	28
	Ratios and Proportional Relationships	18,662	35	35	30
	Statistics and Probability	3,423	12	55	33
	Performance Task	728	21	66	13
7	Expressions and Equations	15,545	20	46	34
	Geometry	4,716	9	51	40
	Number System	17,074	24	50	27
	Ratios and Proportional Relationships	18,100	21	50	29
	Statistics and Probability	2,774	14	51	35
	Performance Task	472	11	57	32
8	Expressions and Equations I	13,026	28	49	23
	Expressions and Equations II	7,975	27	42	31
	Functions	13,152	32	42	27
	Geometry	9,204	21	47	32
	Number System	7,343	20	36	43
	Performance Task	423	21	61	18

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

Table A–11: Mathematics Percentage of Students in Performance Categories by IAB Block Labels (Grade 11)

Grade	Block	Number Tested	% Below	% At/Near	% Above
11	Algebra – Linear Functions				
	Algebra – Quadratic Functions				
	Geometry – Congruence	1	100	0	0
	Geometry – Measurement and Modeling				
	Geometry – Right Triangles and Trigonometric Ratios				
	Interpreting Functions				
	Number and Quantity				
	Seeing Structure in Expressions and Polynomial Expressions				
	Statistics and Probability				
	Performance Task				

Appendix B: Student Performance Across 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				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 3																
All Students	38,942	54	2438	89	38,097	52	2432	91	37,525	53	2435	90	36,516	54	2437	91
Female	19,139	58	2447	88	18,506	56	2442	89	18,417	57	2443	88	17,890	58	2445	89
Male	19,803	50	2430	90	19,591	48	2423	91	19,108	49	2427	91	18,626	51	2429	92
African American	4,874	31	2392	81	4,841	30	2388	83	4,764	33	2395	84	4,603	34	2395	86
AmerIndian/Alaskan	90	48	2422	78	97	37	2399	83	110	50	2422	87	101	48	2416	83
Asian	2,151	74	2480	84	2,049	71	2472	84	2,022	73	2479	85	1,945	73	2481	87
Hispanic/Latino	9,854	33	2395	82	9,847	31	2390	85	10,287	32	2392	84	10,122	35	2397	87
Pacific Islander	47	38	2420	92	33	61	2444	84	46	46	2438	85	29	45	2413	70
White	20,601	67	2465	82	19,903	65	2459	83	18,889	67	2464	80	18,236	67	2464	82
Two or More Races	1,325	57	2450	87	1,327	55	2443	91	1,407	58	2445	90	1,480	58	2447	93
LEP	3,554	16	2361	70	4,011	18	2361	76	4,153	18	2360	76	4,287	22	2369	79
Special Education	4,332	17	2357	78	4,490	16	2349	78	4,871	16	2355	78	5,018	18	2358	80
Grade 4																
All Students	38,450	56	2480	96	39,228	54	2477	96	38,376	55	2479	97	37,727	55	2478	99
Female	18,805	59	2490	94	19,281	58	2487	93	18,646	59	2488	95	18,486	58	2487	96
Male	19,645	52	2471	97	19,947	50	2468	97	19,730	52	2470	99	19,239	51	2470	101
African American	4,955	31	2427	87	4,939	32	2428	88	4,854	34	2431	90	4,820	34	2432	91
AmerIndian/Alaskan	102	42	2446	98	86	47	2465	84	105	41	2451	85	104	49	2463	87
Asian	1,996	74	2526	91	2,109	76	2530	88	2,010	75	2525	89	2,015	76	2530	91
Hispanic/Latino	9,383	33	2430	89	10,078	33	2430	90	10,195	35	2432	93	10,477	35	2432	93
Pacific Islander	29	55	2486	89	42	43	2457	92	37	65	2502	93	42	55	2476	78
White	20,825	70	2511	85	20,623	67	2506	86	19,781	68	2509	87	18,857	68	2510	89
Two or More Races	1,160	59	2493	95	1,351	58	2489	92	1,394	59	2490	100	1,412	58	2489	97
LEP	2,962	14	2384	78	3,372	15	2386	80	3,776	18	2392	83	3,999	18	2391	84
Special Education	4,934	17	2390	84	5,006	17	2389	85	5,174	17	2388	86	5,443	18	2389	87

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

Group	2015–2016				2016–2017				2017-2018				2018-2019			
	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD	N	% Prof	Scale Score	SD
Grade 5																
All Students	39,010	59	2517	97	38,748	56	2512	100	39,594	58	2517	98	38,605	58	2516	100
Female	19,273	64	2531	94	19,028	61	2524	97	19,454	63	2528	95	18,733	63	2528	97
Male	19,737	53	2504	98	19,720	52	2501	102	20,140	54	2506	100	19,871	54	2506	102
African American	4,840	33	2461	90	5,019	31	2454	91	5,034	36	2467	90	4,955	36	2466	94
AmerIndian/Alaskan	112	54	2501	95	104	38	2480	97	82	55	2489	100	111	43	2479	96
Asian	2,003	77	2563	89	1,992	75	2564	96	2,109	79	2571	90	2,003	78	2568	90
Hispanic/Latino	9,201	37	2467	92	9,580	34	2461	93	10,458	38	2470	94	10,371	38	2470	95
Pacific Islander	43	63	2525	109	29	69	2526	74	49	43	2495	101	36	67	2526	101
White	21,826	72	2547	86	20,830	71	2544	89	20,476	72	2547	87	19,683	72	2547	89
Two or More Races	985	62	2528	96	1,194	62	2526	96	1,386	63	2529	95	1,446	61	2526	103
LEP	2,694	13	2411	75	2,779	9	2400	76	3,186	13	2410	79	3,387	14	2415	80
Special Education	5,070	17	2420	84	5,464	16	2416	86	5,520	18	2423	86	5,647	18	2420	88
Grade 6																
All Students	39,071	55	2536	98	39,180	54	2534	98	39,019	54	2534	101	39,588	55	2538	99
Female	18,963	60	2548	95	19,355	59	2547	95	19,152	59	2546	97	19,412	60	2550	96
Male	20,108	50	2525	100	19,825	49	2522	99	19,866	50	2522	103	20,175	51	2526	101
African American	4,881	31	2482	91	4,889	31	2483	89	5,034	32	2484	92	5,069	34	2493	91
AmerIndian/Alaskan	95	47	2527	94	105	47	2521	94	119	36	2498	99	80	41	2510	86
Asian	1,990	73	2580	90	1,980	74	2585	91	1,931	77	2591	93	2,059	79	2597	88
Hispanic/Latino	8,794	31	2481	94	9,438	31	2481	94	9,938	32	2482	95	10,575	35	2490	95
Pacific Islander	32	50	2541	105	44	45	2523	107	32	56	2533	91	45	40	2507	97
White	22,299	68	2565	87	21,699	67	2564	87	20,706	68	2565	89	20,320	68	2567	89
Two or More Races	980	56	2542	95	1,025	57	2547	95	1,259	58	2542	99	1,440	58	2547	97
LEP	2,112	6	2411	75	2,315	5	2406	73	2,502	6	2406	73	2,710	7	2415	75
Special Education	5,193	15	2438	87	5,415	14	2438	84	5,839	15	2436	89	5,759	15	2442	86

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

Group	2015–2016				2016–2017				2017-2018				2018-2019			
	N	**% Prof**	**Scale Score**	**SD**	**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	39,165	56	2559	105
Female	19,410	61	2573	96	19,056	60	2568	99	19,421	61	2572	100	19,200	61	2574	100
Male	20,675	50	2546	101	20,156	50	2544	104	19,970	49	2541	107	19,961	51	2546	107
African American	4,917	29	2502	89	4,933	30	2499	96	4,895	31	2501	97	5,068	33	2507	98
AmerIndian/Alaskan	113	43	2537	95	100	46	2539	96	95	52	2544	107	117	38	2521	98
Asian	1,994	77	2613	91	1,982	74	2607	95	1,942	76	2612	94	1,922	77	2619	97
Hispanic/Latino	8,836	32	2505	95	8,956	32	2501	99	9,757	33	2502	101	10,134	36	2510	101
Pacific Islander	43	56	2555	117	34	59	2574	111	46	59	2560	122	29	59	2573	103
White	23,119	67	2587	89	22,182	68	2586	90	21,546	68	2588	92	20,584	70	2591	93
Two or More Races	1,063	59	2566	101	1,025	56	2561	99	1,110	57	2564	104	1,311	59	2567	105
LEP	2,074	5	2430	71	2,110	5	2421	77	2,410	5	2421	79	2,429	6	2425	78
Special Education	5,232	15	2460	86	5,368	15	2455	91	5,632	15	2454	92	6,086	17	2460	93
Grade 8																
All Students	39,351	55	2574	100	40,139	54	2569	103	39,427	56	2575	103	39,372	56	2574	104
Female	19,157	62	2589	96	19,440	60	2585	98	19,178	62	2591	99	19,362	62	2591	101
Male	20,194	49	2559	102	20,699	48	2554	104	20,245	50	2560	104	20,006	50	2558	105
African American	5,068	32	2520	92	4,978	30	2513	94	4,932	33	2522	95	4,917	34	2522	96
AmerIndian/Alaskan	94	44	2556	93	108	44	2544	92	98	38	2546	96	100	50	2563	102
Asian	1,925	76	2626	93	1,973	76	2627	94	1,975	76	2629	95	1,917	78	2635	92
Hispanic/Latino	8,546	33	2519	95	9,068	32	2516	99	9,258	34	2522	98	9,883	34	2521	100
Pacific Islander	26	58	2585	106	41	61	2590	100	37	62	2595	109	48	58	2574	121
White	22,770	67	2601	90	22,921	65	2597	93	22,056	69	2605	92	21,345	69	2605	94
Two or More Races	922	59	2582	100	1,050	57	2578	102	1,071	56	2581	102	1,162	57	2581	103
LEP	1,791	4	2436	68	1,857	3	2428	71	2,112	5	2437	72	2,225	3	2432	69
Special Education	5,171	15	2473	85	5,358	14	2470	89	5,557	16	2476	89	5,790	16	2474	88

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

Group	2014–2015				2015–2016				2016–2017				2017-2018				2018-2019			
	N	% Prof	Scale Score	SD	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	36,460	55	2443	86
Female	18,701	47	2426	77	19,109	52	2438	78	18,464	53	2439	79	18,393	53	2439	81	17,877	54	2441	83
Male	19,548	49	2428	83	19,761	53	2439	84	19,552	54	2440	86	19,079	55	2442	87	18,583	56	2445	88
African American	4,943	21	2379	71	4,860	27	2391	75	4,826	29	2393	77	4,751	30	2395	79	4,597	31	2397	79
AmerIndian/Alaskan	111	36	2406	85	90	51	2431	77	96	42	2417	67	110	45	2427	77	101	47	2429	76
Asian	1,961	71	2477	80	2,147	78	2491	76	2,042	76	2490	78	2,024	79	2496	77	1,944	79	2497	81
Hispanic/Latino	9,176	24	2385	73	9,833	31	2398	75	9,817	33	2401	77	10,270	33	2400	78	10,107	35	2404	81
Pacific Islander	32	34	2416	70	46	46	2421	77	33	52	2441	77	46	50	2441	72	29	59	2432	71
White	20,829	62	2453	71	20,569	67	2463	72	19,881	66	2464	74	18,866	68	2467	74	18,202	69	2470	76
Two or More Races	1,197	49	2433	79	1,325	56	2446	77	1,321	58	2448	83	1,405	56	2448	84	1,480	57	2448	87
LEP	3,117	11	2358	68	3,546	20	2377	70	4,005	24	2385	75	4,158	24	2380	77	4,286	28	2388	79
Special Education	4,384	15	2350	80	4,324	18	2360	82	4,484	18	2361	81	4,865	19	2361	83	5,028	19	2364	82
Grade 4																				
All Students	38,829	44	2470	80	38,387	48	2478	82	39,162	50	2482	85	38,307	51	2484	85	37,675	52	2486	87
Female	19,180	43	2469	76	18,773	47	2476	78	19,254	49	2480	81	18,618	50	2482	80	18,467	51	2484	83
Male	19,649	45	2471	84	19,614	49	2480	86	19,908	51	2483	89	19,689	52	2485	90	19,206	54	2489	91
African American	4,783	17	2419	70	4,938	21	2427	72	4,927	25	2432	78	4,839	26	2434	79	4,805	28	2437	80
AmerIndian/Alaskan	115	34	2452	74	102	36	2450	87	86	43	2474	74	104	42	2462	80	104	49	2475	72
Asian	2,002	70	2523	79	1,992	73	2533	82	2,106	77	2543	78	2,007	78	2541	78	2,013	80	2547	78
Hispanic/Latino	8,929	21	2426	72	9,372	24	2434	74	10,055	29	2439	79	10,178	30	2443	79	10,454	31	2445	81
Pacific Islander	41	46	2468	96	29	55	2488	77	41	46	2465	85	37	49	2491	88	42	48	2480	69
White	21,971	57	2494	71	20,794	62	2504	72	20,598	64	2508	75	19,747	65	2511	75	18,848	67	2515	76
Two or More Races	988	46	2480	83	1,160	51	2488	81	1,349	53	2491	82	1,395	53	2491	87	1,409	56	2495	87
LEP	2,942	11	2400	70	2,954	12	2405	69	3,370	15	2411	73	3,773	19	2418	76	3,992	21	2420	79
Special Education	4,695	11	2392	76	4,916	13	2401	75	4,998	15	2402	80	5,169	16	2402	82	5,448	17	2404	83

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

Group	2014–2015				2015–2016				2016–2017				2017-2018				2018-2019			
	N	**% Prof**	**Scale Score**	**SD**	**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	38,514	47	2513	94
Female	18,980	35	2492	83	19,242	40	2500	86	18,990	42	2504	89	19,439	44	2510	89	18,690	45	2512	90
Male	20,064	38	2495	91	19,699	42	2502	93	19,666	44	2506	96	20,101	46	2510	96	19,823	48	2513	98
African American	4,889	11	2434	75	4,830	14	2440	77	4,994	16	2445	81	5,031	19	2453	82	4,940	21	2456	85
AmerIndian/Alaskan	96	20	2468	69	112	32	2488	84	101	32	2480	89	82	29	2488	78	110	28	2477	84
Asian	2,019	60	2547	87	1,999	68	2562	87	1,987	70	2570	90	2,107	74	2577	85	1,997	75	2578	86
Hispanic/Latino	8,550	15	2444	78	9,173	18	2452	80	9,545	21	2458	83	10,442	24	2466	85	10,344	27	2469	88
Pacific Islander	30	33	2499	85	43	37	2511	103	29	48	2506	83	49	33	2475	99	36	47	2509	97
White	22,499	49	2520	77	21,798	54	2530	79	20,805	57	2535	82	20,449	59	2539	82	19,644	61	2543	83
Two or More Races	961	35	2498	86	986	43	2512	91	1,195	46	2515	93	1,380	48	2520	90	1,443	49	2519	98
LEP	2,586	5	2410	70	2,688	6	2415	69	2,770	7	2417	72	3,188	9	2425	77	3,375	13	2434	79
Special Education	4,958	7	2409	77	5,055	9	2416	78	5,453	10	2418	82	5,511	12	2422	82	5,632	12	2422	85
Grade 6																				
All Students	39,870	37	2513	100	38,965	41	2521	104	39,031	44	2526	106	38,946	44	2527	107	39,488	45	2530	109
Female	19,372	37	2516	94	18,921	41	2523	99	19,287	44	2530	101	19,115	45	2531	102	19,374	47	2534	104
Male	20,498	37	2511	105	20,044	41	2519	108	19,744	43	2523	111	19,830	43	2523	112	20,113	44	2527	113
African American	4,841	12	2449	88	4,860	14	2452	95	4,864	18	2461	97	5,020	19	2464	100	5,051	22	2471	101
AmerIndian/Alaskan	121	21	2483	92	95	31	2499	94	103	37	2511	102	118	31	2495	107	81	31	2497	94
Asian	1,979	65	2584	95	1,988	66	2588	99	1,976	71	2602	99	1,929	73	2608	100	2,055	78	2616	95
Hispanic/Latino	8,577	15	2456	95	8,769	17	2461	97	9,397	20	2467	100	9,918	22	2472	101	10,537	24	2476	103
Pacific Islander	40	53	2537	111	32	41	2530	117	44	39	2524	126	32	47	2532	93	44	34	2511	93
White	23,299	48	2542	86	22,243	53	2553	89	21,627	57	2559	92	20,674	58	2561	92	20,286	59	2564	94
Two or More Races	1,013	39	2520	100	978	40	2525	101	1,020	45	2538	102	1,255	46	2536	108	1,434	46	2537	108
LEP	2,230	4	2402	88	2,107	4	2402	86	2,307	5	2405	88	2,495	5	2407	88	2,697	6	2409	92
Special Education	5,042	7	2408	95	5,158	7	2412	96	5,391	8	2413	97	5,832	9	2415	100	5,725	10	2419	102

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

Group	**2014–2015**				**2015–2016**				**2016–2017**				**2017-2018**				**2018-2019**			
	N	**% Prof**	**Scale Score**	**SD**	**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	39,002	46	2547	115
Female	18,952	38	2532	101	19,352	42	2540	102	18,969	42	2542	106	19,382	45	2546	110	19,125	46	2550	111
Male	20,049	39	2528	111	20,609	42	2536	112	20,064	43	2541	115	19,883	44	2539	117	19,873	46	2544	119
African American	5,026	14	2466	94	4,895	14	2467	95	4,906	16	2469	97	4,873	18	2473	100	5,038	20	2478	104
AmerIndian/Alaskan	88	18	2491	92	113	29	2509	89	100	27	2508	102	95	38	2521	112	116	31	2513	109
Asian	1,901	68	2605	101	1,988	71	2617	103	1,983	70	2618	106	1,939	73	2628	106	1,917	76	2636	110
Hispanic/Latino	8,270	16	2468	98	8,798	19	2477	101	8,883	20	2479	102	9,719	21	2481	104	10,072	24	2487	106
Pacific Islander	25	32	2525	101	43	44	2546	119	33	48	2569	122	46	39	2550	141	29	55	2567	113
White	22,816	50	2560	93	23,063	54	2569	93	22,106	56	2575	97	21,486	58	2578	99	20,525	61	2584	100
Two or More Races	875	40	2537	103	1,061	44	2544	108	1,022	40	2540	109	1,107	44	2550	113	1,305	50	2560	117
LEP	2,053	4	2412	87	2,057	5	2415	89	2,091	5	2416	88	2,405	5	2417	91	2,406	5	2419	88
Special Education	4,957	7	2421	93	5,189	9	2427	99	5,334	9	2430	97	5,607	9	2427	99	6,042	11	2435	101
Grade 8																				
All Students	39,764	37	2541	114	39,181	40	2551	116	39,955	42	2554	120	39,294	43	2558	120	39,216	44	2558	123
Female	19,282	38	2546	108	19,069	42	2557	110	19,350	43	2560	114	19,100	44	2564	115	19,290	45	2565	118
Male	20,482	36	2536	120	20,112	39	2546	121	20,605	40	2549	125	20,190	42	2553	125	19,922	42	2552	128
African American	5,073	12	2468	94	5,043	15	2479	100	4,950	15	2475	103	4,909	18	2483	105	4,890	19	2483	106
AmerIndian/Alaskan	106	23	2504	102	94	20	2509	107	109	28	2520	98	98	23	2518	107	98	37	2532	120
Asian	1,791	64	2621	113	1,922	69	2635	113	1,970	72	2645	114	1,975	72	2646	114	1,914	74	2653	116
Hispanic/Latino	8,203	15	2476	102	8,504	17	2485	103	9,008	19	2489	108	9,209	20	2493	108	9,811	21	2493	109
Pacific Islander	37	32	2521	112	26	31	2551	127	41	59	2593	116	37	57	2589	127	47	38	2575	130
White	23,706	48	2573	104	22,679	52	2585	104	22,831	54	2589	107	21,997	56	2595	107	21,295	57	2597	110
Two or More Races	848	35	2543	112	913	43	2559	115	1,046	43	2561	123	1,069	43	2563	118	1,161	43	2564	127
LEP	1,935	4	2416	89	1,779	3	2419	85	1,845	4	2418	90	2,101	4	2426	89	2,202	3	2418	83
Special Education	4,921	6	2429	94	5,131	7	2437	95	5,297	8	2438	101	5,527	8	2438	100	5,712	8	2438	101

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	36,516	79	89	69	65	88	71	83	58	54	82
Female	17,890	78	89	69	65	88	70	82	57	54	83
Male	18,626	79	90	69	65	88	71	84	58	54	82
African American	4,603	79	90	69	65	84	71	85	58	54	76
AmerIndian/Alaskan	101	77	92	70	63	83	68	87	58	55	70
Asian	1,945	81	88	69	65	90	74	80	57	53	87
Hispanic/Latino	10,122	79	90	69	65	84	71	86	58	54	76
Pacific Islander	29	69	89*	68	63	66*	60	71*	62	54	49*
White	18,236	78	87	69	65	88	70	77	57	54	84
Two or More Races	1,480	80	88	69	66	89	72	83	57	54	84
LEP	4,287	80	91	69	65	80	73	87	58	54	68
Special Education	5,025	83	92	69	65	84	76	89	57	53	73
Grade 4											
All Students	37,727	77	89	60	62	88	70	83	47	51	82
Female	18,486	77	88	60	62	88	69	82	47	51	82
Male	19,239	78	90	59	62	87	70	85	47	51	82
African American	4,820	78	90	60	62	85	70	86	47	51	75
AmerIndian/Alaskan	104	73	92	59	62	86	65	83	47	55	69
Asian	2,015	80	88	60	62	91	73	78	47	51	87
Hispanic/Latino	10,477	78	90	60	62	85	70	86	48	51	75
Pacific Islander	42	71	82*	58	65	80	61	73*	47	55	70
White	18,857	77	86	60	62	88	69	78	47	51	84
Two or More Races	1,412	77	89	60	62	89	69	82	48	51	83
LEP	3,999	81	92	60	62	79	75	90	48	50	62
Special Education	5,447	83	93	59	62	84	78	91	46	50	73
Grade 5											
All Students	38,605	79	90	64	72	86	71	84	52	63	80
Female	18,733	79	89	64	72	86	71	82	52	63	80
Male	19,871	79	90	64	72	86	72	85	52	63	79
African American	4,955	79	91	64	72	82	72	86	53	63	72
AmerIndian/Alaskan	111	78	91	63	73	81	70	85	52	66	72
Asian	2,003	81	86	65	73	89	73	77	52	63	85
Hispanic/Latino	10,371	79	91	64	72	83	71	86	53	63	72
Pacific Islander	36	79	92*	67*	72	84	70	86*	52*	61	78
White	19,683	78	88	64	72	86	70	79	52	63	81
Two or More Races	1,446	80	89	64	72	88	72	84	52	63	82
LEP	3,387	83	93	64	72	76	77	90	53	59	55
Special Education	5,651	84	93	64	72	81	78	91	53	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	39,588	78	89	68	73	85	69	81	57	64	77
Female	19,412	77	88	68	73	85	69	78	58	64	77
Male	20,175	78	89	68	73	84	70	83	57	64	76
African American	5,069	78	89	68	73	82	69	83	59	63	68
AmerIndian/Alaskan	80	76	91	69	69	85*	67	81	60	61	71*
Asian	2,059	80	87	67	73	88	72	75	55	64	83
Hispanic/Latino	10,575	79	90	68	73	81	71	84	58	64	68
Pacific Islander	45	79	86	71	73	85*	70	84	60	60	74*
White	20,320	77	86	68	73	85	68	75	57	65	78
Two or More Races	1,440	77	87	68	72	86	69	79	57	64	78
LEP	2,710	86	93	67	71	73	81	91	57	55	51
Special Education	5,760	83	92	68	71	81	76	89	58	59	65
Grade 7											
All Students	39,165	78	89	67	75	84	70	83	55	67	76
Female	19,200	78	87	67	74	85	69	80	55	67	77
Male	19,961	79	90	67	75	84	71	85	55	67	75
African American	5,068	79	90	67	75	80	71	85	56	65	67
AmerIndian/Alaskan	117	78	90	67	74	82	70	84	58	65	66
Asian	1,922	81	88	67	75	89	73	79	55	65	84
Hispanic/Latino	10,134	79	90	67	74	81	71	85	56	66	67
Pacific Islander	29	77	84*	71*	80*	76*	68	75*	59*	67*	73*
White	20,584	78	86	67	75	85	69	78	54	67	77
Two or More Races	1,311	78	90	67	74	84	70	83	56	66	76
LEP	2,429	87	93	67	72	74*	82	92	55	57	45*
Special Education	6,088	83	92	67	74	81	77	89	55	63	64
Grade 8											
All Students	39,372	79	88	70	77	84	71	82	59	70	76
Female	19,362	79	87	70	77	85	71	79	59	70	78
Male	20,006	80	89	70	77	83	72	83	59	70	74
African American	4,917	80	89	70	77	81	72	84	59	69	68
AmerIndian/Alaskan	100	78	87	65	77	86	70	80	55	70	76
Asian	1,917	81	84	70	77	87	73	75	57	69	82
Hispanic/Latino	9,883	80	90	70	77	81	73	85	59	69	69
Pacific Islander	48	86	91	78*	82	94	79	88	61*	77	86
White	21,345	79	86	70	77	85	70	76	58	70	77
Two or More Races	1,162	80	87	70	79	86	71	78	60	71	78
LEP	2,225	88	93	69	74	68*	84	92	56	53	31*
Special Education	5,792	83	91	69	77	81	77	88	58	66	65

*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	36,460	84	90	73	79	90	77	85	63	72	86
Female	17,877	83	90	74	79	90	76	85	63	72	85
Male	18,583	84	91	73	79	91	77	86	63	71	86
African American	4,597	84	92	73	79	86	77	88	64	70	80
AmerIndian/Alaskan	101	83	89	75	80	89	76	85	64	72	82
Asian	1,944	87	89	73	79	94	82	81	61	72	91
Hispanic/Latino	10,107	83	92	74	79	88	77	87	64	71	80
Pacific Islander	29	80	100*	74*	76	88*	72	77*	64*	72	83*
White	18,202	83	88	73	79	91	76	80	63	72	86
Two or More Races	1,480	83	90	71	78	91	77	84	62	71	86
LEP	4,286	84	92	74	78	86	78	89	64	69	78
Special Education	5,033	87	94	73	79	88	82	92	63	69	81
Grade 4											
All Students	37,675	84	90	80	79	90	78	84	72	71	86
Female	18,467	84	89	80	79	90	77	83	72	71	85
Male	19,206	85	91	80	79	91	79	85	72	71	87
African American	4,805	84	91	80	78	87	78	86	72	69	79
AmerIndian/Alaskan	104	83	90	80	78	91	75	83	71	74	76
Asian	2,013	87	88	80	79	94	82	77	71	71	92
Hispanic/Latino	10,454	84	91	80	78	88	78	86	73	70	80
Pacific Islander	42	81	88*	86	67	88*	74	77*	77	60	82*
White	18,848	84	88	80	79	90	78	79	72	71	86
Two or More Races	1,409	85	90	80	80	91	79	84	72	72	88
LEP	3,992	86	92	80	78	87	80	88	73	69	77
Special Education	5,451	88	94	80	77	89	82	91	71	68	80
Grade 5											
All Students	38,514	83	91	77	71	91	77	86	68	61	86
Female	18,690	83	90	77	71	90	76	85	69	61	86
Male	19,823	84	92	77	71	91	78	87	68	61	87
African American	4,940	85	92	77	71	87	79	89	68	59	79
AmerIndian/Alaskan	110	82	92	74	69	81	76	87	67	57	80
Asian	1,997	86	89	78	72	93	81	82	68	61	92
Hispanic/Latino	10,344	84	92	77	71	88	77	88	68	60	81
Pacific Islander	36	86	91	81*	73	99*	79	90	68*	69	84*
White	19,644	82	88	78	72	91	75	81	69	62	87
Two or More Races	1,443	84	91	77	72	92	78	86	68	61	88
LEP	3,375	87	93	77	71	87	81	91	68	60	75
Special Education	5,637	89	95	77	70	87	84	93	66	59	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	39,488	83	92	78	72	90	77	87	69	62	85
Female	19,374	83	91	78	72	89	76	86	69	62	84
Male	20,113	84	92	77	72	90	77	87	69	61	86
African American	5,051	85	93	77	71	86	78	89	70	61	76
AmerIndian/Alaskan	81	82	92	78	73	78	74	84	71	61	73
Asian	2,055	87	89	78	73	94	81	79	69	62	92
Hispanic/Latino	10,537	84	93	77	72	85	78	89	69	61	77
Pacific Islander	44	78	88	72	66	86*	70	85	64	57	78*
White	20,286	82	89	78	72	90	75	81	69	62	86
Two or More Races	1,434	83	90	77	72	91	76	86	69	61	86
LEP	2,697	90	95	75	70	84	86	93	67	55	75
Special Education	5,728	90	95	77	71	87	85	93	68	58	80
Grade 7											
All Students	39,002	84	91	76	75	91	77	87	67	65	86
Female	19,125	83	91	76	75	90	76	85	67	65	86
Male	19,873	84	92	76	75	91	78	88	66	65	87
African American	5,038	85	93	75	74	86	79	90	67	63	78
AmerIndian/Alaskan	116	83	91	75	76	96	77	87	67	68	81
Asian	1,917	87	90	75	74	95	82	83	66	64	93
Hispanic/Latino	10,072	85	92	76	74	88	78	89	66	64	80
Pacific Islander	29	83	89*	77*	69*	90	76	86*	62*	58*	87
White	20,525	82	89	76	75	91	75	82	67	66	86
Two or More Races	1,305	84	91	77	74	91	78	86	68	64	87
LEP	2,407	90	95	75	73	84	86	93	63	60	70
Special Education	6,042	89	95	75	73	88	85	93	64	61	79
Grade 8											
All Students	39,216	83	91	71	71	91	76	87	61	61	87
Female	19,290	82	90	72	71	90	75	85	61	61	86
Male	19,922	83	92	71	71	91	77	88	61	61	87
African American	4,890	85	93	72	71	85	79	90	60	60	75
AmerIndian/Alaskan	98	83	94	75	71	88	76	88	66	61	81
Asian	1,914	86	88	71	71	95	81	81	60	62	93
Hispanic/Latino	9,811	84	92	71	71	87	78	89	61	59	80
Pacific Islander	47	80	86	71	65*	89	74	80	66	46*	90
White	21,295	81	89	71	71	91	74	82	62	61	87
Two or More Races	1,161	83	90	70	70	93	76	85	62	59	88
LEP	2,204	92	95	68	70	94	89	95	52	56	74
Special Education	5,718	90	95	70	71	90	86	94	59	57	81

*The classification index is based on n<10.