Connecticut Smarter Balanced Summative Assessments

2015–2016 Technical Report

Addendum to the Smarter Balanced Technical Report

CSDE

CONNECTICUT STATE
DEPARTMENT OF EDUCATION

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

LIST OF APPENDICES

1. OVERVIEW

The Smarter Balanced Assessment Consortium (SBAC) is a state-led enterprise intended to provide leadership and resources to improve teaching and learning by creating and maintaining a system of valid, reliable, and fair next-generation assessments (Smarter Balanced assessments) aligned to the *Common Core State Standards* (CCSS) in English language arts/literacy (ELA/L) and mathematics for grades 3–8 and 11. Connecticut is among 18 member states (plus the U.S. Virgin Islands) of the Consortium that lead the development of assessments in ELA/L and mathematics. The system includes both summative assessments—using computer adaptive testing (CAT) technologies—for accountability purposes and optional interim assessments for instructional use to provide meaningful feedback and actionable data that teachers and other educators can use to help students succeed.

The Connecticut State Board of Education formally adopted the CCSS in ELA/L and mathematics on July 7, 2010. All students in Connecticut, including students with significant cognitive disabilities who are eligible to take the Connecticut Alternate Assessment, an AA-AAAS, are taught to the same academic content standards. The Connecticut CCSS define the knowledge and skills students need to be proficient in order 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 human-scored items.

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

- **Computer Adaptive Test:** An online adaptive test that provides an individualized assessment for each student.
- **Performance Task:** A task that challenges students to apply their knowledge and skills to respond to real-world problems. Performance tasks can best be described as collections of questions and activities that are coherently connected to a single theme or scenario. They are used to better measure capacities such as depth of understanding, research skills, and complex analysis that cannot be adequately assessed with selected-response or constructed-response items. Some performance task items can be scored by the computer, but most are hand-scored.

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

- Replaced the summative ELA/L and mathematics assessments in grade 11 with the SAT Reading, Writing and Language, and mathematics tests.
- Removed the summative field test items and off-grade items from the ELA/L and mathematics CAT item pool.

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

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

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

This report provides a technical summary of the 2015–2016 summative assessments in ELA/L and mathematics administered in grades 3–8 under the Connecticut Smarter Balanced assessments. The report includes eight chapters covering an overview, test administration, the 2015–2016 operational administration, validity, reliability, scoring, reporting and interpreting scores, and the quality control process. The data included in this report are based on Connecticut data for the summative assessment only. For the interim assessments, the number of students who took ICAs and IABs is provided in Appendix A.

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

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

2. Testing Administration

2.1 Testing Windows

The 2015–2016 Smarter Balanced assessments testing window spanned approximately three months for the summative assessments and nine months for the interim assessments. The paper-and-pencil fixed-form tests for summative assessments were administered concurrently during the three-month online summative window. Table 1 shows the testing windows for both online and paper-and-pencil assessments.

Table 1. 2015–2016 Testing Windows

Tests	Grade	Start Date	End Date	Mode
Summative Assessments	3–8	03/15/2016	06/10/2016	Online Adaptive Tests
	3–8	03/15/2016	06/10/2016	Paper Fixed-Form Tests
Interim Comprehensive Assessments	3–8, 11	10/09/2015	06/10/2016	Online Fixed-Form Tests
Interim Assessment Blocks	3–8, 11	10/09/2015	06/10/2016	Online Fixed-Form Tests

2.2 Test Options and Administrative Roles

The Smarter Balanced assessments are administered primarily online. To ensure that all eligible students in the tested grades were given the opportunity to take the Smarter Balanced assessments, a number of assessment options were available for the 2015–2016 administration to accommodate students' needs. Table 2 lists the testing options that were offered in 2015–2016. A testing option is selected by content area. Once an option is selected, it would apply to all tests in the content area. Once the testing option is selected, it would apply to all tests in the content area.

Table 2. Summary of Tests and Testing Options in 2015–2016

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

To ensure standardized administration conditions, Teachers (TEs) and Test Administrators (TAs) follow procedures outlined in the *Smarter ELA/L and Mathematics Online, Summative Test Administration Manual* (TAM). TEs and TAs must review the TAM prior to the beginning of testing to ensure that the testing room is prepared appropriately (e.g., removing certain classroom posters, arranging desks). Make-up procedures should be established for any students who are absent on the day(s) of testing. TEs and TAs follow required administration procedures and directions. TEs and TAs 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 (DCs), School Test Coordinators (SCs), Teachers (TEs), and Test Administrators (TAs). The main responsibilities of these key personnel are described below. More detailed descriptions can be found in the TAM provided online at this URL: http://ct.portal.airast.org/resources/.

District Administrator (DA)

The DA is a District Administrator who may add users with District Test Coordinator (DC) roles in TIDE. For example, a Director of Special Education may need DC 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 DCs. Their primary responsibility is to coordinate the administration of the Smarter Balanced assessment in the district.

District Test Coordinator (DC)

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

DCs are responsible for the following:

- Reviewing all Smarter Balanced policies and test administration documents
- Reviewing scheduling and test requirements with SCs, TEs, and TAs
- Working with SCs and Technology Coordinators (TC) to ensure that all systems, including the secure browser, are properly installed and functional
- Importing users (SCs, TEs, and TAs) into TIDE
- Verifying all student information and eligibility in TIDE
- Scheduling and administering training sessions for all SCs, TEs, TAs, and TCs
- Ensuring that all personnel are trained on how to properly administer the Smarter Balanced assessments
- Monitoring the secure administration of the test
- 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 (SC)

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

SCs are responsible for the following:

- Based on test administration windows, establishing a testing schedule with DCs, TEs, and TAs
- Working with technology staff to ensure timely computer setup and installations
- 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 test
- Monitoring testing progress during the testing window and ensuring that all students participate, as appropriate
- Investigating and reporting all testing improprieties, irregularities, and breaches reported by the TEs and TAs
- Attending to any secure material according to CSDE and Smarter Balanced policies

Teacher (TE)

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

Test Administrator (TA)

A TA is primarily responsible for administering the Smarter Balanced assessments. The TA's role does not allow access to student results and is designed for test administrators, such as technology staff, who administer tests but should 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 SCs and DCs as appropriate
- Administering the Smarter Balanced assessments

- Reporting all potential test security incidents to the SCs and DCs in a manner consistent with Smarter Balanced, CSDE, and district policies

2.2.2 Online Administration

Within the state's testing window, schools can set testing schedule, allowing students to test in intervals (e.g., multiple sessions) rather than in one long 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.

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

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

Once an assessment is started, the student must answer all test questions presented on a page before proceeding to the next page. Skipping questions is not permitted. For the online computer adaptive test (CAT), students are allowed to scroll back to review and edit previously answered items, as long as these items are in the same test session and this session has not been paused for more than 20 minutes. Students may review and edit their responses they have previously completed before submitting the assessment. During an active CAT session, if a student reviews and changes the response to a previously answered item, then all following items to which the student already responded remain the same. No new items are assigned to this student for changing the answers. For example, a student paused for 10 minutes after completing item 10. After the pause, the student went back to item 5 and changed the answer. If the response change in item 5 changed the item score from wrong to right, the student's overall score would improve; however, there will be no change in items 6–10. No pause rule is implemented for the performance tasks. The same rules that apply to the CAT for reviews and changes to responses also apply to performance tasks.

For the summative test, an assessment can be started in one component (but not completed) and completed in another component. For the CAT, the assessment must be completed within 45 calendar days of the start date, otherwise, the assessment opportunity will expire. For the Performance Tasks, the assessment must be completed within 20 calendar days of the start date.

During a test session, TEs/TAs may pause the test for a student or group of students for a break. It is up to the TEs/TAs to determine an appropriate stopping point; however, for ELA/L and mathematics CAT, the assessments cannot be paused for more than 20 minutes to ensure the integrity of the test scores or testing. If an assessment is paused for more than 20 minutes, the student must restart a new test session and starts from where the student left off. Previous responses and editing are no longer available.

The TAs or TEs must remain in the room at all times during a test session to monitor student testing. Once the test session ends, the TAs or TEs must ensure that each student has successfully logged out of the system, collect and handouts or scratch paper that students used during the assessment to securely shred them.

2.2.3 Paper-and-Pencil Test Administration

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

The DA at the district with student(s) who need to take the paper-and-pencil version must submit a request on behalf of the student who need to take the paper-and-pencil test for test materials. If the request is approved, the testing contractor will ship the appropriate test booklets and the *Paper-and-Pencil Test Administration Manual* to the district.

Separate test booklets are used for the ELA/L and mathematics. The items from the CAT and the Performance Task components are combined into one test booklet, including two sessions for CAT and one session for performance task in both content areas. The TEs and TAs were asked not to administer the ELA performance task on the paper test.

After the student has completed the assessments, the DA returns the test booklets to the testing vendor. The testing vendor scans the answer document and scores the test, including the hand-scored items.

2.2.4 Braille Test Administration

In SY 2015–2016, the online braille test was also available. The interface is described below in several formats:

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

- Mathematics items are presented to students in Nemeth braille through the CAT or the performance task via a braille embosser.

- Mathematics items are presented to students in Nemeth braille through a fixed-form CAT test. TAs could decide whether to administer the online fixed form braille test or the online Braille CAT test.

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

Before administering the online summative assessments using the braille interface, TEs/TAs must ensure that the technical requirements are met. These requirements apply to the student's computer, the TEs/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, DCs, and SCs oversee all aspects of testing at their schools and serve as the main points of contact, while TEs and TAs administer the online assessments. The online TA Certification Course, webinars, user guides, manuals, and training sites are used to train TEs and TAs about the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for the administration are provided online.

2.3.1 Online Training

Multiple training opportunities were offered to the key staff through the Internet.

TA Certification Course

All school personnel who serve as TEs and TAs are encouraged to complete an online TA Certification Course to administer assessments. This web-based course is about 30–45 minutes long and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to actually start test sessions under different scenarios. Throughout the training and at the end of the course, participants are required to answer multiple-choice questions about the information provided.

Webinars

The following three webinars were offered to the field:

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

TIDE and How to Start/Monitor Online Testing and Test Settings: The webinar provides an overview of how to navigate the Test Information Distribution Engine (TIDE) and Test Delivery System (TDS), including how to set student settings in TIDE and how to start and monitor a test session using the TA Interface.

Online Reporting System (ORS): The webinar provides an overview of the ORS, including how to retrieve student results for the Smarter Balanced spring 2016 summative assessments, manage rosters, and batch-print individual student reports.

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

Practice and Training Test Site

In January 2015, separate training sites were opened for TEs/TAs and students and remained open throughout the 2015–2016 school year. TEs/TAs can practice administering assessments and starting and ending test sessions on the TA Training Site, and students can practice taking an online assessment on the Student Practice and Training Site. The Smarter Balanced assessment practice tests mirror the corresponding summative assessments for ELA/L and mathematics. Each test provides students with a grade-specific testing experience, including a variety of question types and difficulty levels (approximately 30 items each in ELA/L and mathematics), as well as a performance task.

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

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

Manuals and User Guides

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

The *Test Coordinator Manual* provides information for DCs and SCs regarding policies and procedures for the 2016 Smarter Balanced assessments in ELA/L and mathematics.

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

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

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

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

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

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

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

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

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

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

All manuals and user guides pertaining to the 2015–2016 online testing were available on the portal, and DAs, DCs, and SCs can use the manuals and user guides to train TAs and TEs in test administration policies and procedures.

Brochures and Quick Guides

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

How to Activate a Test Session for the Interim Assessments: This document provides a quick step-by-step of 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.

ORS Log-In Quick Guide: This quick guide provides step-by-step instructions for how to log-in to the Online Reporting System (ORS) to view score reports.

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

Request Additional Orders Brochure: This brochure provides instructions for how to request additional orders for paper testing materials in TIDE. This includes orders for materials related to the paper test administration for Large Print and braille tests for the Smarter Balanced assessments.

Test Delivery System Brochure: This brochure provides an overview of the Test Delivery System (TDS) for students.

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

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

Understanding Rosters Brochure: This brochure describes how the view, create, modify, and print rosters in TIDE.

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

Training Modules

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

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

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

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

Performance Task Overview Module: This module provides an overview of the performance task component and the purpose of the classroom activity as it pertains to the performance task.

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

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

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

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

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

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

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

2.3.2 District Training Workshops

District Test Coordinator (DC) Workshops were held on January 20−22, 2016, at the Institute of Technology and Business Development (ITBD) in New Britain. Training was provided for the administration of the Smarter Balanced assessments for ELA/L and mathematics. During the training, DCs were provided with information to support training of the SCs, TEs, and TAs.

2.4 TEST SECURITY

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

2.4.1 Student-Level Testing Confidentiality

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

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

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

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

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

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

All students, including home-schooled students, must be enrolled or registered at their testing schools in order to take the online, paper-and-pencil, or Braille assessments. Student enrollment information, including demographic data, is generated using a CSDE file and uploaded nightly via a secured file transfer site to the online testing system during the testing period.

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

After a test session, only staff with the administrative roles of DAs, DCs, SCs, or TEs can view their students’ scores. TAs do not have access to student scores.

2.4.2 System Security

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

A hierarchy of control: As described in Section 2.2, DAs, DCs, SCs, TAs, and TEs have defined roles and access to the testing system. When the TIDE window opens, CSDE provides a verified list of DAs to the testing contractor who uploads the information into TIDE. DAs are then responsible for selecting and entering the DC's and SC's information into TIDE, and the SC is responsible for entering TAs' and TEs' information in TIDE. Throughout the year, the DA, DC, and SC 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 SCs, TAs, and TEs receive separate passwords through their personal e-mail addresses assigned by the school.

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

2.4.3 Security of the Testing Environment

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

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

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

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

Room Preparation

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

Seating Arrangements

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

After the Test

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

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

2.4.4 Test Security Violations

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

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

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

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

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

2.5 STUDENT PARTICIPATION

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

2.5.1 Home-Schooled Students

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

2.5.2 Exempt Students

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

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

2.6 ONLINE TESTING FEATURES AND TESTING ACCOMMODATIONS

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

The Smarter Balanced *Usability, Accessibility, and Accommodations 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. The *Guidelines* focus on universal tools, designated supports, and accommodations for the Smarter Balanced assessments of ELA/L and mathematics. At the same time, the *Guidelines* support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments.

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

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

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

For additional information about the availability of designated supports and accommodations, refer to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* for complete information http://www.smarterbalanced.org/wp-content/uploads/2015/09/Usability-Accessibility-Accomodations-Guidelines.pdf.

2.6.1 Online Universal Tools for ALL Students

Universal tools are access features of an assessment or exam that are embedded or non-embedded components of the test administration system. Universal tools are available to all students based on their preference and selection and have been preset in TIDE. In SY 2015–2016 test administration, the following features of universal tools are available for *all* students to access. For specific information on how to access and use these features, refer to the *Test Administrator User Guide* at this URL: http://ct.portal.airast.org.

Embedded Universal Tools

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

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

Pause: The student can pause the assessment and return to the test question that the student was on. 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 indicated that it would be appropriate.

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

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

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

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

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

Cross out response options: by using the strikethrough function.

Mark a question for review: to return to it later. However, for the CAT, if the assessment is paused for more than 20 minutes, students will not be allowed to return to marked test questions.

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

Non-Embedded Universal Tools

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

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

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

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

2.6.2 Designated Supports and Accommodations

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

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

Embedded Designated Supports

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

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

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

Translated test directions for math: Translation of test directions is a language support available before beginning the actual test items. Students can see test directions in another language. As an embedded designated support, translated test directions are automatically a part of the stacked translation designated support.

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

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

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

Non-Embedded Designated Supports

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

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

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

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

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

Read aloud (for mathematics items and ELA/L items, but not for 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.

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

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

Translated test directions: This is a PDF file of directions translated in each of the languages currently supported. A bilingual adult can read the file to the student.

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

Embedded Accommodations

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

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

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

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

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

Non-Embedded Accommodations

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 adapted keyboards, large keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches.

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

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

Print on demand: Paper copies of passages, stimuli, and/or items are printed for students. For those students needing a paper copy of a passage or stimulus, permission for the students to request printing must first be set in TIDE. For those students needing a paper copy of one or more items, the School Test Coordinator must fill out a Verification of Student Need Form and contact CSDE to have the accommodation set for the student.

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

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

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

Table 3 presents a list of universal tools, designated supports, and accommodations that were offered in the 2015–2016 administration. Tables 4–9 provide the number of students who were offered the accommodations and designated supports.

Table 3. SY 2015–2016 Universal Tools, Designated Supports, and Accommodations

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

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

[1] For calculator-allowed items only in grades 6-8 and 11
[2] For ELA/L performance task full-writes
[3] Includes embedded ruler, embedded protractor
[4] Includes bold, italic, underline, indent, cut, paste, spell check, bullets, undo/redo
[5] For ELA/L PT stimuli, ELA/L PT and CAT items (not ELA/L CAT reading passages), and math stimuli and items: Must be set in TIDE before test begins.
[6] For math items
[7] For math items
[8] For math test
[9] For ELA/L listening items and math items
[10] For ELA/L listening items
[11] For ELA/L reading passages. Must be set in TIDE by state-level user.
[12] For ELA/L performance task full writes
[13] For ELA/L performance task full writes
[14] For ELA/L performance task full writes
[15] For ELA/L items (not ELA reading passages) and math items
[16] For ELA/L non-writing items and math items
[17] For math items on the paper/pencil test
[18] Includes adapted keyboards, large keyboard, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches
[19] For calculator-allowed items only in grades 6-8 and 11
[20] For math items beginning in grade 4
[21] For ELA reading passages, all grades

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

Accommodations	**Grade**					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	9	4	3	6	7	9
Closed Captioning	18	26	26	20	27	24
Streamlined Mode	86	71	68	52	44	25
Text-to-Speech: Passage and Items	384	411	415	609	724	642
Non-Embedded Accommodations						
Alternate Response Options	11	5	6	4	3	3
Large Print	9	4	3	3	4	2
Read Aloud Stimuli	49	34	18	21	27	15
Scribe Items (Writing)	11	5	4	2	1	1
Speech-to-Text	47	76	68	76	58	57

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	75	75	63	36	30	15
	LEP	13	12	9	1	1	
	IDEA Eligible	17	16	24	14	8	6
Masking	Overall	151	181	182	89	118	126
	LEP	47	37	45	26	37	41
	IDEA Eligible	106	125	124	65	79	87
Text-to-Speech: Items	Overall	4,503	4,633	4,402	2,897	2,435	2,024
	LEP	2,039	1,793	1,719	829	700	623
	IDEA Eligible	1,974	2,476	2,479	1,933	1,606	1,278

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	4	9	4	3	3	1
	LEP	1	5	1			
	IDEA Eligible	1	2	2	2		1
Color Overlay	Overall	5	12	15	5	4	4
	LEP	1	5	1	1		
	IDEA Eligible	4	2	11	4	2	3
Magnification	Overall	15	22	12	11	18	10
	LEP	3	7	2	1	1	2
	IDEA Eligible	5	9	6	6	12	9
Noise Buffers	Overall	27	28	19	9	4	4
	LEP	6	7	2			2
	IDEA Eligible	13	9	10	5	2	3
Read Aloud Items	Overall	127	77	62	34	31	28
	LEP	63	30	27	5	5	3
	IDEA Eligible	71	55	41	29	26	28
Scribe Items (Non-Writing)	Overall	7	5	4	1	2	2
	LEP			2			
	IDEA Eligible	7	4	3	1	2	2
Separate Setting	Overall	2,328	2,598	2,524	2,023	1,885	2,054
	LEP	498	514	474	279	254	267
	IDEA Eligible	1,579	1,847	1,835	1,584	1,458	1,591
Translated Test Directions	Overall	180	221	232	207	208	225
	LEP	166	209	218	188	190	214
	IDEA Eligible	21	20	29	24	17	17

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

Accommodations	**Grade**					
	3	**4**	**5**	**6**	**7**	**8**
Embedded Accommodations						
American Sign Language	9	4	3	6	8	9
Streamlined Mode	87	71	69	53	43	24
Non-Embedded Accommodations						
Abacus	1	1		2	9	4
Alternate Response Options	11	4	8	4	4	4
Calculator	18	24	34	153	215	254
Multiplication Table		1,630	1,945	1,777	1,570	1,252
Speech-to-Text	42	72	67	68	58	60

Table 8. Mathematics Total Students with Allowed Embedded Designated Supports

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	74	74	64	36	18	22
	LEP	13	12	9	1		2
	IDEA Eligible	16	15	24	13	8	9
Masking	Overall	150	187	182	87	119	126
	LEP	47	38	46	26	37	41
	IDEA Eligible	106	126	122	63	80	88
Translation (Glossary): Spanish	Overall	448	372	384	309	283	205
	LEP	440	368	378	295	275	201
	IDEA Eligible	33	39	51	36	26	24
Translation (Glossary): Other Languages	Overall	75	63	47	60	41	44
	LEP	75	62	47	57	41	44
	IDEA Eligible	2	1	2	2	1	
Text-to-Speech: Stimuli and Items	Overall	5,365	5,434	5,095	3,564	3,156	2,648
	LEP	2,178	1,869	1,707	823	696	593
	IDEA Eligible	2,539	3,083	3,056	2,623	2,384	1,963

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

Designated Supports	**Subgroup**	**Grade**					
		3	**4**	**5**	**6**	**7**	**8**
Color Contrast	Overall	4	9	4	3	4	
	LEP	1	5	1			
	IDEA Eligible	1	2	2	2	2	
Color Overlay	Overall	5	13	14	5	3	3
	LEP	1	5	1	1		
	IDEA Eligible	4	3	10	4	1	3
Translation (Glossary): Spanish	Overall	154	157	184	174	153	195
	LEP	146	148	175	158	145	186
	IDEA Eligible	17	20	26	17	7	13
Translation (Glossary): Other Languages	Overall	17	16	15	15	9	14
	LEP	16	15	14	14	8	13
	IDEA Eligible			1		1	1
Magnification	Overall	11	17	7	11	20	10
	LEP	2	2	1	1	1	2
	IDEA Eligible	5	10	3	6	14	9
Noise Buffers	Overall	24	23	16	9	3	3
	LEP	5	2	1			1
	IDEA Eligible	13	9	8	5	2	2
Read Aloud Items & Stimuli	Overall	180	97	90	57	35	41
	LEP	77	35	36	10	7	4
	IDEA Eligible	112	74	64	49	28	37
Read Aloud Items & Stimuli (Spanish)	Overall	51	36	48	20	3	11
	LEP	49	35	45	19	3	11
	IDEA Eligible	5	2	9	3		2

Designated Supports	Subgroup	Grade					
		3	4	5	6	7	8
Scribe Items	Overall	5	4	3	2	2	2
	LEP			1			
	IDEA Eligible	5	3	2	2	2	2
Separate Setting	Overall	2,259	2,524	2,471	2,014	1,883	2,036
	LEP	494	469	449	282	248	260
	IDEA Eligible	1,533	1,820	1,813	1,582	1,464	1,584
Translated Test Directions	Overall	179	198	209	139	149	166
	LEP	166	188	195	126	135	156
	IDEA Eligible	20	18	29	23	13	17

2.7 DATA FORENSICS PROGRAM

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

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

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

2.7.1 Changes in Student Performance

Cross-year comparisons are performed starting with the second year of the Smarter Balanced assessment using a regression model. The 2015-16 scores were regressed on the 2014-15 scores controlling for the number of days between the two test end days. The number of days between test end days was used to control the instruction time between the two test scores.

A large score gain or loss between grades is detected by examining the residuals for outliers. The residuals are computed as observed value minus predicted value. Studentized t residuals were computed to detect unusual residuals. An unusual increase or decrease in student scores is flagged when studentized t residuals are greater than |3|.

For aggregate units (testing session, test administrator, and school), unusual changes in an aggregate performance between test administrations are based on the average studentized t residuals for the students in the aggregate unit. For each aggregate unit, a critical t value is computed and flagged when t was greater than |3|,

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

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

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

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

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

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

2.7.2 Item Response Time

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

The expectation is that the item response time will be shorter than the average time if students have a prior knowledge of items. An example of unusual item response time is a test record for an individual who scores very well on the test even though the average time spent for each item was far less than that required of students statewide. If students already know the answers to the questions, the response time will be much shorter than the response time for those items where the student has no prior knowledge of the item content. Conversely, if a TA helps students by “coaching” them to change their responses during the test, the testing time could be longer than expected.

The average and the standard deviation of test-taking time are computed across all students for each opportunity. Students and aggregate units are flagged if the test-taking time is greater than |3| standard deviations of the state average. The state average and standard deviation is computed based on all students when the analysis was performed. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit.

2.7.3 Inconsistent Item Response Pattern (Person Fit)

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

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

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

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

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

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

where *s* = standard deviation of l_z values in an aggregate unit and *n* = number of students in an aggregate unit. The QA report includes a list of the flagged aggregate units with the number of flagged students in the aggregate unit (e.g., test session, test administrator, and school).

3. SUMMARY OF 2015–2016 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT POPULATION

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

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

Group	G3	G4	G5	G6	G7	G8
All Students	38,942	38,450	39,010	39,071	40,085	39,351
Female	19,139	18,805	19,273	18,963	19,410	19,157
Male	19,803	19,645	19,737	20,108	20,675	20,194
American Indian/Alaska Native	90	102	112	95	113	94
Asian	2,151	1,996	2,003	1,990	1,994	1,925
African American	4,874	4,955	4,840	4,881	4,917	5,068
Hispanic/Latino	9,854	9,383	9,201	8,794	8,836	8,546
White	20,601	20,825	21,826	22,299	23,119	22,770
Multiple Ethnicities	1,325	1,160	985	980	1,063	922
LEP	3,554	2,962	2,694	2,112	2,074	1,791
IDEA	4,332	4,934	5,070	5,193	5,232	5,171

Table 11. Number of Students in Summative Mathematics Assessment

Group	G3	G4	G5	G6	G7	G8
All Students	38,870	38,387	38,941	38,965	39,961	39,181
Female	19,109	18,773	19,242	18,921	19,352	19,069
Male	19,761	19,614	19,699	20,044	20,609	20,112
American Indian/Alaska Native	90	102	112	95	113	94
Asian	2,147	1,992	1,999	1,988	1,988	1,922
African American	4,860	4,938	4,830	4,860	4,895	5,043
Hispanic/Latino	9,833	9,372	9,173	8,769	8,798	8,504
White	20,569	20,794	21,798	22,243	23,063	22,679
Multiple Ethnicities	1,325	1,160	986	978	1,061	913
LEP	3,546	2,954	2,688	2,107	2,057	1,779
IDEA	4,324	4,916	5,055	5,158	5,189	5,131

3.2 SUMMARY OF STUDENT PERFORMANCE

Tables 12–15 present a summary of overall student performance in the 2015–2016 summative test for all students and by subgroups, including the average and the standard deviation of overall scale scores, the percentage of students in each achievement level, and the percentage of proficient students. Figures 1–2 compare the percentage of proficient students in 2014–2015 and 2015–2016 for all students and subgroups (cohort comparisons). The average and the standard deviation of scale scores, and the percentage of proficient students in both years are provided in Appendix B.

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	38,942	2438	89	23	23	23	31	54
Female	19,139	2447	88	19	23	24	34	58
Male	19,803	2430	90	26	24	23	27	50
American Indian or Alaska Native	90	2422	78	24	28	30	18	48
Asian	2,151	2480	84	11	15	25	50	74
African American	4,874	2392	81	41	28	19	13	31
Hispanic or Latino	9,854	2395	82	39	29	19	13	33
Native Hawaiian/Pacific Islander	47	2420	92	23	38	11	28	38
White	20,601	2465	82	12	20	26	41	67
Multiple Ethnicities	1,325	2450	87	18	25	23	34	57
LEP	3,554	2361	70	54	29	13	4	16
IDEA Eligible	4,332	2357	78	59	24	11	6	17
Grade 4								
All Students	38,450	2480	96	27	18	23	32	56
Female	18,805	2490	94	23	18	24	36	59
Male	19,645	2471	97	30	18	23	29	52
American Indian or Alaska Native	102	2446	98	41	17	18	25	42
Asian	1,996	2526	91	13	12	23	51	74
African American	4,955	2427	87	48	21	18	13	31
Hispanic or Latino	9,383	2430	89	46	21	20	13	33
Native Hawaiian/Pacific Islander	29	2486	89	24	21	28	28	55
White	20,825	2511	85	14	16	26	43	70
Multiple Ethnicities	1,160	2493	95	23	18	22	37	59
LEP	2,962	2384	78	68	18	11	3	14
IDEA Eligible	4,934	2390	84	65	18	12	6	17
Grade 5								
All Students	39,010	2517	97	23	18	30	28	59
Female	19,273	2531	94	18	17	32	33	64
Male	19,737	2504	98	27	19	29	24	53
American Indian or Alaska Native	112	2501	95	27	19	35	20	54
Asian	2,003	2563	89	11	12	30	47	77
African American	4,840	2461	90	43	23	23	10	33
Hispanic or Latino	9,201	2467	92	41	22	26	11	37
Native Hawaiian/Pacific Islander	43	2525	109	23	14	30	33	63
White	21,826	2547	86	12	16	34	38	72
Multiple Ethnicities	985	2528	96	20	18	30	32	62
LEP	2,694	2411	75	65	22	12	1	13
IDEA Eligible	5,070	2420	84	62	21	13	4	17

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

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	39,071	2536	98	22	23	33	22	55
Female	18,963	2548	95	18	22	35	25	60
Male	20,108	2525	100	26	24	31	19	50
American Indian or Alaska Native	95	2527	94	22	31	29	18	47
Asian	1,990	2580	90	10	16	35	38	73
African American	4,881	2482	91	40	29	25	7	31
Hispanic or Latino	8,794	2481	94	40	28	24	7	31
Native Hawaiian/Pacific Islander	32	2541	105	25	25	19	31	50
White	22,299	2565	87	12	20	38	30	68
Multiple Ethnicities	980	2542	95	19	25	34	22	56
LEP	2,112	2411	75	73	21	6	0	6
IDEA Eligible	5,193	2438	87	61	24	12	3	15
Grade 7								
All Students	40,085	2559	100	23	22	35	20	55
Female	19,410	2573	96	18	21	37	24	61
Male	20,675	2546	101	27	23	33	17	50
American Indian or Alaska Native	113	2537	95	26	31	30	13	43
Asian	1,994	2613	91	9	15	37	39	77
African American	4,917	2502	89	43	28	24	6	29
Hispanic or Latino	8,836	2505	95	41	27	25	7	32
Native Hawaiian/Pacific Islander	43	2555	117	28	16	30	26	56
White	23,119	2587	89	12	20	41	26	67
Multiple Ethnicities	1,063	2566	101	20	21	37	22	59
LEP	2,074	2430	71	76	19	4	0	5
IDEA Eligible	5,232	2460	86	62	23	13	2	15
Grade 8								
All Students	39,351	2574	100	21	24	37	18	55
Female	19,157	2589	96	16	22	41	21	62
Male	20,194	2559	102	26	25	34	15	49
American Indian or Alaska Native	94	2556	93	21	35	32	12	44
Asian	1,925	2626	93	8	16	41	35	76
African American	5,068	2520	92	37	31	26	6	32
Hispanic or Latino	8,546	2519	95	38	29	27	6	33
Native Hawaiian/Pacific Islander	26	2585	106	23	19	35	23	58
White	22,770	2601	90	12	21	43	24	67
Multiple Ethnicities	922	2582	100	19	22	37	22	59
LEP	1,791	2436	68	79	18	3	0	4
IDEA Eligible	5,171	2473	85	60	26	13	2	15

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

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 3								
All Students	38,870	2,438	81	24	24	30	23	53
Female	19,109	2,438	78	23	25	31	21	52
Male	19,761	2,439	84	24	23	29	24	53
American Indian or Alaska Native	90	2,431	77	24	24	33	18	51
Asian	2,147	2,491	76	9	13	31	47	78
African American	4,860	2,391	75	44	29	20	7	27
Hispanic or Latino	9,833	2,398	75	40	29	22	9	31
Native Hawaiian/Pacific Islander	46	2,421	77	24	30	30	15	46
White	20,569	2,463	72	12	21	36	31	67
Multiple Ethnicities	1,325	2,446	77	20	23	32	24	56
LEP	3,546	2,377	70	52	28	16	4	20
IDEA Eligible	4,324	2,360	82	61	21	13	5	18
Grade 4								
All Students	38,387	2,478	82	21	31	28	20	48
Female	18,773	2,476	78	21	32	28	18	47
Male	19,614	2,480	86	22	29	27	22	49
American Indian or Alaska Native	102	2,450	87	32	31	24	13	36
Asian	1,992	2,533	82	8	19	28	45	73
African American	4,938	2,427	72	43	36	15	6	21
Hispanic or Latino	9,372	2,434	74	39	37	18	6	24
Native Hawaiian/Pacific Islander	29	2,488	77	10	34	34	21	55
White	20,794	2,504	72	10	28	35	28	62
Multiple Ethnicities	1,160	2,488	81	18	31	27	24	51
LEP	2,954	2,405	69	55	33	10	3	12
IDEA Eligible	4,916	2,401	75	59	28	10	4	13
Grade 5								
All Students	38,941	2,501	89	31	28	20	21	41
Female	19,242	2,500	86	31	30	20	20	40
Male	19,699	2,502	93	31	27	20	22	42
American Indian or Alaska Native	112	2,488	84	36	32	15	17	32
Asian	1,999	2,562	87	12	20	22	46	68
African American	4,830	2,440	77	59	27	9	4	14
Hispanic or Latino	9,173	2,452	80	53	29	12	7	18
Native Hawaiian/Pacific Islander	43	2,511	103	35	28	7	30	37
White	21,798	2,530	79	17	29	25	28	54
Multiple Ethnicities	986	2,512	91	27	30	18	25	43
LEP	2,688	2,415	69	73	21	5	1	6
IDEA Eligible	5,055	2,416	78	72	19	6	3	9

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

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

Group	Number Tested	Scale Score Mean	Scale Score SD	% Level 1	% Level 2	% Level 3	% Level 4	% Proficient
Grade 6								
All Students	38,965	2521	104	30	30	21	20	41
Female	18,921	2523	99	28	31	22	19	41
Male	20,044	2519	108	31	29	20	20	41
American Indian or Alaska Native	95	2499	94	36	34	19	12	31
Asian	1,988	2588	99	13	20	22	44	66
African American	4,860	2452	95	57	29	10	4	14
Hispanic or Latino	8,769	2461	97	52	31	12	5	17
Native Hawaiian/Pacific Islander	32	2530	117	38	22	13	28	41
White	22,243	2553	89	16	30	27	26	53
Multiple Ethnicities	978	2525	101	28	32	20	20	40
LEP	2,107	2402	86	80	16	3	1	4
IDEA Eligible	5,158	2412	96	73	19	5	2	7
Grade 7								
All Students	39,961	2538	108	29	29	23	19	42
Female	19,352	2540	102	27	31	24	18	42
Male	20,609	2536	112	31	28	22	20	42
American Indian or Alaska Native	113	2509	89	42	29	21	8	29
Asian	1,988	2617	103	11	18	24	46	71
African American	4,895	2467	95	56	29	11	4	14
Hispanic or Latino	8,798	2477	101	51	31	14	5	19
Native Hawaiian/Pacific Islander	43	2546	119	28	28	19	26	44
White	23,063	2569	93	17	29	29	25	54
Multiple Ethnicities	1,061	2544	108	28	29	24	20	44
LEP	2,057	2415	89	79	16	3	2	5
IDEA Eligible	5,189	2427	99	73	18	6	3	9
Grade 8								
All Students	39,181	2551	116	35	25	19	21	40
Female	19,069	2557	110	32	26	21	21	42
Male	20,112	2546	121	37	24	18	21	39
American Indian or Alaska Native	94	2509	107	49	31	10	11	20
Asian	1,922	2635	113	14	17	21	48	69
African American	5,043	2479	100	62	24	10	5	15
Hispanic or Latino	8,504	2485	103	59	24	11	6	17
Native Hawaiian/Pacific Islander	26	2551	1274	35	35	4	27	31
White	22,679	2585	104	22	26	25	28	52
Multiple Ethnicities	913	2559	115	33	24	21	22	43
LEP	1,779	2419	85	86	11	2	1	3
IDEA Eligible	5,131	2438	95	78	15	5	2	7

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

Figure 1. ELA/L %Proficient in 2014–2015 and 2015–2016

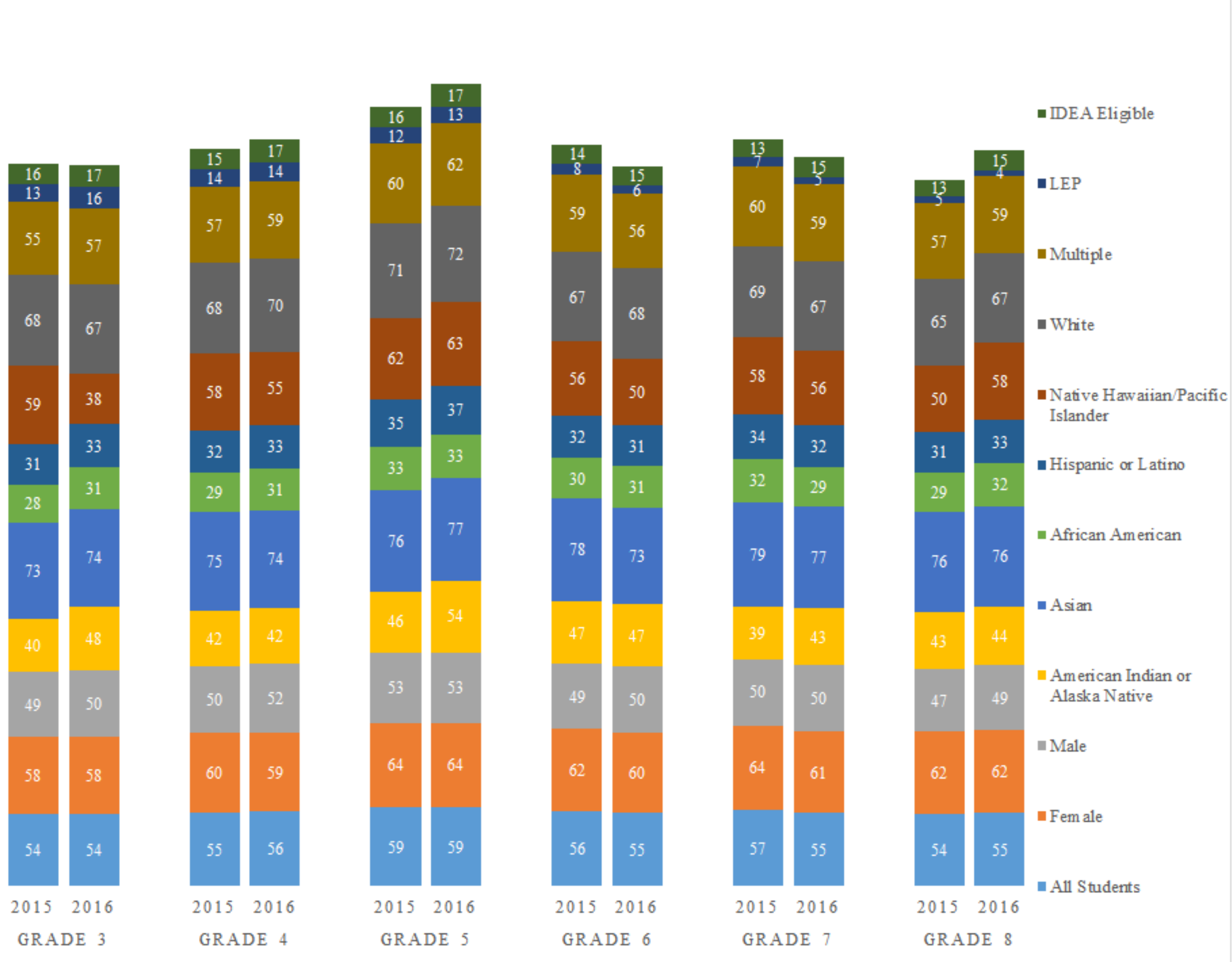
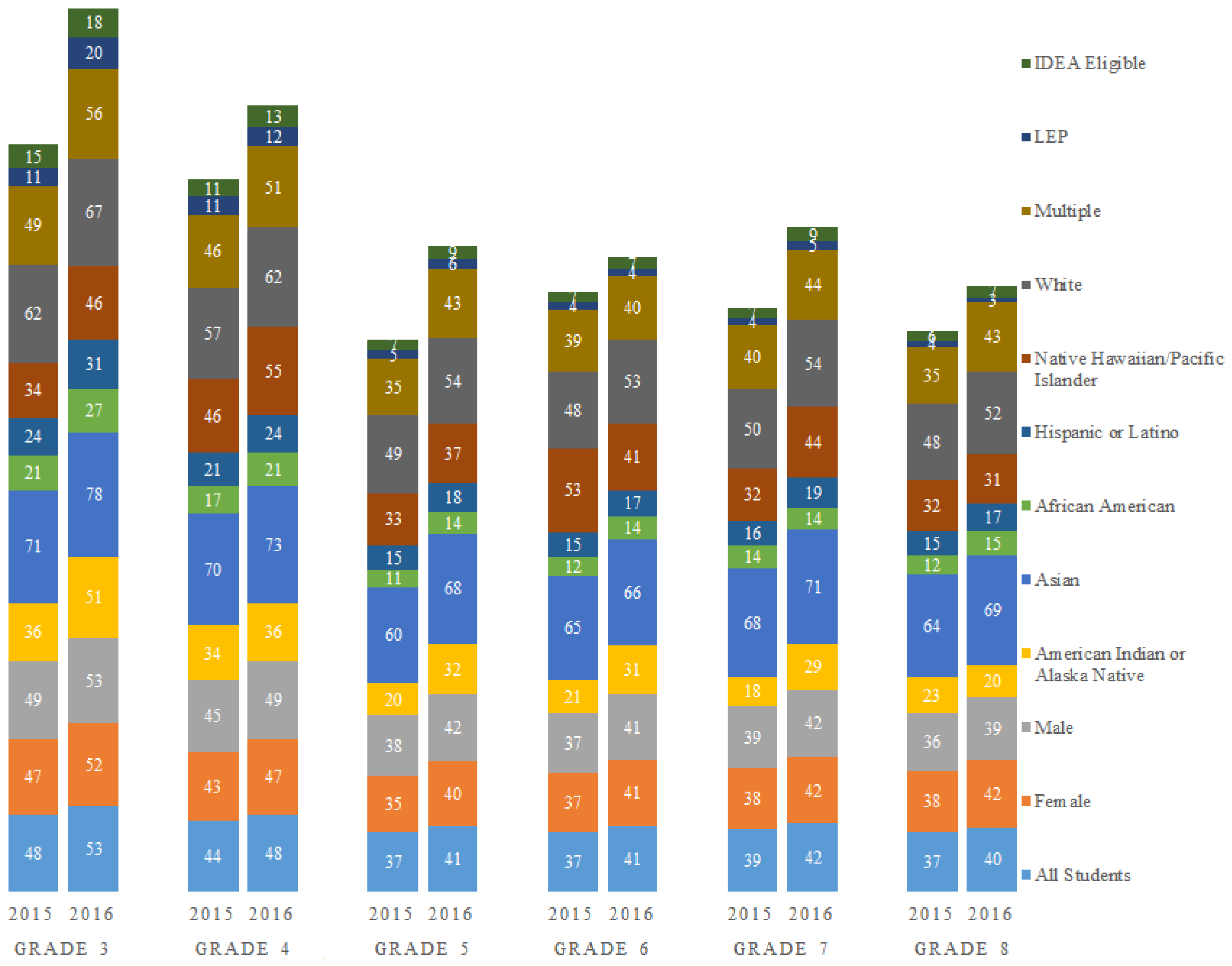


Figure 2. Mathematics %Proficient in 2014–2015 and 2015–2016



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

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

Grade	Performance Category	Claim 1 Reading	Claim 2 & 4: Writing & Research	Claim 3 Listening
3	Below	25	28	14
	At/Near	46	43	62
	Above	29	30	24
4	Below	26	26	14
	At/Near	43	45	64
	Above	31	30	22
5	Below	25	21	16
	At/Near	43	44	60
	Above	32	36	25
6	Below	29	21	13
	At/Near	48	48	66
	Above	23	32	21
7	Below	25	22	14
	At/Near	47	48	65
	Above	28	30	20
8	Below	25	25	14
	At/Near	44	46	67
	Above	31	29	20

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

Grade	Performance Category	Claim 1	Claim 2 & 4	Claim 3
3	Below	30	26	17
	At/Near	35	45	51
	Above	35	29	32
4	Below	35	29	27
	At/Near	34	46	44
	Above	31	25	28
5	Below	42	36	33
	At/Near	32	41	46
	Above	26	23	21
6	Below	41	34	26
	At/Near	35	45	52
	Above	24	21	22
7	Below	39	31	24
	At/Near	34	46	52
	Above	26	23	24
8	Below	42	22	26
	At/Near	33	53	53
	Above	25	24	21

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

3.3 TEST TAKING TIME

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

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

Tables 18 and 19 present an average testing time and the percentage of students for testing time by hourly intervals for the overall test, the CAT component, and the PT component.

Table 18. ELA/L Test Taking Time

Grade	Average Testing Time (hh:mm)	% Students in Each Testing Time Category				
		Less than an hour	1–2 hours	2–3 hours	3–4 hours	More than 4 hours
Overall Test (CAT Component)						
3	1:31	18	65	15	2	0
4	1:36	14	65	18	2	1
5	1:30	16	69	14	1	0
6	1:32	15	67	15	2	0
7	1:20	24	67	8	1	0
8	1:18	27	65	7	1	0

Table 19. Mathematics Test Taking Time

Grade	Average Testing Time (hh:mm)	% Students in Each Testing Time Category				
		Less than an hour	1–2 hours	2–3 hours	3–4 hours	More than 4 hours
Overall Test						
3	1:52	9	56	26	7	2
4	1:51	11	54	26	7	2
5	2:09	6	45	33	11	5
6	1:59	7	53	30	8	3
7	1:38	14	63	19	3	1
8	1:42	14	58	22	4	1
CAT Component						
3	1:12	42	50	7	1	0
4	1:14	40	50	8	1	0
5	1:15	37	54	8	1	0
6	1:13	38	55	6	1	0
7	1:12	38	56	5	1	0
8	1:11	40	53	6	1	0
PT Component						
3	0:40	84	15	1	0	0
4	0:37	88	12	0	0	0
5	0:54	68	28	3	1	0
6	0:46	78	20	2	0	0
7	0:26	96	4	0	0	0
8	0:31	92	7	0	0	0

3.4 STUDENT ABILITY–ITEM DIFFICULTY DISTRIBUTION FOR THE 2015–2016 OPERATIONAL ITEM POOL

Figures 3 and 4 display the empirical distribution of the Connecticut student scale scores in the 2015–2016 administration and the distribution of the summative item difficulty parameters in the operational pool. The student ability distribution is shifted to the left in all grades and subjects, more pronounced in the mathematics upper grades, indicating that the pool includes more difficult items than the ability of students in the tested population. The pool includes difficult items to measure high performing students accurately but needs additional easy items to better measure low performing students. The Smarter Balanced Assessment Consortium plans to add additional easy items to the pool, and augment the pool in proportion to the test blueprint constraints (e.g., content, Depth-of-Knowledge (DoK), item type, and item difficulties).

Figure 3. Student Ability–Item Difficulty Distribution for ELA/L

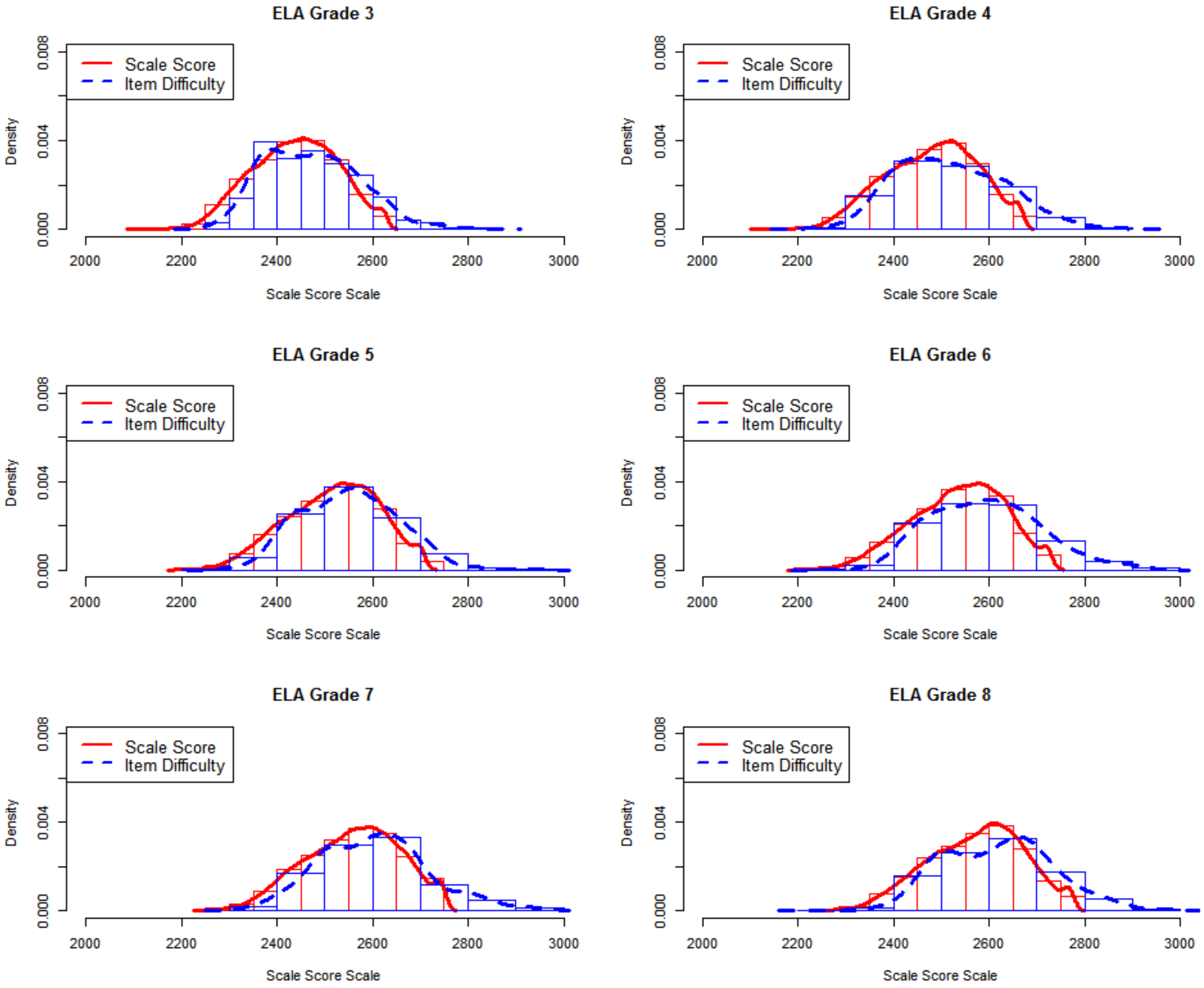
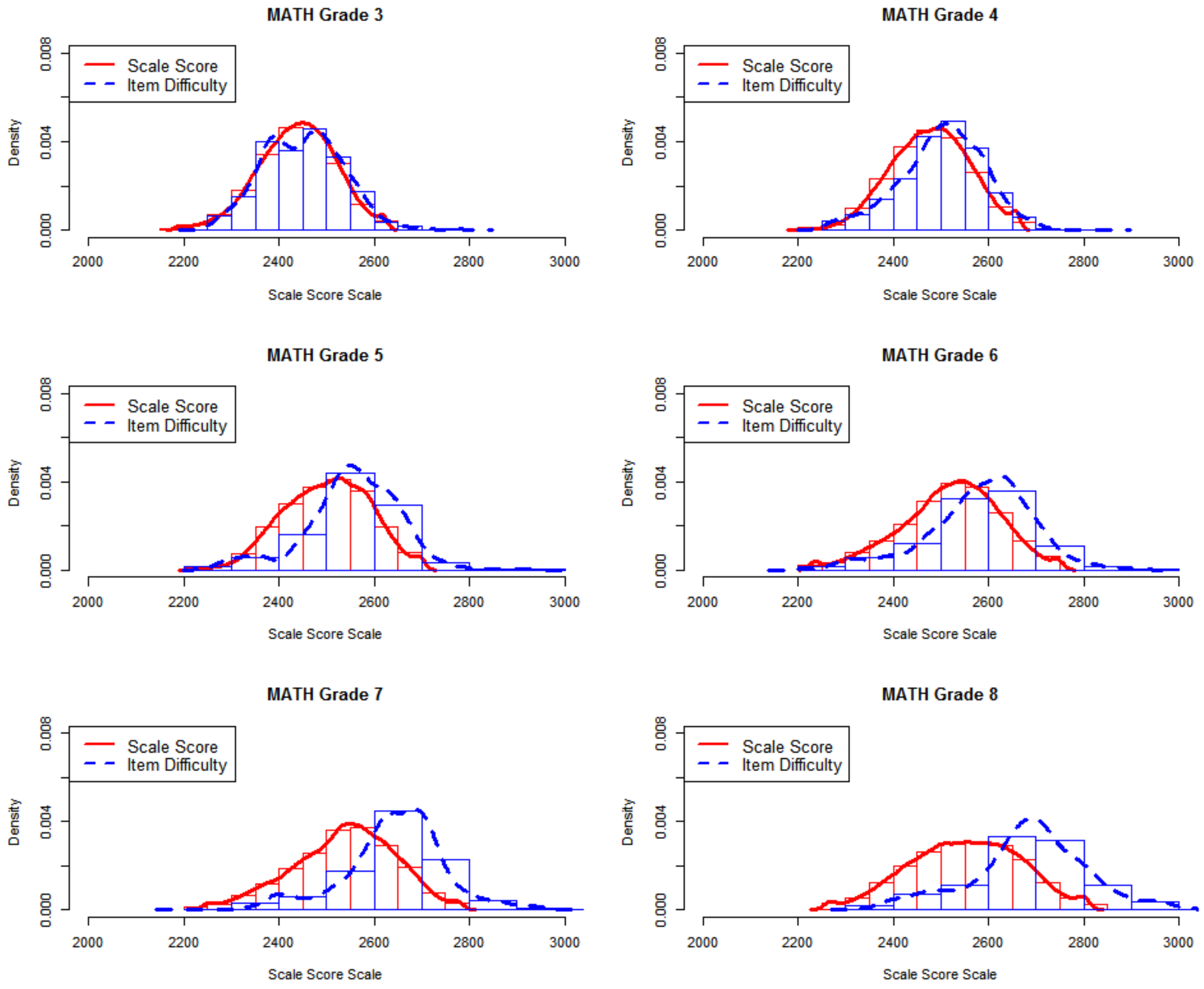


Figure 4. Student Ability–Item Difficulty Distribution for Mathematics



4. VALIDITY

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

Validity evidence provided in this chapter is as follows:

- Test content
- Internal structure
- Relations to other variable

Evidence on test content validity is provided with the blueprint match rates for the delivered tests. Evidence on internal structure is examined in the results of inter-correlations among reporting category scores. For relations to other variables, the relationships between ELA/L and mathematics scores between years were examined using 2014–2015 and 2015–2016 Connecticut summative test data.

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

4.1 EVIDENCE ON TEST CONTENT

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

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match-to-ability. The Smarter Balanced blueprints (Smarter Balanced Assessment Consortium, 2015) specify a range of items to be administered in each claim, content domain/standards, and targets. Moreover, blueprints constrain the DoK and item and passage types. In blueprints, all content blueprint elements are configured to obtain a strictly enforced range of items administered. The algorithm also seeks to satisfy target-level constraints, but these ranges are not strictly enforced. In ELA/L, the blueprints also specify the number of passages in reading (claim 1) and listening (claim 3) claims.

Tables 20 and 21 present the percentages of tests aligned with the test blueprint constraints for ELA/L CAT. Table 20 provides the blueprint match rates for item and passage requirements for each claim. For DoK and item type constraints, the Smarter Balanced blueprint specifies the minimum number of items, not the maximum. Table 21 presents the percentages of tests that satisfied the DoK and item type constraints for each claim. All tests met the requirements, except for the claim 2 DoK2 requirement in grades 3, 7, and 8, which each administered one DoK2 item fewer than required in claim 2.

Tables 22–25 provide the percentages of tests aligned with the test blueprint constraints for mathematics CAT. Tables 22–24 provide the blueprint match rates for claims and content domains within each claim.

The fidelity to the DoK and target constraints is shown in Table 25. In mathematics, all tests met the blueprint requirements for claims, but there were a few exceptions in content domains. A few tests administered one item fewer or one item more than the minimum or maximum item requirement for content domains. For the DoK and target constraints, all tests satisfied the requirements, except for grade 5 and 6. In grade 5, two percent of all delivered tests administered one DoK3 or DoK4 item fewer than required in claim 2 and 4 combined. In grade 6, one percent of all delivered tests administered one Target A or D item fewer than required in target A and D combined within claim 3.

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

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

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

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

DoK and Item Type Constraints	**Minimum Required Items**	**%Blueprint Match**					
		Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
Claim 1 DoK2	7	100%	100%	100%	100%	100%	100%
Claim 1 DoK3 or higher	2	100%	100%	100%	100%	100%	100%
Claim 2 DoK2	4	97%	100%	100%	100%	90%	99%
Claim 2 DoK3 or higher	1	100%	100%	100%	100%	100%	100%
Claim 2 Brief Write	1	100%	100%	100%	100%	100%	100%
Claim 3 DoK2 or higher	3	100%	100%	100%	100%	100%	100%

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

Claim	**Content Domain**	**Grade 3**			**Grade 4**			**Grade 5**		
		Min	**Max**	**%BP Match**	**Min**	**Max**	**%BP Match**	**Min**	**Max**	**%BP Match**
1	ALL	20	20	100%	20	20	100%	20	20	100%
	P	15	15	99%	15	15	100%	15	15	100%
	S	5	5	99%	5	5	100%	5	5	100%
2	ALL	3	3	100%	3	3	100%	3	3	100%
	G	0	2	100%	0	2	100%	0	2	100%
	MD	0	2	100%	0	2	100%	0	2	100%
	NBT	0	2	100%	0	2	100%	0	2	100%
	NF	0	2	100%	1	3	100%	1	3	100%
	OA	0	2	100%	0	2	100%	0	2	100%
3	ALL	8	8	100%	8	8	100%	8	8	100%
	G							0	3	100%
	MD	0	4	100%				0	4	100%
	NBT				0	4	100%	0	4	100%
	NF	2	6	100%	2	6	100%	2	6	100%
	OA	0	4	100%	0	4	100%			
	OTHER				0	2	100%			
4	ALL	3	3	100%	3	3	100%	3	3	100%
	G	0	1	100%	0	1	100%	0	1	100%
	MD	1	2	100%	0	2	100%	1	2	100%
	NBT	0	1	100%	0	1	100%	0	1	100%
	NF	0	1	100%	0	2	100%	1	2	100%
	OA	1	2	100%	0	2	100%	0	1	100%

Legend:

ALL	Total item requirement in a claim.	N	Number and quantity
1-P	Primary target set	NBT	Number and operations in Base ten
1-S	Secondary target set	NF	Number and operations—fractions
A	Algebra	OA	Operations and algebraic thinking
G	Geometry	OTHER	Other content domains
MD	Measurement and data		

Table 23. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain: Grades 6–7 Mathematics

Claim	**Content Domain**	**Segment**	**Grade 6**			**Grade 7**		
			Min	**Max**	**%BP Match**	**Min**	**Max**	**%BP Match**
1	ALL	Calc	6	6	100%	10	10	100%
	P	Calc	3	3	100%	6	6	100%
	S	Calc	3	3	100%	4	4	100%
	ALL	NoCalc	13	13	100%	10	10	100%
	P	NoCalc	11	11	100%	9	9	100%
	S	NoCalc	2	2	100%	1	1	100%
2	ALL	Calc	3	3	100%	3	3	100%
	EE	Calc	0	2	100%	0	2	100%
	G	Calc	0	2	100%	0	2	100%
	NS	Calc	0	2	100%	0	2	100%
	RP	Calc	0	2	100%	0	2	100%
	SP	Calc	0	2	100%	0	2	100%
	OTHER	Calc	0	2	100%	0	2	100%
3	ALL	Calc	7	7	100%	8	8	100%
	EE	Calc	0	5	100%	1	5	100%
	NS	Calc	2	6	100%	1	5	100%
	RP	Calc	0	5	100%	1	5	100%
	ALL	NoCalc	1	1	100%			
	EE	NoCalc	0	1	100%			
	NS	NoCalc	0	1	100%			
	RP	NoCalc	0	1	100%			
4	ALL	Calc	3	3	100%	3	3	100%
	EE	Calc	0	1	99%	0	1	100%
	G	Calc	0	1	100%	0	1	100%
	NS	Calc	0	1	99%	0	1	100%
	RP	Calc	0	1	100%	0	1	100%
	SP	Calc	0	1	100%	0	1	100%
	OTHER	Calc	0	1	100%	0	1	100%

Legend:

ALL	Total item requirement in a claim.	NS	The number system
1-P	Primary target set	OTHER	Other content domains
1-S	Secondary target set	RP	Ratios and proportional relationships
EE	Expressions and equations	SP	Statistics and probability
G	Geometry	Calc	Segment with calculator use
		NoCalc	Segment without calculator use

Table 24. Percentage of Delivered Tests Meeting Blueprint Requirements for Each Claim and Content Domain: Grade 8 Mathematics

Grade 8					
Claim	**Content Domain**	**Segment**	**Min**	**Max**	**%BP Match**
1	ALL	Calc	14	14	100%
	P	Calc	11	11	100%
	S	Calc	3	3	100%
	ALL	NoCalc	6	6	100%
	P	NoCalc	4	4	100%
	S	NoCalc	2	2	100%
2	ALL	Calc	3	3	100%
	EE	Calc	0	2	100%
	F	Calc	0	2	100%
	G	Calc	0	2	100%
	NS	Calc	0	2	100%
	SP	Calc	0	2	100%
	OTHER	Calc	0	2	100%
3	ALL	Calc	8	8	100%
	EE	Calc	1	5	99%
	F	Calc	1	5	100%
	G	Calc	1	5	100%
4	ALL	Calc	3	3	100%
	EE	Calc	1	2	100%
	F	Calc	0	1	97%
	G	Calc	0	1	100%
	NS	Calc	0	1	100%
	SP	Calc	0	1	100%
	OTHER	Calc	0	1	100%

Legend:

ALL	Total item requirement in a claim.	N	Number and quantity
1-P	Primary target set	NBT	Number and operations in Base ten
1-S	Secondary target set	NF	Number and operations—fractions
A	Algebra	NS	The number system
EE	Expressions and equations	OA	Operations and algebraic thinking
F	Functions	OTHER	Other content domains
G	Geometry	RP	Ratios and proportional relationships
MD	Measurement and data	SP	Statistics and probability
Calc	Segment with calculator use	NoCalc	Segment without calculator use

Table 25. Mathematics Percentage of Delivered Tests Meeting Blueprint Requirements for Depth-of-Knowledge and Targets

DoK and Target Constraints	Minimum Required Items				%Blueprint Match					
	G3-5	G6	G7	G8	G3	G4	G5	G6	G7	G8
Segment 1										
Claim 1 DoK1	5	2	3	4	100%	100%	100%	100%	100%	100%
Claim 1 DoK2 or higher	7	2	4	5	100%	100%	100%	100%	100%	100%
Claim 2 Target A	2	2	2	2	100%	100%	100%	100%	100%	100%
Claim 2 Target B,C,D	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim 2/4 DoK3 or higher	2	2	2	2	100%	100%	98%	100%	100%	100%
Claim 3 DoK3 or higher	2	1	2	2	100%	100%	100%	100%	100%	100%
Claim 3 Target A,D	3	3	2	2	100%	100%	100%	99%	100%	100%
Claim 3 Target B,E	3	2	3	3	100%	100%	100%	100%	100%	100%
Claim 3 Target C,F	2				100%	100%	100%			
Claim 3 Target C,F,G		2	1	1				100%	100%	100%
Claim 4 Target A,D	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim 4 Target B,E	1	1	1	1	100%	100%	100%	100%	100%	100%
Claim 4 Target C,F	1	1	1	1	100%	100%	100%	100%	100%	100%
Segment 2										
Claim 1 DoK1		3	3	2				100%	100%	100%
Claim 1 DoK2 or higher		5	4	4				100%	100%	100%

Table 26 summarizes the target coverage, the number of unique targets administered in each delivered test by claim. The table includes the number of targets specified in the blueprints and the mean and range of the number of targets administered to students. Since the test blueprint is not required to cover all targets in each test, it is expected that the number of targets covered varies across tests. Although the target coverage varies somewhat across individual tests, all targets are covered at an aggregate level, across all tests combined.

Table 26. 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	10	4	1	3	8-13	3-4	1-1	3-3
4	14	5	1	3	11	4	1	3	8-13	3-5	1-1	3-3
5	14	5	1	3	11	5	1	3	8-14	4-5	1-1	3-3
6	14	5	1	3	10	5	1	3	8-11	4-5	1-1	3-3
7	14	5	1	3	10	5	1	3	8-11	4-5	1-1	3-3
8	14	5	1	3	10	4	1	3	8-11	3-4	1-1	3-3
Mathematics												
3	11	4	6	6	10	2	5	3	9-10	2-2	3-6	3-3
4	12	4	6	6	10	2	6	3	10-10	2-2	3-6	3-3
5	11	4	6	6	9	2	6	3	8-9	2-2	3-6	3-3
6	10	4	7	6	10	2	4	3	8-10	2-2	3-6	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	3-3

An adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability and meeting the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test difficulty). However, scores from the test should be comparable, and each test form should measure the same content, albeit with a different set of test items, ensuring the comparability of assessments in content and scores. The blueprint match and target coverage results demonstrate that test forms conform to the same content as specified, thus providing evidence of content comparability. In other words, while each form is unique with respect to its items, all forms align with the same curricular expectations set forth in the test blueprints.

4.2 Evidence on Internal Structure

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

The correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal), are presented in Tables 27 and 28. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability. The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$, where $r_{x'y'}$ is the correlation between x and y corrected for attenuation, r_{xy} is the observed correlation between x and y, r_{xx} is the reliability coefficient for x, and r_{yy} is the reliability coefficient for y. When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Table 27. Correlations among Reporting Categories for ELA/L

Grade	Reporting Categories	Observed and Disattenuated Correlation		
		Claim 1	Claims 2 & 4	Claim 3
3	Claim 1: Reading		0.96	0.98
	Claim 2 & 4: Writing & Research	0.75		0.98
	Claim 3: Listening	0.63	0.66	
4	Claim 1: Reading		0.97	0.99
	Claim 2 & 4: Writing & Research	0.77		0.99
	Claim 3: Listening	0.64	0.65	
5	Claim 1: Reading		0.96	0.98
	Claim 2 & 4: Writing & Research	0.75		0.97
	Claim 3: Listening	0.66	0.66	
6	Claim 1: Reading		0.97	1.00
	Claim 2 & 4: Writing & Research	0.73		1.00
	Claim 3: Listening	0.60	0.64	
7	Claim 1: Reading		0.99	1.00
	Claim 2 & 4: Writing & Research	0.75		1.00
	Claim 3: Listening	0.63	0.64	
8	Claim 1: Reading		0.98	1.00
	Claim 2 & 4: Writing & Research	0.77		1.00
	Claim 3: Listening	0.65	0.65	

Table 28. Correlations among Reporting Categories for Mathematics

Grade	Reporting Categories	Observed and Disattenuated Correlation		
		Claim 1	**Claims 2 & 4**	**Claim 3**
3	Claim 1		1.00	1.00
	Claim 2 & 4	0.79		1.00
	Claim 3	0.78	0.74	
4	Claim 1		1.00	0.99
	Claim 2 & 4	0.79		1.00
	Claim 3	0.81	0.75	
5	Claim 1		1.00	0.99
	Claim 2 & 4	0.78		1.00
	Claim 3	0.77	0.73	
6	Claim 1		1.00	1.00
	Claim 2 & 4	0.82		1.00
	Claim 3	0.76	0.73	
7	Claim 1		1.00	1.00
	Claim 2 & 4	0.80		1.00
	Claim 3	0.81	0.74	
8	Claim 1		1.00	1.00
	Claim 2 & 4	0.73		1.00
	Claim 3	0.77	0.67	

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

4.3 EVIDENCE ON RELATIONS TO OTHER VARIABLES

Validity evidence based on relations to other variables can address a variety of questions. At its core, this type of validity addresses the relationship between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity (Campbell & Fiske, 1959). Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute—scores from two tests measuring the same attribute should be correlated. Conversely, evidence for discriminant validity is obtained when test scores are not correlated with measures of construct irrelevant attributes.

The convergent and discriminant validity was examined based on the relationships between ELA/L and mathematics scores in 2014–2015 and 2015–2016. It was expected that the correlation between two tests measuring the same content (e.g., correlations between ELA/L scores) would be higher than the correlation between tests measuring different contents (e.g., correlation between ELA/L and mathematics scores).

In Table 29, the reliability coefficients are in boldface on diagonal, the correlations between students' scores for the same subject in two years are underlined (convergent validity), and the correlations between ELA/L and mathematics scores within and between years are in a rectangle (discriminant validity). The correlations between two grades for the same subject and between subjects for are computed for grades 4 through 8 only since grades 3 does not have lower grade score to correlate with.

As expected, the coefficients were in the order of reliability coefficients (numbers in boldface), correlations between same subject scores in two years (numbers underlined), and correlations between different subject scores (numbers in rectangles).

The correlations for the same subject scores in two different grades were higher than the correlations between two subject scores within and between grades. The correlation coefficients for the same subject scores ranged from 0.83 to 0.84 for ELA/L and from 0.86 to 0.88 for mathematics. The correlation between ELA/L and mathematics scores within and between grades ranged from 0.75 to 0.82 in grades 3-8. The observed pattern of correlations within and between subjects conforms to the criteria expected for convergent and discriminant validity (Campbell & Fiske, 1959).

Table 29. Relationships between ELA/L and Mathematics Scores

Grade	Year/Subject	N	2015 ELA/L	2016 ELA/L	2015 Math	2016 Math
3	2015 ELA/L	37,885	**0.92**			
3	2016 ELA/L	38,860	n/a	**0.91**		
3	2015 Math	37,885	0.82	n/a	**0.94**	
3	2016 Math	38,860	n/a	0.80	n/a	**0.94**
4	2015 ELA/L	36,339	**0.92**			
4	2016 ELA/L	36,339	0.83	**0.90**		
4	2015 Math	36,546	0.81	0.77	**0.94**	
4	2016 Math	36,546	0.77	0.81	0.87	**0.94**
5	2015 ELA/L	36,909	**0.92**			
5	2016 ELA/L	36,909	0.84	**0.91**		
5	2015 Math	37,110	0.81	0.76	**0.93**	
5	2016 Math	37,110	0.78	0.80	0.87	**0.93**
6	2015 ELA/L	36,877	**0.91**			
6	2016 ELA/L	36,877	0.83	**0.89**		
6	2015 Math	37,040	0.80	0.75	**0.93**	
6	2016 Math	37,040	0.79	0.81	0.86	**0.93**
7	2015 ELA/L	37,948	**0.92**			
7	2016 ELA/L	37,948	0.83	**0.89**		
7	2015 Math	38,081	0.81	0.78	**0.91**	
7	2016 Math	38,081	0.78	0.81	0.88	**0.93**
8	2015 ELA/L	37,132	**0.92**			
8	2016 ELA/L	37,132	0.83	**0.90**		
8	2015 Math	37,217	0.81	0.76	**0.92**	
8	2016 Math	37,217	0.77	0.79	0.86	**0.91**

5. RELIABILITY

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

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

5.1 MARGINAL RELIABILITY

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

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

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

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

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

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

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

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

Table 30. Marginal Reliability for ELA/L and Mathematics

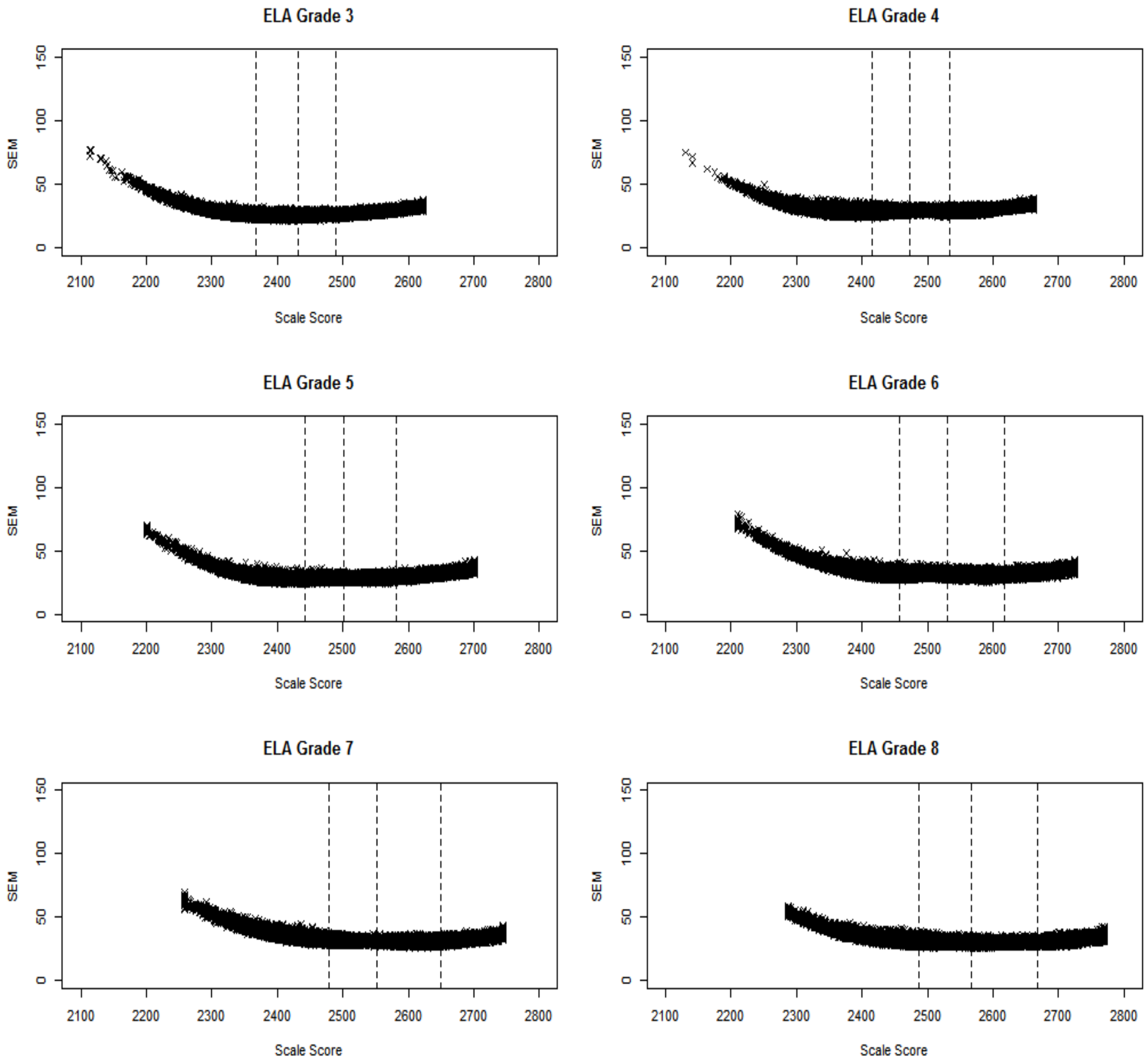
Grade	N	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
ELA/L							
3	38,942	38	40	0.91	2438	89	27
4	38,450	38	40	0.90	2480	96	30
5	39,010	38	41	0.91	2517	97	30
6	39,071	38	41	0.89	2536	98	33
7	40,085	38	41	0.89	2559	100	33
8	39,351	40	41	0.90	2574	100	32
Mathematics							
3	38,870	39	40	0.94	2438	81	19
4	38,387	37	40	0.94	2478	82	20
5	38,941	38	40	0.93	2501	89	23
6	38,965	38	39	0.93	2521	104	27
7	39,961	38	40	0.93	2538	108	28
8	39,181	38	40	0.91	2551	116	34

5.2 STANDARD ERROR CURVES

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

Overall, the standard error curves suggest that students are measured with a high degree of precision given that the standard errors are consistently low. However, larger standard errors are observed at the lower ends of the score distribution relative to the higher ends. This occurs because the item pools currently have a shortage of very easy items that are better targeted toward these lower-achieving students. Content experts use this information to consider how to further target and populate item pools.

Figure 5. Conditional Standard Error of Measurement for ELA/L



ELA Grade 3

ELA Grade 4

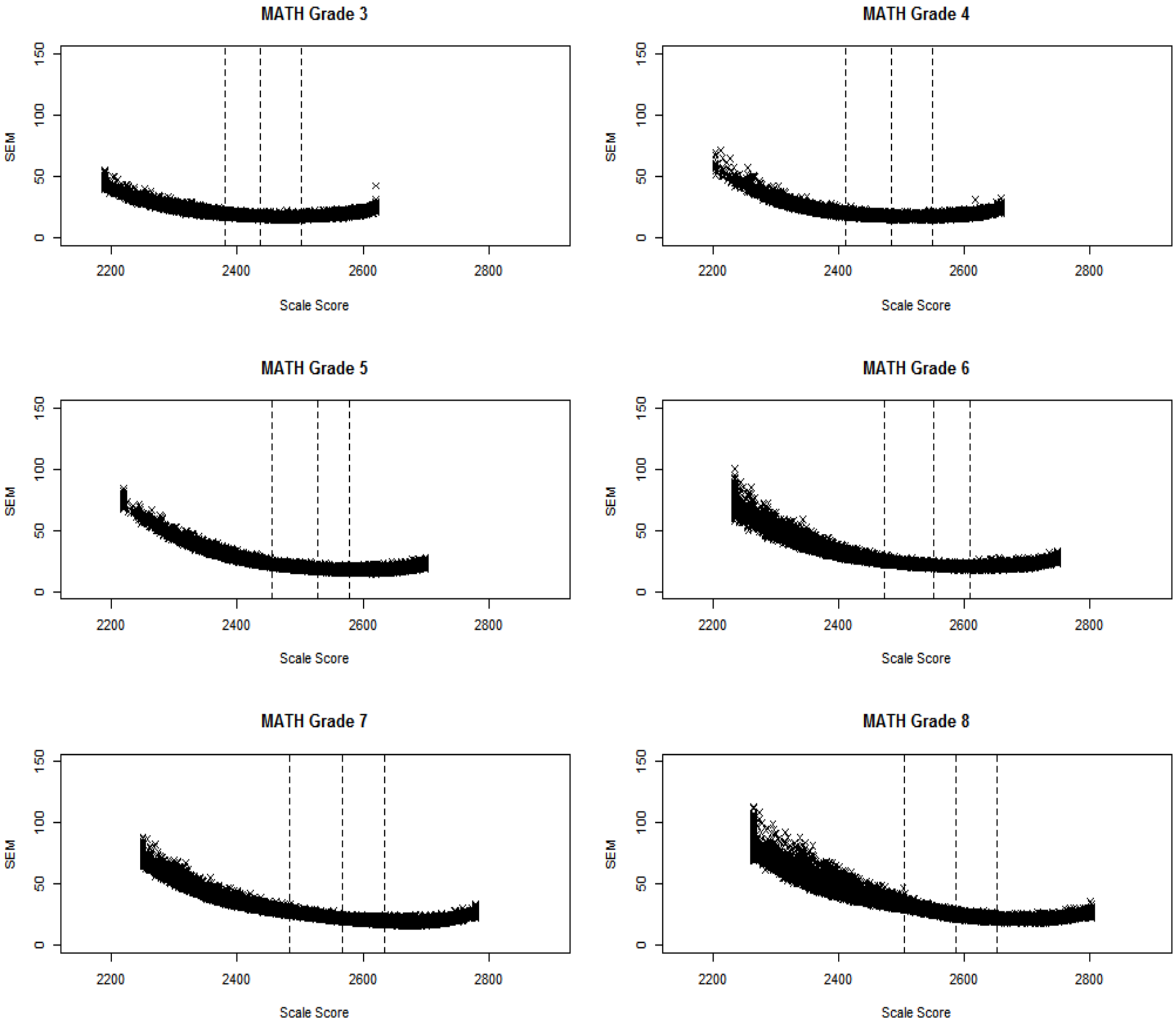
ELA Grade 5

ELA Grade 6

ELA Grade 7

ELA Grade 8

Figure 6. Conditional Standard Error of Measurement for Mathematics



The SEMs presented in the figures above are summarized in Tables 31 and 32. Table 31 provides the average conditional SEM for all scores and scores in each achievement level. Table 32 presents the average conditional SEMs at the each cut score and the difference in average conditional SEMs between two cut scores. As shown in Figures 5 and 6, the greatest average conditional SEM is in Level 1 in both ELA/L and mathematics. Average conditional SEMs at all cut scores are similar in ELA/L, but larger in Level 2 cut scores in mathematics.

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

Grade	Level 1	Level 2	Level 3	Level 4	Average CSEM
ELA/L					
3	30	25	26	28	27
4	31	29	29	30	30
5	31	28	28	31	30
6	37	32	31	33	33
7	38	32	31	33	33
8	36	30	30	32	32
Mathematics					
3	24	18	17	19	19
4	25	18	17	19	20
5	30	20	18	18	23
6	36	23	21	21	27
7	38	24	21	20	28
8	46	29	23	22	34

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

Grade	L2 Cut	L3 Cut	L4 Cut	|L2-L3|	|L3-L4|	|L2-L4|
ELA/L						
3	26	25	26	1	1	0
4	28	29	29	1	0	1
5	27	28	29	1	1	2
6	32	32	31	0	1	1
7	32	31	31	1	0	1
8	31	30	31	1	1	0
Mathematics						
3	20	17	17	3	0	3
4	20	17	17	3	0	3
5	22	18	18	4	0	4
6	25	21	20	4	1	5
7	27	22	20	5	2	7
8	33	25	22	8	3	11

5.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

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

For a fixed-form test, the accuracy and consistency of classifications are estimated on a single-form's test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the CAT, because the adaptive testing algorithm constructs a test form unique to each student, the classification indexes are computed based on all sets of items administered across students using an IRT based method (Guo, 2006).

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

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

For the *i*th student, the student's estimated ability is $\hat{\theta}_i$ with SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed, as $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, 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, the probability of at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

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

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

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

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

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

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

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = \left(a_j, b_j, c_j\right)$ 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, $c_j = 0$), 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}, 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 33 provides the proportion of classification accuracy and consistency for overall and by achievement level.

The overall classification index ranged from 0.76 to 0.84 for the accuracy and from 0.67 to 0.77 for the consistency across all grades and subjects. For achievement levels, the classification index is higher in L1 and L4 than in L2 and L3. The higher accuracy at L1 and L4 is due to the intervals used to compute the classification probability to classify students into L1 $[-\infty,$L2 cut] or L4 [L4 cut, ∞] is wider than the intervals used in L2 [L2 cut, L3 cut] and L3 [L3 cut, L4 cut]. The misclassification probability tends to be higher for narrow intervals.

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

Table 33. Classification Accuracy and Consistency by Achievement Levels

Grade	Achievement Level	ELA/L		Mathematics	
		% Accuracy	% Consistency	% Accuracy	% Consistency
3	Overall	0.78	0.70	0.82	0.76
	L1	0.88	0.82	0.90	0.84
	L2	0.69	0.58	0.73	0.64
	L3	0.65	0.54	0.79	0.72
	L4	0.88	0.82	0.89	0.84
4	Overall	0.77	0.69	0.84	0.77
	L1	0.89	0.83	0.89	0.83
	L2	0.61	0.48	0.81	0.73
	L3	0.62	0.51	0.79	0.71
	L4	0.87	0.82	0.89	0.84
5	Overall	0.78	0.70	0.83	0.76
	L1	0.89	0.83	0.91	0.86
	L2	0.63	0.51	0.78	0.69
	L3	0.72	0.62	0.71	0.61
	L4	0.86	0.79	0.89	0.84
6	Overall	0.76	0.67	0.82	0.75
	L1	0.87	0.80	0.91	0.86
	L2	0.65	0.54	0.77	0.69
	L3	0.71	0.62	0.72	0.61
	L4	0.83	0.74	0.89	0.83
7	Overall	0.77	0.69	0.83	0.76
	L1	0.87	0.80	0.91	0.86
	L2	0.65	0.54	0.77	0.69
	L3	0.75	0.67	0.75	0.66
	L4	0.84	0.75	0.90	0.84
8	Overall	0.78	0.70	0.82	0.75
	L1	0.87	0.80	0.90	0.85
	L2	0.69	0.58	0.72	0.62
	L3	0.77	0.70	0.72	0.62
	L4	0.83	0.74	0.90	0.85

5.4 RELIABILITY FOR SUBGROUPS

The reliability of test scores and achievement levels are also computed by subgroups. Tables 34 and 35 present the marginal reliability coefficients by the subgroups. The reliability coefficients are similar across subgroups, but somewhat lower for Limited English Proficiency (LEP) and IDEA subgroups, a large percentage of whom received Level 1 with large SEMs. The classification indexes by subgroups are provided in Appendix C.

Table 34. Marginal Reliability Coefficients for Overall and by Subgroup for ELA/L

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.91	0.90	0.91	0.89	0.89	0.90
Female	0.91	0.90	0.90	0.88	0.88	0.89
Male	0.91	0.90	0.91	0.89	0.89	0.90
American Indian/ Alaska Native	0.88	0.91	0.90	0.88	0.88	0.89
Asian	0.89	0.89	0.89	0.87	0.87	0.88
African American	0.89	0.89	0.89	0.86	0.85	0.88
Hispanic/Latino	0.89	0.89	0.90	0.87	0.87	0.88
Pacific Islander	0.91	0.89	0.92	0.91	0.92	0.91
White	0.89	0.88	0.88	0.86	0.87	0.88
Multiple Ethnicities	0.90	0.90	0.91	0.88	0.89	0.89
Limited English Proficiency	0.84	0.84	0.83	0.76	0.72	0.72
IDEA	0.86	0.86	0.86	0.82	0.82	0.84

Table 35. Marginal Reliability Coefficients for Overall and by Subgroup for Mathematics

Subgroup	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
All Students	0.94	0.94	0.93	0.93	0.93	0.91
Female	0.94	0.94	0.93	0.93	0.93	0.91
Male	0.95	0.95	0.94	0.94	0.94	0.92
American Indian/ Alaska Native	0.93	0.94	0.92	0.92	0.90	0.88
Asian	0.94	0.95	0.95	0.94	0.95	0.94
African American	0.92	0.91	0.87	0.89	0.88	0.84
Hispanic/Latino	0.92	0.92	0.89	0.89	0.90	0.85
Pacific Islander	0.79	0.94	0.95	0.95	0.95	0.93
White	0.93	0.93	0.93	0.93	0.93	0.92
Multiple Ethnicities	0.94	0.95	0.94	0.93	0.93	0.92
Limited English Proficiency	0.90	0.89	0.81	0.79	0.80	0.63
IDEA	0.92	0.90	0.85	0.85	0.85	0.76

5.5 RELIABILITY FOR CLAIM SCORES

The marginal reliability coefficients and the measurement errors are also computed for the claim scores. In mathematics, claims 2 and 4 are combined to have enough items to generate a score. Because the precision of scores in claims is not sufficient to report scores, given a small number of items, the scores on each claim are reported using one of the three achievement categories, taking into account the SEM of the claim score: (1) Below standard, (2) At/Near standard, or (3) Above standard. Tables 36 and 37 present the marginal reliability coefficients for each claim score in ELA/L and mathematics, respectively.

Table 36. Marginal Reliability Coefficients for Claim Scores in ELA/L

Grade	Reporting Categories	Number of Items Specified in Test Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Claim 1: Reading	14	16	0.75	2439	101	51
	Claims 2 & 4: Writing & Research	16	16	0.82	2431	98	41
	Claim 3: Listening	8	8	0.55	2445	115	77
4	Claim 1: Reading	14	16	0.78	2476	107	50
	Claims 2 & 4: Writing & Research	16	16	0.80	2477	103	46
	Claim 3: Listening	8	8	0.53	2487	123	85
5	Claim 1: Reading	14	16	0.77	2509	109	52
	Claims 2 & 4: Writing & Research	16	16	0.80	2525	101	46
	Claim 3: Listening	8	9	0.59	2508	130	84
6	Claim 1: Reading	14	16	0.73	2514	117	61
	Claims 2 & 4: Writing & Research	16	16	0.77	2543	103	49
	Claim 3: Listening	8	9	0.50	2554	126	89
7	Claim 1: Reading	14	16	0.75	2553	112	56
	Claims 2 & 4: Writing & Research	16	16	0.77	2557	107	52
	Claim 3: Listening	8	9	0.50	2567	123	87
8	Claim 1: Reading	16	16	0.78	2570	109	52
	Claims 2 & 4: Writing & Research	16	16	0.79	2570	111	51
	Claim 3: Listening	8	9	0.51	2583	118	82

Table 37. Marginal Reliability Coefficients for Claim Scores in Mathematics

Grade	**Reporting Categories**	**Number of Items Specified in Test Blueprint**		**Marginal Reliability**	**Scale Score Mean**	**Scale Score SD**	**Average CSEM**
		Min	**Max**				
3	Claim 1	20	20	0.90	2440	84	27
	Claims 2 and 4	8	11	0.69	2428	97	54
	Claim 3	9	11	0.67	2434	98	56
4	Claim 1	20	20	0.90	2479	85	28
	Claims 2 and 4	8	10	0.67	2467	103	59
	Claim 3	9	10	0.75	2473	96	48
5	Claim 1	20	20	0.88	2501	93	32
	Claims 2 and 4	8	10	0.61	2486	119	74
	Claim 3	9	10	0.69	2493	110	62
6	Claim 1	19	19	0.88	2520	110	38
	Claims 2 and 4	9	10	0.68	2509	123	70
	Claim 3	10	11	0.66	2518	118	69
7	Claim 1	20	20	0.88	2539	112	39
	Claims 2 and 4	10	10	0.67	2523	127	73
	Claim 3	8	10	0.70	2533	123	67
8	Claim 1	20	20	0.86	2549	122	45
	Claims 2 and 4	8	10	0.53	2523	160	110
	Claim 3	9	10	0.62	2542	135	83

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

6. SCORING

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

6.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

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

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

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

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

in the case of items with two or more points,

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

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

Standard Error of Measurement

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

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

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

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

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

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

6.2 RULES FOR TRANSFORMING THETA TO VERTICAL SCALE SCORES

The student's performance in each subject is summarized in an overall test score referred to as a *scale score*. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \theta + b$. The scaling constants a and b are provided by the Smarter Balanced Assessment consortium. Table 38 lists the scaling constants for each subject for the theta-to-scale score linear transformation. Scale scores are rounded to an integer.

Table 38. Vertical Scaling Constants on the Reporting Metric

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

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

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

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

The scale scores are mapped into four achievement levels using three achievement standards (i.e., cut scores). Table 39 provides three achievement standards for each grade and content area.

Table 39. 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 40 presents the lowest obtainable score (LOT or LOSS) and the highest obtainable score (HOT or HOSS) in both theta and scale score metrics. Estimated theta's lower than LOT or higher than HOT are truncated to the LOT and HOT values, and assign LOSS and HOSS associated with the LOT and HOT. LOT and HOT were applied to all tests and all scores (total and subscores). The standard error for LOT and HOT are computed using the LOT and HOT ability estimates given the administered items.

Table 40. Lowest and Highest Obtainable Scores

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

6.4 SCORING ALL CORRECT AND ALL INCORRECT CASES

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

6.5 Rules for Calculating Strengths and Weaknesses for Reporting Categories (Claim Scores)

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

For summative tests, the specific rules are as follows:

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

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

6.6 Target Scores

The target-level reports are not possible to produce for a fixed-form test because the number of items included per target (i.e., benchmark) is too few to produce a reliable score at the target level. A typical fixed-form test includes only one or two items per target. Even when aggregated, these data reflect the benchmark narrowly because they reflect only one or two ways of measuring the target. However, an adaptive test offers a tremendous opportunity for target-level data at the class, school, and district area level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given target. A target score is an aggregate of the differences in student overall proficiency and the differences in the difficulty of the items measuring a target in a class, school, or district area. Target scores are computed for attempted tests based on the responded items. Target scores are computed within each claim (three claims) in ELA/L and Claim 1 only in mathematics.

Target scores will be computed as following:

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

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

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

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

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

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

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

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

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 performs better, worse, or as expected on this target. In some cases, insufficient information will be available and that will be indicated as well.

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

- If $\bar{\delta}_{Tg} - se(\bar{\delta}_{Tg}) \geq 0.07$, then performance is 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.

6.7 HUMAN SCORING

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

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

6.7.1 Reader Selection

MI maintains a large pool of qualified, experienced readers at each scoring center, as well as distributive readers who work remotely from their homes. MI only needs to inform the readers that a project is pending and invite them to return. MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. MI employs many of these experienced readers for SBAC project and recruit new ones as well.

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

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

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

MI requires all hand-scoring project staff (scoring directors, team leaders, readers, and clerical staff) to sign a confidentiality/nondisclosure agreement before receiving any training or secure project materials. The employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or the scoring methods to any person.

6.7.2 Reader Training

All readers hired for Smarter Balanced assessment hand-scoring are trained using the rubric(s), anchor sets, and training/qualifying sets provided by SBAC. These sets were created during the original field-test scoring in 2014 and approved by SBAC. The same anchor sets are used each year. The only changes made to anchor sets across the years include occasional updates to annotations and removal of individual responses, as determined during annual meetings between the vendors and SBAC. Additionally, several of the brief writes anchor sets were revised between the 2015 and 2016 test administrations. Readers are placed into a scoring group that corresponds to the subject/grade that they are deemed best suited to score (based on work history, results of the placement assessments, and performance on past scoring projects). They are trained on a specific item type (i.e., brief write, reading, research, full write, and/or mathematics). Within each group, readers are divided into teams consisting of one team leader and 10–15 readers. Each team leader and reader is assigned a unique number for easy identification of their scoring work throughout the scoring session.

MI's Virtual Scoring Center (VSC) online training interface presents rubrics, scoring guides, and training/qualifying sets in three modes:

- In-person training with a scoring director
- Distance webinar training with a live trainer
- Remote self-training

Regardless of mode, the same training protocol is followed.

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

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

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

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

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

Training design varies slightly depending on Smarter Balanced item type:

- Full writes: readers train and qualify on baseline sets for each grade and writing purpose (Grade 3 Narrative, Grade 6 Argumentative, etc.), then take qualifying sets for each item in that grade and purpose.
- Brief writes, reading, and research: readers train and qualify on a baseline set within a specific grade band and target.
- Mathematics: readers train on baseline items, which qualify the readers for that item as well as any items associated with it; for items with no associated items, training is for the specific item.

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

6.7.3 Reader Statistics and Analyses

One concern regarding the scoring of any open-response assessment is the reliability and accuracy of the scoring. MI appreciates and shares this concern and continually develops new and technically sound methods of monitoring reliability. Reliable scoring starts with detailed scoring rubrics and training materials, and thorough training sessions by experienced trainers. Quality results are achieved by daily monitoring of each reader. Unbiased scoring is ensured because the only identifying information on the student response is the identification number. Unless the students sign their names, write about their hometowns, or in some way provide other identifying information, the readers have no knowledge of them.

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

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

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

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

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

6.7.4 Reader Monitoring and Retraining

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

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

Each student response is scored holistically by a trained and qualified reader using the scoring scales developed and approved by SBAC, with a second read conducted on 15% of responses for each item for reliability purposes. Responses are selected randomly for second reading and scored by readers who are unaware that the response has been read before. The second reader is also not aware of the score the response received. MI's QA/reliability procedures allow their hand-scoring staff to identify struggling readers very early and begin retraining immediately. While retraining these readers, MI also monitors their scoring intensively to ensure that all responses are scored accurately. In fact, MI's monitoring is also used as a retraining method. MI shows readers responses that the readers have scored incorrectly, explains the correct scores, and has the readers change the scores.

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

6.7.5 Reader Validity Checks

Approved responses are loaded into the VSC system as validity responses. A small set of validity responses are provided by SBAC for all vendors to use, and these are supplemented with responses selected and approved by MI scoring management. The "true" or range finding scores for these responses are entered into a validity database. These responses are imbedded into live scoring on an ongoing basis to be scored by the readers. A validity report is generated that includes the response identification number, the score(s) assigned by the readers, and the "true" scores. A daily and project-to-date summary of percentages of correct scores and low/high considerations at each score point is also provided. If it is determined that a validity response and/or item is performing poorly, scoring management reviews the validity responses to ensure that the true scores have been entered correctly. If so, then retraining is conducted with the readers using the validity data as a guide for how to focus the retraining. If the true scores have been entered incorrectly, then the database is updated to show the correct true scores.

6.7.6 Reader Dismissal

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

6.7.7 Reader Agreements

The inter-reader reliability is computed based on scorable responses (numeric scores) scored by two independent readers only, excluding non-scorable responses (e.g., off topic, off purpose, or foreign language responses) which were scored by the leadership readers, not by two independent readers. The inter-reader reliability is based on the combined data across 10 states (Delaware, Hawaii, Idaho, New Hampshire, Oregon, South Dakota, Vermont, Washington, West Virginia, and Connecticut) and the U.S.

Virgin Islands because the number of responses with two independent readers, after removing responses with condition codes, is too small to compute inter-reader reliability by state.

In ELA/L, the short answer items are scored in 0–2. In mathematics, the maximum score points of the hand-scored items range from 1–3. In an adaptive test, because items are selected adapting to a student's ability while meeting the test blueprint, item usages vary across items. Tables 41 and 42 provide a summary of the inter-reader reliability based on items with a sample size greater than 50. The inter-reader reliability is presented with %exact agreement, minimum and maximum %exact agreements, combined %exact and %adjacent agreement, and quadratic weighted Kappa (QWK).

Table 41. ELA/L Reader Agreements for Short-Answer Items

Grade	# of Items	%Exact			%(Exact+ Adjacent)	QWK
		Average	**Min**	**Max**		
3	38	75	59	91	99	0.66
4	53	76	61	93	99	0.70
5	55	73	54	88	98	0.70
6	44	71	61	89	98	0.62
7	53	72	57	92	98	0.65
8	59	69	55	93	98	0.63

Table 42. Mathematics Reader Agreements

Grade	Score Points	# of Items	%Exact			%(Exact+ Adjacent)	QWK
			Average	**Min**	**Max**		
3	1	13	93	88	99	100	0.84
4	1	8	83	74	96	100	0.61
5	1	8	94	90	99	100	0.80
6	1	18	96	90	100	100	0.91
7	1	10	96	93	100	100	0.83
8	1	15	89	79	97	100	0.75
3	2	27	89	76	99	99	0.87
4	2	37	88	75	98	98	0.83
5	2	44	88	79	99	99	0.82
6	2	31	85	71	95	98	0.80
7	2	30	88	76	100	99	0.82
8	2	24	87	81	97	99	0.82
3	3	4	95	94	96	99	0.97
4	3	4	86	84	87	99	0.92
5	3	8	85	79	99	96	0.80
7	3	3	78	70	82	99	0.88

7. REPORTING AND INTERPRETING SCORES

The Online Reporting System (ORS) generates a set of online score reports that include the information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete tests and the tests are hand scored. Because the score report on students' performance are updated each time students complete tests and they are hand-scored, authorized users (e.g., school principals, teachers) can view students' performance on the tests and use them to improve student learning. In addition to individual students' score reports, the Online Reporting System also produces aggregate score reports by class, schools, districts, and states. It should be noted that the ORS does not produce aggregate score reports for state. The timely accessibility of aggregate score reports could help users monitor students testing in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year. Additionally, the ORS provides participation data that helps monitor the student participation rate.

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

7.1 ONLINE REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

7.1.1 Types of Online Score Reports

The ORS is designed to help educators and students answer questions regarding how well students have performed on ELA/L and mathematics assessments. The ORS is the online tool to provide educators and other stakeholders with timely, relevant score reports. The ORS for the Smarter Balanced assessments has been designed with stakeholders, who are not technical measurement experts, in mind, ensuring that test results are presented as easy to read and understand by using simple language so that users can quickly understand assessment results and make inferences about student achievement. The ORS is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar elements.

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

Generally, the ORS provides two categories of online score reports: (1) aggregate score reports and (2) student score reports. Table 43 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Online Reporting System User Guide*, located in a help button on the ORS.

Table 43. 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 percent of students with Level 3 or 4 (overall students and by subgroup) • Average scale score and standard error of average scale score (overall students and by subgroup) • Percent of students at each achievement level on overall test and by claims (overall students and by subgroup) • Performance category in each target (overall students)[1] • Participation rate (overall students)[2] • On-demand student roster report
Student	• Total scale score and standard error of measurement • Achievement level on overall and claim scores with achievement level descriptors • Average scale scores and standard errors of average scale scores for student's school, and district

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

The aggregate score reports at a selected aggregate level are provided for overall students and by subgroups. Users can see student assessment results by any of the subgroups. Table 44 presents the types of subgroups and subgroup categories provided in ORS.

Table 44. Types of Subgroups

Subgroup	**Subgroup Category**
Gender	Male Female
IDEA Indicator	Special Education Unknown
Limited English Proficiency (LEP) Status	Yes Unknown
Ethnicity	American Indian or Alaska Native Asian Black or African American Two or More Races Hispanic or Latino White Native Hawaiian or Other Pacific Islander

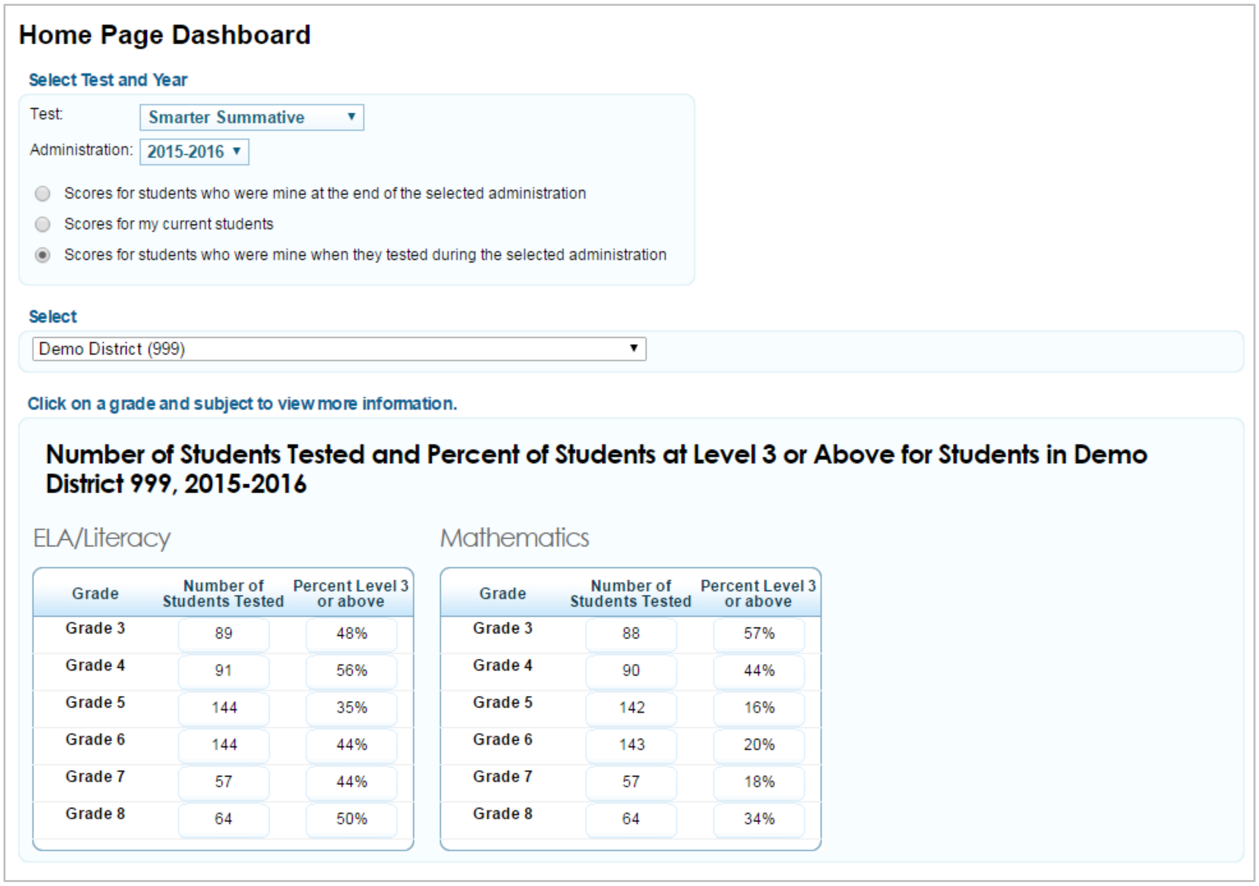
7.1.2 The Online Reporting System

7.1.2.1 Home Page

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

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

Exhibit 1. Home Page: District Level



Home Page Dashboard

Select Test and Year

Test: Smarter Summative ▼

Administration: 2015-2016 ▼

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

○ Scores for my current students

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

Select

Demo District (999) ▼

Click on a grade and subject to view more information.

Number of Students Tested and Percent of Students at Level 3 or Above for Students in Demo District 999, 2015-2016

ELA/Literacy

Grade	Number of Students Tested	Percent Level 3 or above
Grade 3	89	48%
Grade 4	91	56%
Grade 5	144	35%
Grade 6	144	44%
Grade 7	57	44%
Grade 8	64	50%

Mathematics

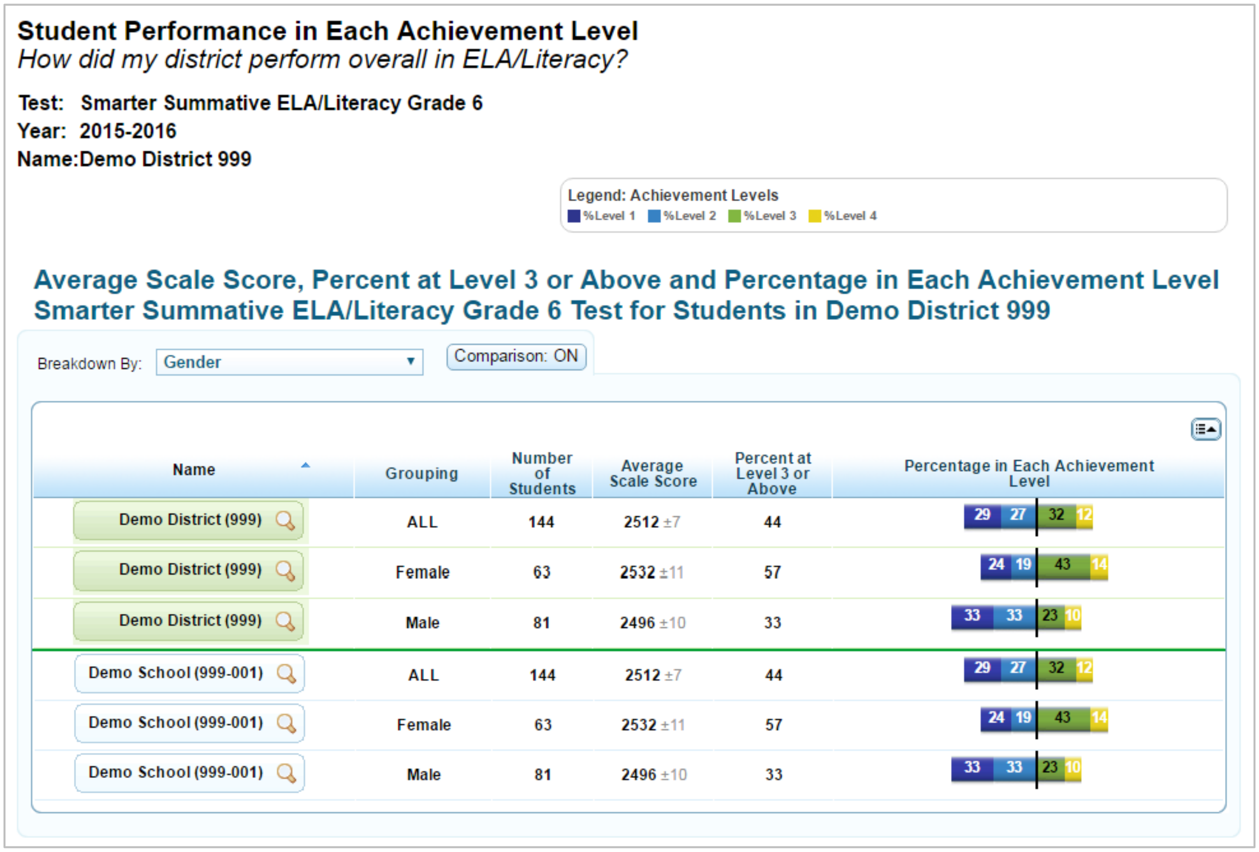
Grade	Number of Students Tested	Percent Level 3 or above
Grade 3	88	57%
Grade 4	90	44%
Grade 5	142	16%
Grade 6	143	20%
Grade 7	57	18%
Grade 8	64	34%

7.1.2.2 Subject Detail Page

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

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

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



Student Performance in Each Achievement Level
How did my district perform overall in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo District 999

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

Average Scale Score, Percent at Level 3 or Above and Percentage in Each Achievement Level
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo District 999

Breakdown By: Gender ▾ Comparison: ON

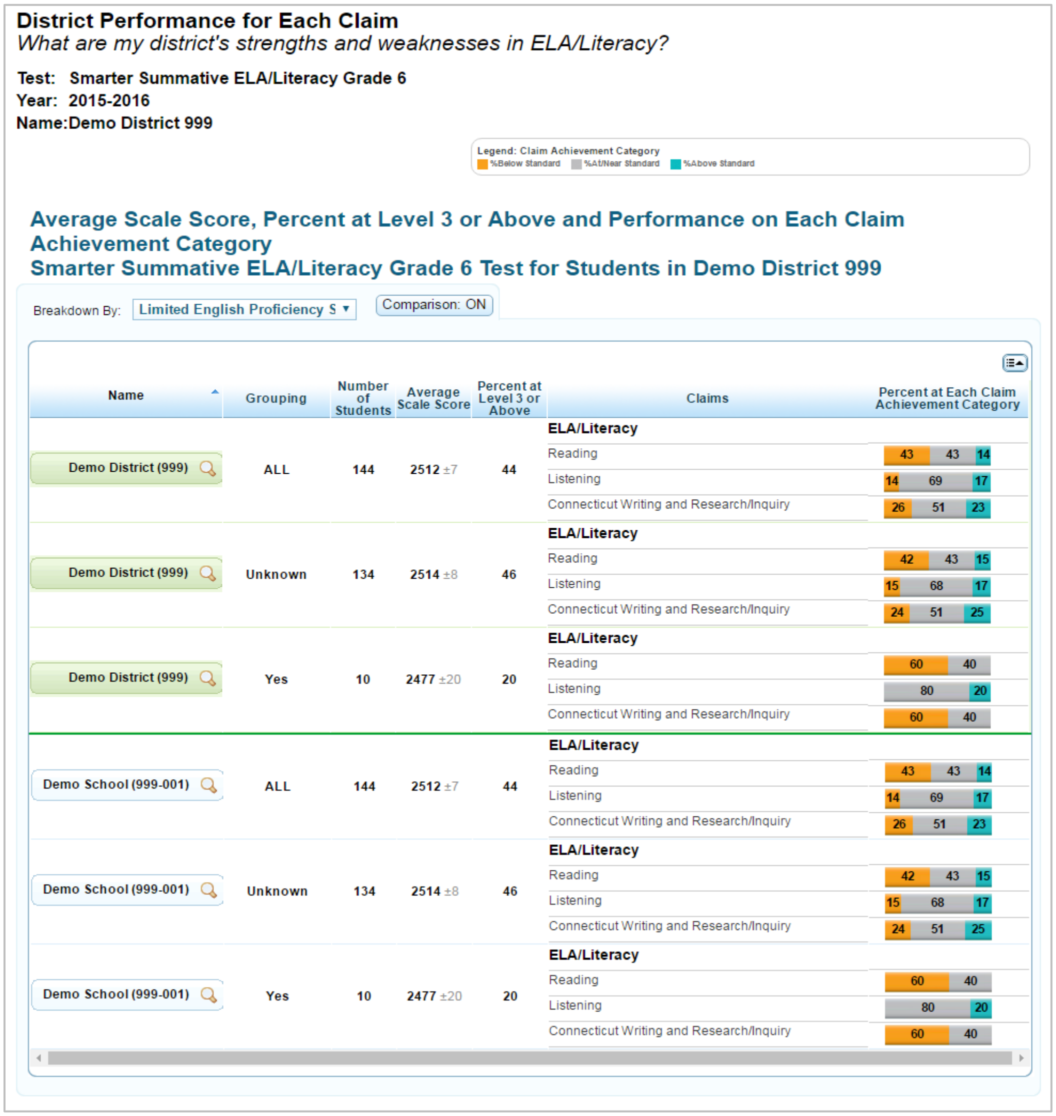
Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Percentage in Each Achievement Level
Demo District (999)	ALL	144	2512 ±7	44	29 27 32 12
Demo District (999)	Female	63	2532 ±11	57	24 19 43 14
Demo District (999)	Male	81	2496 ±10	33	33 33 23 10
Demo School (999-001)	ALL	144	2512 ±7	44	29 27 32 12
Demo School (999-001)	Female	63	2532 ±11	57	24 19 43 14
Demo School (999-001)	Male	81	2496 ±10	33	33 33 23 10

7.1.2.3 Claim Detail Page

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

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

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



District Performance for Each Claim
What are my district's strengths and weaknesses in ELA/Literacy?

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo District 999

Legend: Claim Achievement Category
🟧 %Below Standard ⬜ %At/Near Standard 🟦 %Above Standard

Average Scale Score, Percent at Level 3 or Above and Performance on Each Claim Achievement Category
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo District 999

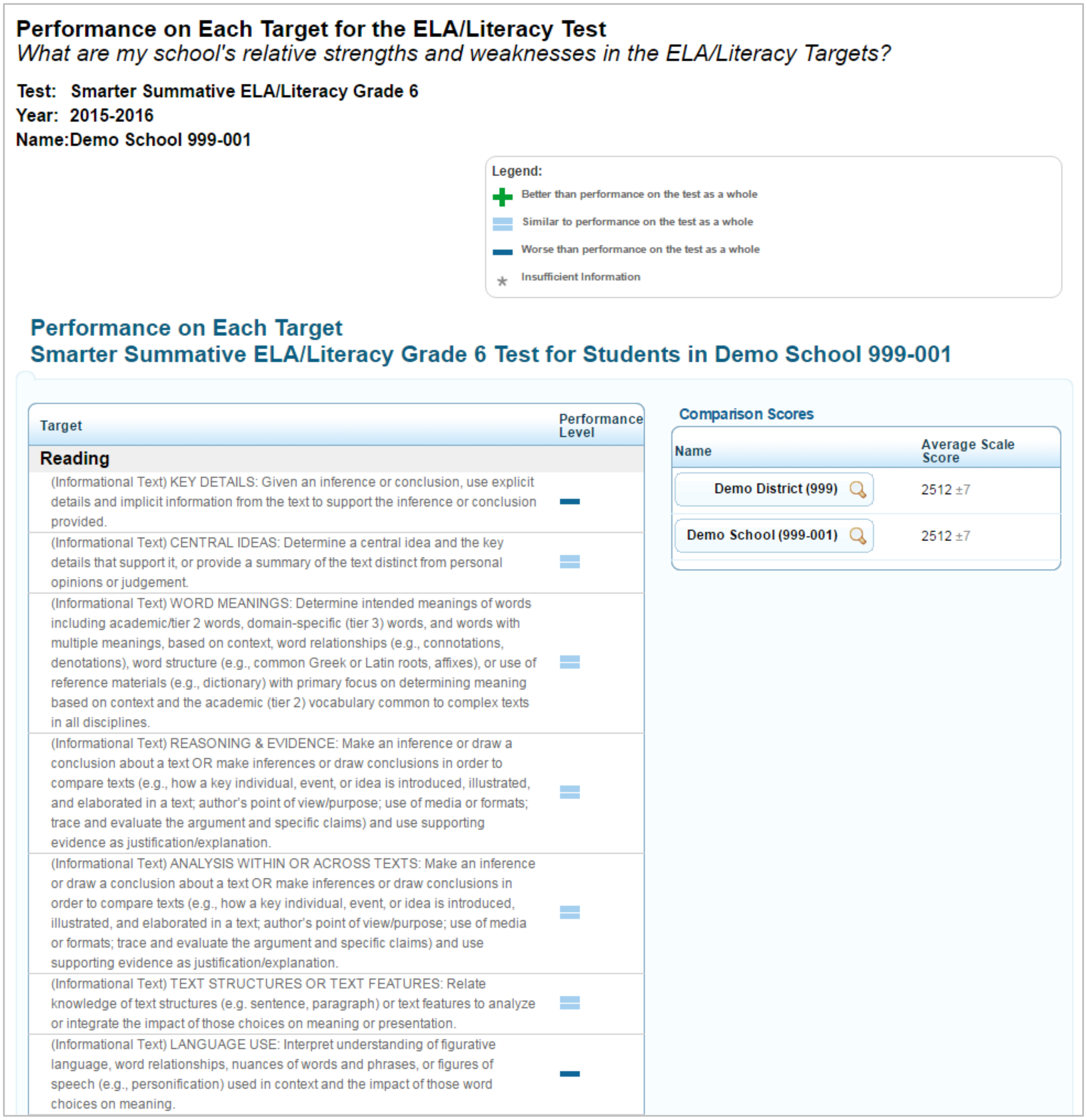
Breakdown By: Limited English Proficiency S ▾ Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent at Level 3 or Above	Claims	Percent at Each Claim Achievement Category
Demo District (999)	ALL	144	2512 ±7	44	**ELA/Literacy**	
					Reading	43 43 14
					Listening	14 69 17
					Connecticut Writing and Research/Inquiry	26 51 23
Demo District (999)	Unknown	134	2514 ±8	46	**ELA/Literacy**	
					Reading	42 43 15
					Listening	15 68 17
					Connecticut Writing and Research/Inquiry	24 51 25
Demo District (999)	Yes	10	2477 ±20	20	**ELA/Literacy**	
					Reading	60 40
					Listening	80 20
					Connecticut Writing and Research/Inquiry	60 40
Demo School (999-001)	ALL	144	2512 ±7	44	**ELA/Literacy**	
					Reading	43 43 14
					Listening	14 69 17
					Connecticut Writing and Research/Inquiry	26 51 23
Demo School (999-001)	Unknown	134	2514 ±8	46	**ELA/Literacy**	
					Reading	42 43 15
					Listening	15 68 17
					Connecticut Writing and Research/Inquiry	24 51 25
Demo School (999-001)	Yes	10	2477 ±20	20	**ELA/Literacy**	
					Reading	60 40
					Listening	80 20
					Connecticut Writing and Research/Inquiry	60 40

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, and (2) average scale scores and standard errors of average scale scores for the selected aggregate unit and the aggregate unit above the selected aggregate. It should be noted that the summaries on target-level student performance are generated for overall students only. That is, the summaries on target-level student performance are not generated by subgroup. Exhibits 4-7 present examples of target detail pages for ELA/L and mathematics at the school level and the teacher level.

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



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

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo School 999-001

Legend:

- ✚ Better than performance on the test as a whole
- ═ Similar to performance on the test as a whole
- − Worse than performance on the test as a whole
- * Insufficient Information

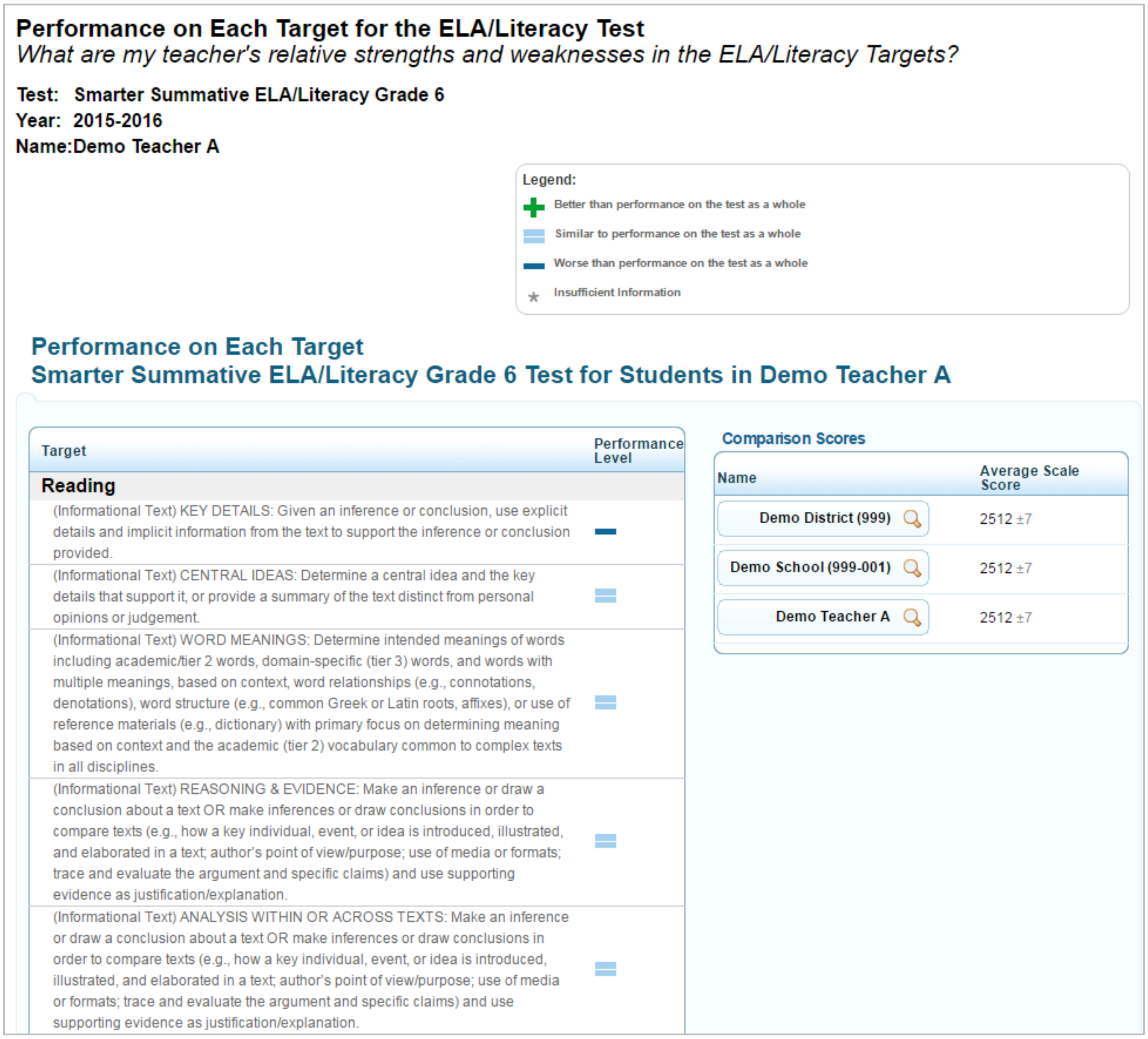
Performance on Each Target
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo School 999-001

Target	Performance Level
Reading	
(Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	−
(Informational Text) CENTRAL IDEAS: Determine a central idea and the key details that support it, or provide a summary of the text distinct from personal opinions or judgement.	═
(Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., connotations, denotations), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary) with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	═
(Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	═
(Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., how a key individual, event, or idea is introduced, illustrated, and elaborated in a text; author's point of view/purpose; use of media or formats; trace and evaluate the argument and specific claims) and use supporting evidence as justification/explanation.	═
(Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g. sentence, paragraph) or text features to analyze or integrate the impact of those choices on meaning or presentation.	═
(Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., personification) used in context and the impact of those word choices on meaning.	−

Comparison Scores

Name	Average Scale Score
Demo District (999)	2512 ±7
Demo School (999-001)	2512 ±7

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



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

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo Teacher A

Legend:

- ✚ Better than performance on the test as a whole
- = Similar to performance on the test as a whole
- – Worse than performance on the test as a whole
- ☆ Insufficient Information

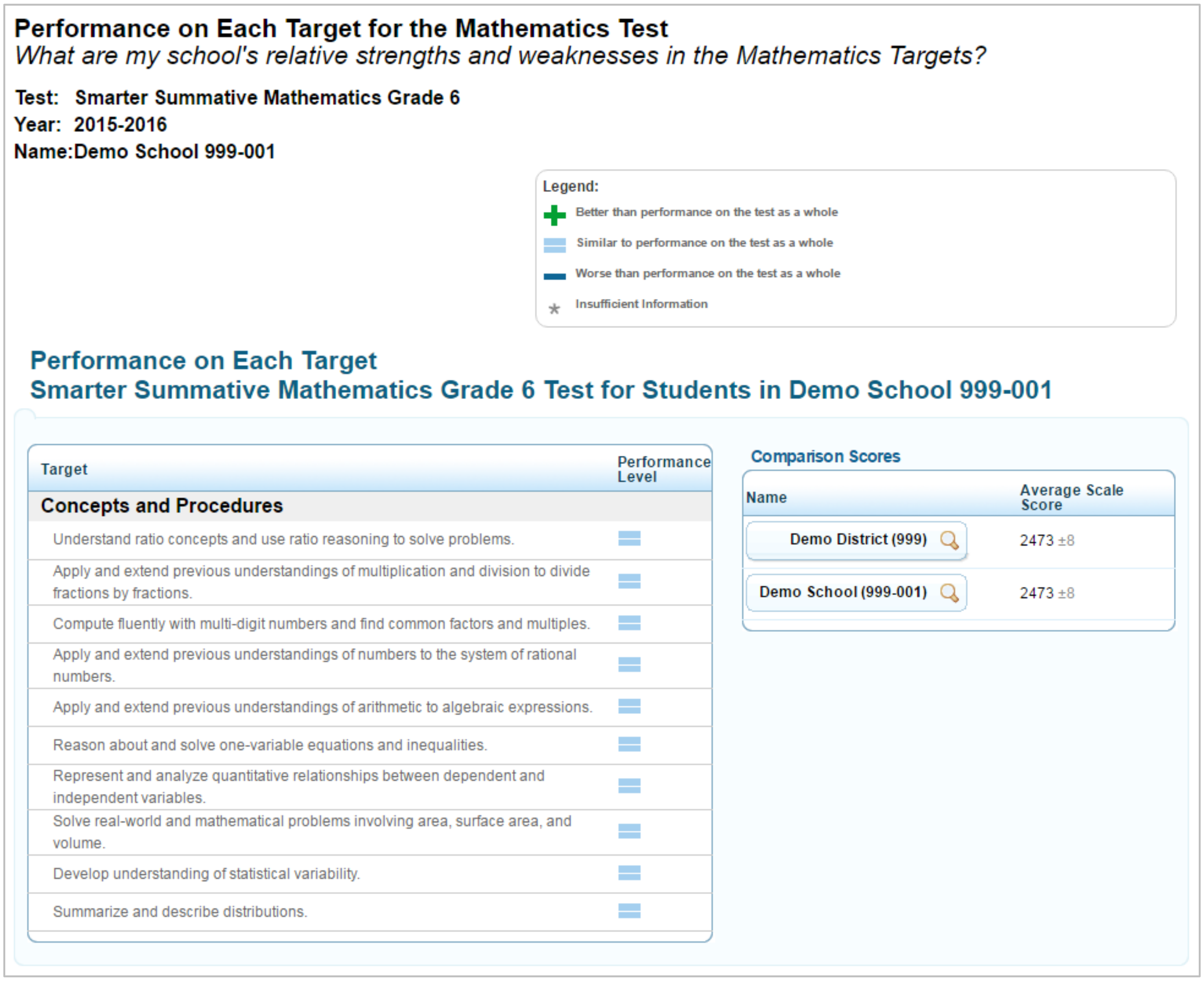
Performance on Each Target
Smarter Summative ELA/Literacy Grade 6 Test for Students in Demo Teacher A

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

Comparison Scores

Name	Average Scale Score
Demo District (999)	2512 ±7
Demo School (999-001)	2512 ±7
Demo Teacher A	2512 ±7

Exhibit 6. Target Detail Page for Mathematics: School Level



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

Test: Smarter Summative Mathematics Grade 6
Year: 2015-2016
Name:Demo School 999-001

Legend:
➕ Better than performance on the test as a whole
= Similar to performance on the test as a whole
▬ Worse than performance on the test as a whole
★ Insufficient Information

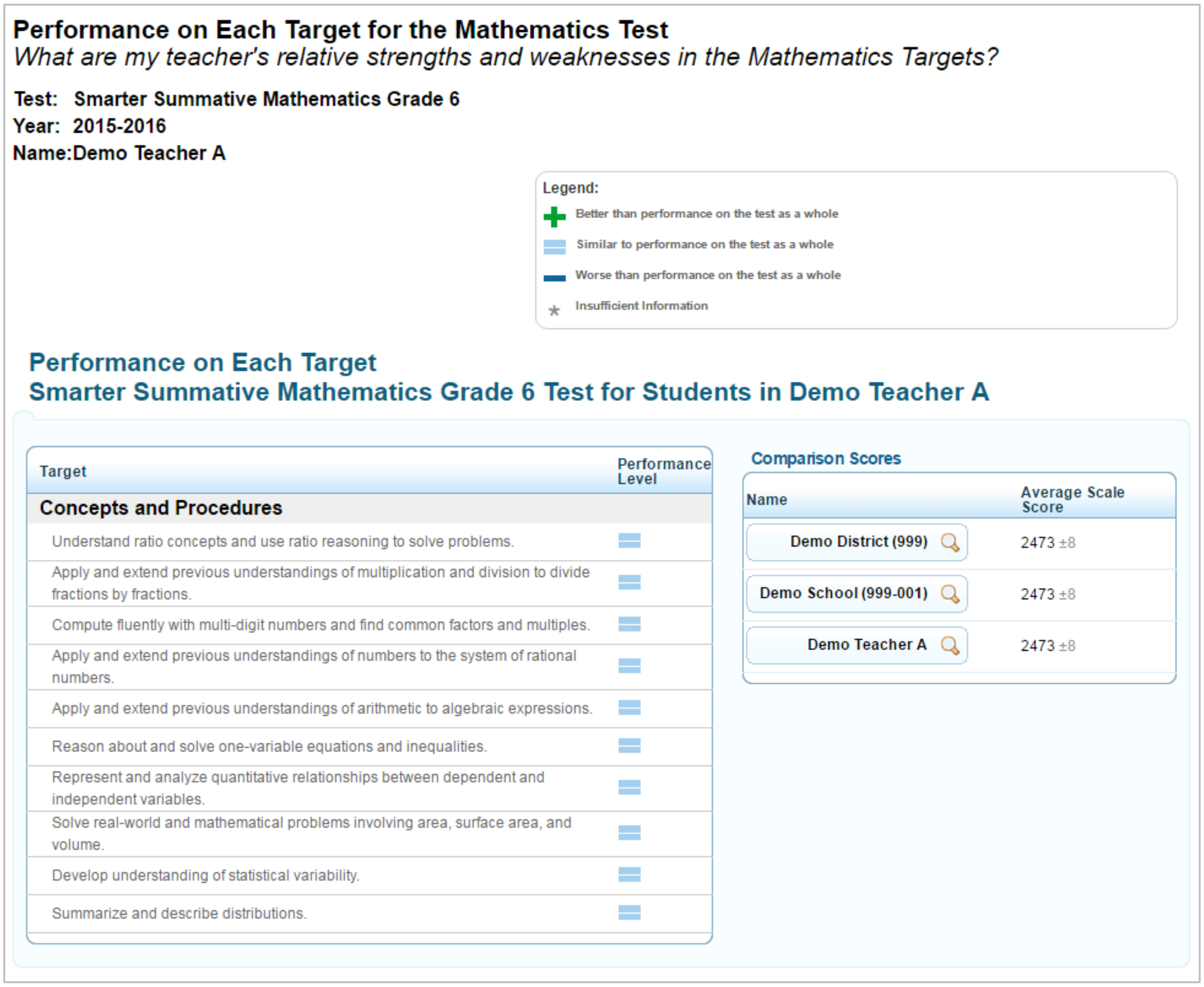
Performance on Each Target
Smarter Summative Mathematics Grade 6 Test for Students in Demo School 999-001

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

Comparison Scores

Name	Average Scale Score
Demo District (999)	2473 ±8
Demo School (999-001)	2473 ±8

Exhibit 7. Target Detail Page for Mathematics: Teacher Level



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

Test: **Smarter Summative Mathematics Grade 6**
Year: **2015-2016**
Name:Demo Teacher A

Legend:
✚ Better than performance on the test as a whole
═ Similar to performance on the test as a whole
▬ Worse than performance on the test as a whole
★ Insufficient Information

Performance on Each Target
Smarter Summative Mathematics Grade 6 Test for Students in Demo Teacher A

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

Comparison Scores

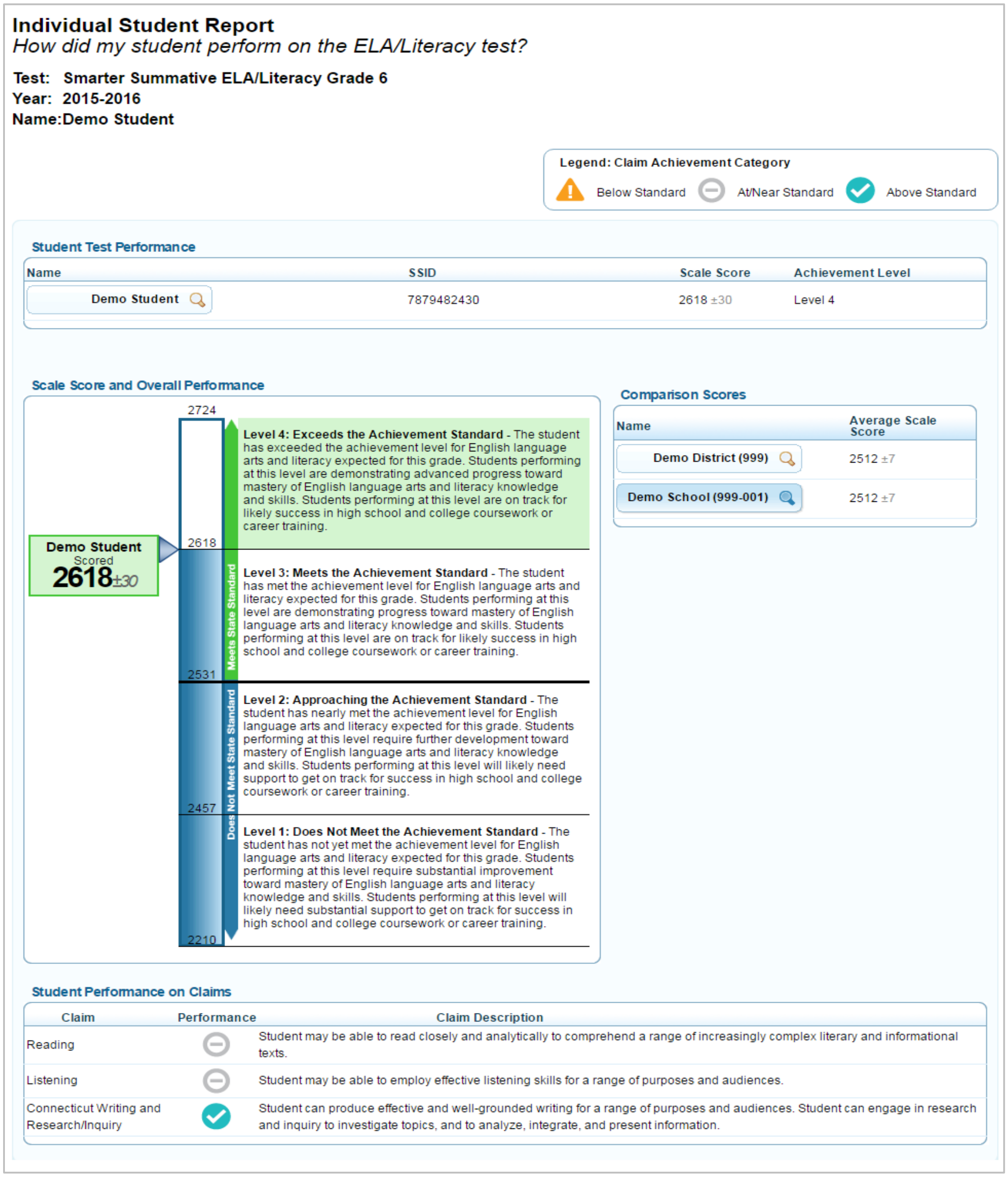
Name	Average Scale Score
Demo District (999)	2473 ±8
Demo School (999-001)	2473 ±8
Demo Teacher A	2473 ±8

7.1.2.5 Student Detail Page

When a student completes a test and the test is hand-scored, an online score report appears in the student detail page in the ORS. The student detail page provides individual student performance on the test. In each subject area, the student detail page provides (1) scale score and standard error of measurement (SEM), (2) achievement level for overall test, (3) achievement category in each claim, (4) average scale scores for student's district, and school.

On the top of the page, the student's name, scale score with SEM, and achievement level are presented. On the left middle section, the student's performance is described in detail using a barrel chart. In the barrel chart, the student's scale score is presented with the SEM using a "±" sign. SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test was administered several times. Further, in the barrel chart, achievement-level descriptors with cut scores at each achievement level are provided, which defines the content area knowledge, skills, and processes that test-takers at each achievement level are expected to possess. On the right middle section, average scale scores and standard errors of the average scale scores for district, and school are displayed so that the student achievement can be compared with the above aggregate levels. It should be noted that the ± next to the student's scale score is the SEM of the scale score whereas the ± next to the average scale scores for aggregate levels represents the standard error of the average scale scores. On the bottom of the page, the student's performance on each reporting category is displayed along with a description of his/her performance on each of the claims. Exhibits 8 and 9 present examples of student detail pages for ELA/L and mathematics.

Exhibit 8. Student Detail Page for ELA/L



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

Test: Smarter Summative ELA/Literacy Grade 6
Year: 2015-2016
Name:Demo Student

Legend: Claim Achievement Category
⚠ Below Standard ⊖ At/Near Standard ✓ Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Demo Student	7879482430	2618 ±30	Level 4

Scale Score and Overall Performance

Demo Student
Scored
2618±30

Level 4: Exceeds the Achievement Standard - The student has exceeded the achievement level 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 level 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 level 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 level 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.

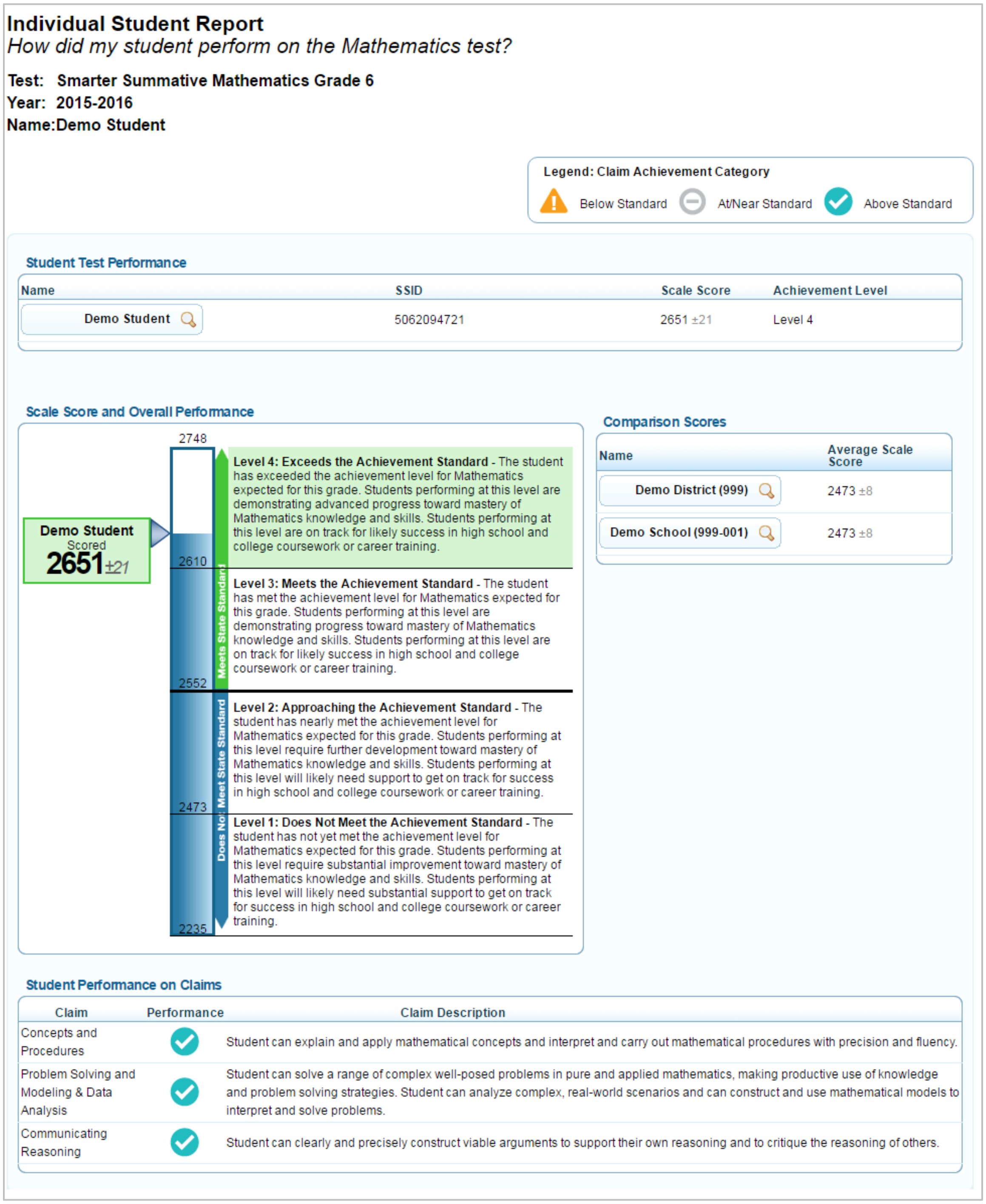
Comparison Scores

Name	Average Scale Score
Demo District (999)	2512 ±7
Demo School (999-001)	2512 ±7

Student Performance on Claims

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

Exhibit 9. Student Detail Page for Mathematics



Individual Student Report
How did my student perform on the Mathematics test?

Test: Smarter Summative Mathematics Grade 6
Year: 2015-2016
Name:Demo Student

Legend: Claim Achievement Category
⚠ Below Standard ⊖ At/Near Standard ✓ Above Standard

Student Test Performance

Name	SSID	Scale Score	Achievement Level
Demo Student	5062094721	2651 ±21	Level 4

Scale Score and Overall Performance

Level 4: Exceeds the Achievement Standard - The student has exceeded the achievement level 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 level 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 level 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 level 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.

Comparison Scores

Name	Average Scale Score
Demo District (999)	2473 ±8
Demo School (999-001)	2473 ±8

Student Performance on Claims

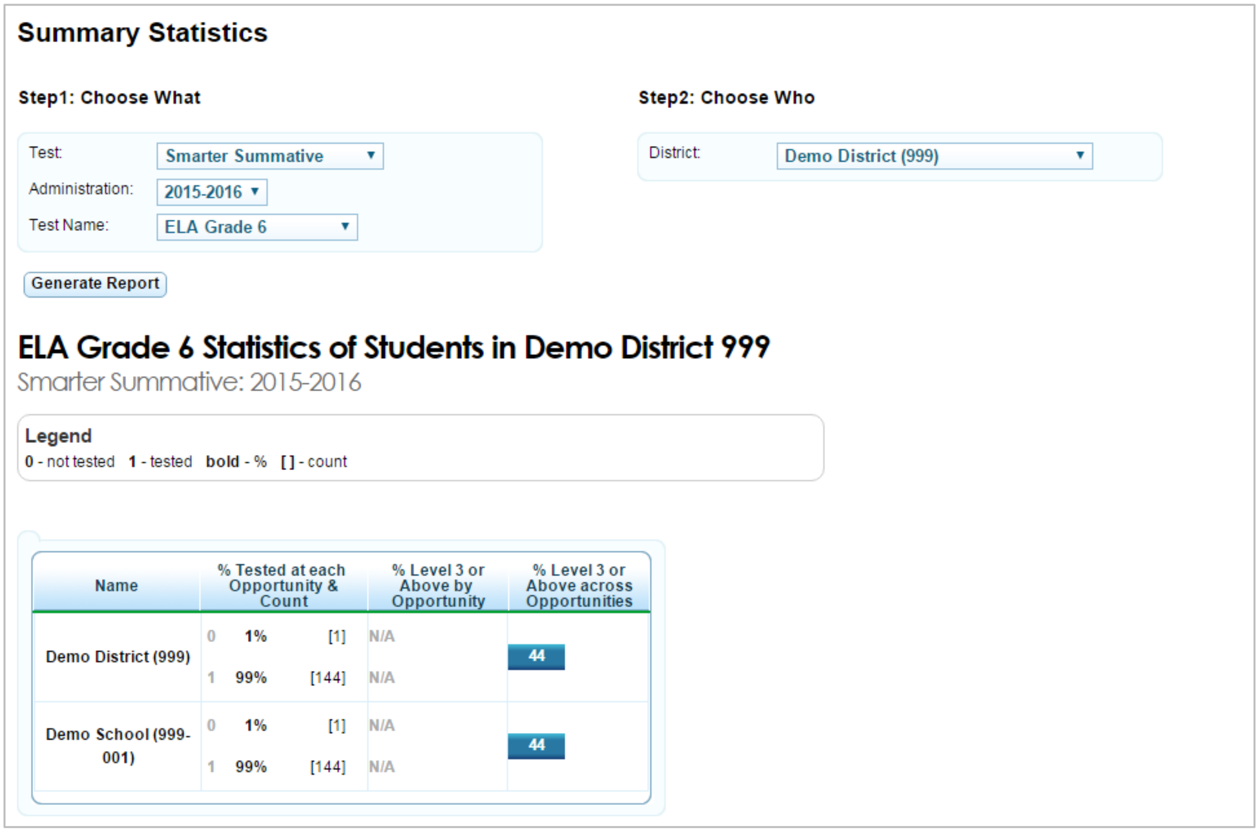
Claim	Performance	Claim Description
Concepts and Procedures	✓	Student can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Problem Solving and Modeling & Data Analysis	✓	Student can solve a range of complex well-posed problems in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

7.1.2.6 Participation Rate

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

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

Exhibit 10. Participation Rate Report at District Level



Summary Statistics

Step1: Choose What

Test: Smarter Summative ▾

Administration: 2015-2016 ▾

Test Name: ELA Grade 6 ▾

Step2: Choose Who

District: Demo District (999) ▾

ELA Grade 6 Statistics of Students in Demo District 999

Smarter Summative: 2015-2016

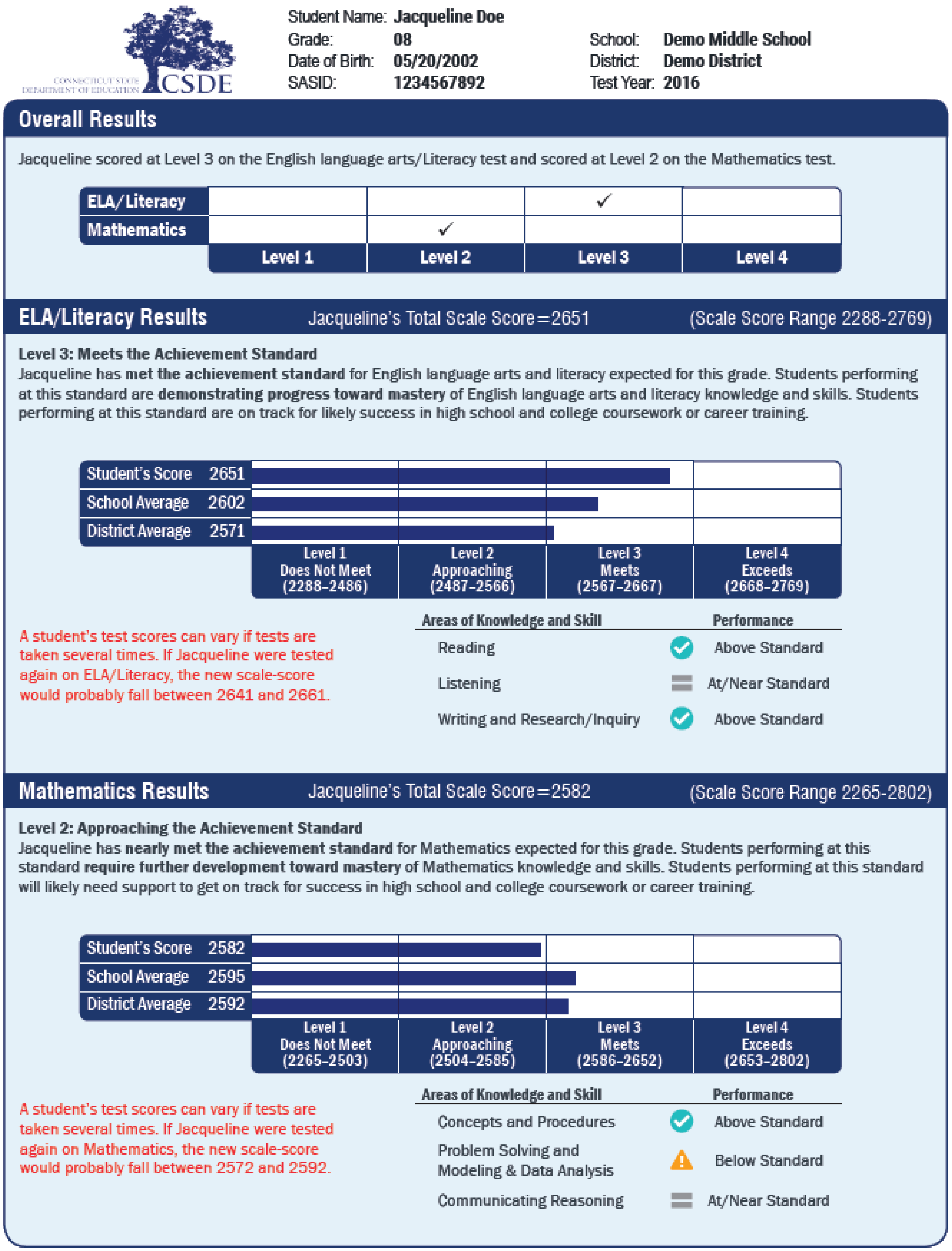
Legend
0 - not tested **1** - tested **bold** - % **[ ]** - count

Name	% Tested at each Opportunity & Count			% Level 3 or Above by Opportunity	% Level 3 or Above across Opportunities
Demo District (999)	0	**1%**	[1]	N/A	44
	1	**99%**	[144]	N/A	
Demo School (999-001)	0	**1%**	[1]	N/A	44
	1	**99%**	[144]	N/A	

7.2 Paper Family Score Reports

After the testing window is closed, parents whose children participated in a test receive a full-color paper score report (hereinafter referred to as a family report) including their child's performance on ELA/L and mathematics. The family report includes information on student performance that is similar to the student detail page from the ORS with additional guidance on how to interpret student achievement results in the family report. An example of a family report is shown in Exhibit 11.

Exhibit 11. Sample Paper Family Score Report



CONNECTICUT STATE DEPARTMENT OF EDUCATION CSDE

Student Name:	**Jacqueline Doe**		
Grade:	**08**	School:	**Demo Middle School**
Date of Birth:	**05/20/2002**	District:	**Demo District**
SASID:	**1234567892**	Test Year:	**2016**

Overall Results

Jacqueline scored at Level 3 on the English language arts/Literacy test and scored at Level 2 on the Mathematics test.

	Level 1	Level 2	Level 3	Level 4
ELA/Literacy			✓	
Mathematics		✓		

ELA/Literacy Results

Jacqueline's Total Scale Score=2651 (Scale Score Range 2288-2769)

Level 3: Meets the Achievement Standard
Jacqueline has **met the achievement standard** for English language arts and literacy expected for this grade. Students performing at this standard are **demonstrating progress toward mastery** of English language arts and literacy knowledge and skills. Students performing at this standard are on track for likely success in high school and college coursework or career training.

Student's Score	2651
School Average	2602
District Average	2571

Level 1 Does Not Meet (2288–2486)	Level 2 Approaching (2487–2566)	Level 3 Meets (2567–2667)	Level 4 Exceeds (2668–2769)

A student's test scores can vary if tests are taken several times. If Jacqueline were tested again on ELA/Literacy, the new scale-score would probably fall between 2641 and 2661.

Areas of Knowledge and Skill		Performance
Reading	✓	Above Standard
Listening	▬	At/Near Standard
Writing and Research/Inquiry	✓	Above Standard

Mathematics Results

Jacqueline's Total Scale Score=2582 (Scale Score Range 2265-2802)

Level 2: Approaching the Achievement Standard
Jacqueline 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 high school and college coursework or career training.

Student's Score	2582
School Average	2595
District Average	2592

Level 1 Does Not Meet (2265–2503)	Level 2 Approaching (2504–2585)	Level 3 Meets (2586–2652)	Level 4 Exceeds (2653–2802)

A student's test scores can vary if tests are taken several times. If Jacqueline were tested again on Mathematics, the new scale-score would probably fall between 2572 and 2592.

Areas of Knowledge and Skill		Performance
Concepts and Procedures	✓	Above Standard
Problem Solving and Modeling & Data Analysis	⚠	Below Standard
Communicating Reasoning	▬	At/Near Standard

7.3 INTERPRETATION OF REPORTED SCORES

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

7.3.1 Scale Score

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

7.3.2 Standard Error of Measurement

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

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

7.3.3 Achievement Level

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

7.3.4 Performance Category for Claims

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

7.3.5 Performance Category for Targets

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

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

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

7.3.6 Aggregated Score

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

7.4 Appropriate Uses for Scores and Reports

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

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

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

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

8. Quality Control Procedure

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

8.1 Adaptive Test Configuration

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

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

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

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

8.1.1 Platform Review

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

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

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

8.1.2 User Acceptance Testing and Final Review

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

8.2 QUALITY ASSURANCE IN DOCUMENT PROCESSING

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

8.3 QUALITY ASSURANCE IN DATA PREPARATION

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

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

8.4 QUALITY ASSURANCE IN HAND SCORING

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

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

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

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

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

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

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

8.4.2 Human-Scoring QA Monitoring Reports

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

8.4.3 Monitoring by Connecticut State Department of Education

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

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

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

This process is also used to notify each Consortium state of possible instances of teacher or proctor interference or student collusion with others. The alert procedure is habitually explained during scorer training sessions. Within the VSC system, if a scorer identifies a response which may require an alert, he or she flags or notes that response as a possible alert and transfers the image to the scoring manager. Scoring management then decides if the response should be forwarded to the client for any necessary action or follow-up.

8.5 QUALITY ASSURANCE IN TEST SCORING

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

A series of Quality Assurance Reports can also be generated at any time during the online assessment window, such as blueprint match rate, item exposure rate, and item statistics, for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved. In addition to these statistics, a cheating analysis report is produced to flag any unlikely patterns of behavior in a testing session as discussed in Section 2.7.

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

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

Table 45 presents an overview of the quality assurance (QA) reports.

Table 45. Overview of Quality Assurance Reports

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

8.5.1 Score Report Quality Check

In the 2015–2016 Smarter Balanced summative assessment, two types of score reports were produced: online reports and printed reports (family reports only).

8.5.1.1 Online Report Quality Assurance

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

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

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DoR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. After scores have passed the QA checks and are uploaded to the DoR, they are passed to the ORS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the ORS until it passes all of the QA system's validation checks. All of the above processes take milliseconds to complete; within less than a second of hand-scores being received by AIR and passing QA validation checks, the composite score will be available in the ORS.

8.5.1.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous quality assurance processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in our reporting specifications document. Upon approval of the specifications, analytic rules are programmed and each program is extensively tested on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implement agreed-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

Much of the statistical processing is repeated, and AIR has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. We write small programs (called macros) that take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in our library for the grades 3–8 and 11 program score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the Director of Score Reporting and the Director of Psychometrics, as well as by the project directors for affected projects.

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

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called VIPP and allows virtually infinite control of the visual appearance of the reports. After designers at AIR create backgrounds, our VIPP programmers write code that indicates where to place all variable information (data, graphics, and text) on the reports. The VIPP code is tested using both artificial and real data. AIR’s data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows the testing of these programs to begin before the statistical programming is complete. In later stages, artificial data are generated according to the input layout and run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables us to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the score reporting team to ensure that design elements are accurately reproduced and data are correctly displayed. Once we receive final data and VIPP programs, the AIR Score Reporting team reviews proofs that contain actual data based on our standard quality assurance documentation. In addition, we compare data independently calculated by AIR psychometricians with data on the reports. A large sample of reports is reviewed by several AIR staff members to make sure that all data are correctly placed on reports. This rigorous review typically is conducted over several days and takes place in a secure location in the AIR building. All reports containing actual data are stored in a locked storage area. Before printing the reports, AIR provides a live data file and individual student reports with sample districts for Department staff review. AIR works closely with the department to resolve questions and correct any problems. The reports are not delivered unless the department approves the sample reports and data file.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 84–105.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 37–46.

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

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Linacre, J. M. (2011). *WINSTEPS Rasch-Model computer program.* Chicago: MESA Press.

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

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

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

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

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

APPENDICES

Appendix A: Number of Students for Interim Assessments

The Interim Comprehensive Assessments (ICA) were fixed-form tests for each grade and subject. No students took ICA for ELA/L. In mathematics, most students took the ICA once, but some students took it twice. Table A–1 presents the number of students who took the ICA once or twice.

Table A–1. Number of Students Who Took ICAs Once or Twice

Grade	**ELA/L**			**Mathematics**		
	Once	**Twice**	**Total**	**Once**	**Twice**	**Total**
3	0	0	0	2	0	2
4	0	0	0	2	0	2
5	0	0	0	0	0	0
6	0	0	0	71	0	71
7	0	0	0	14	0	14
8	0	0	0	10	3	13

For the Interim Assessment Blocks (IAB), there were seven IABs for ELA/L and four IABs in mathematics. Students were allowed to take as many IABs as they wanted. Table A–2 presents the total number of students who took the IABs and the number of students by the number of IABs taken. For example, in grade 3 ELA/L, a total of 7,233 students took IABs, and among 7,233 students, 4,601 students took one IAB, 1,430 students took two IABs, and so on.

Tables A–3 and A–4 disaggregated the number of students in Table A–2 by seven IABs in ELA/L and four IABs in mathematics. For example, 4,601 students in grade 3 ELA/L took one IAB only. Among 4,601 students, none of the students took the Brief Writes IAB.

Table A–2. Number of Students Who Took IABs

Grade	**Total**	**Number of IABs Taken**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
				ELA/L				
3	7,233	4,601	1,430	770	251	173	8	
4	7,468	5,036	1,270	830	263	69		
5	6,314	3,626	1,951	624	25	81	7	
6	6,355	4,032	1,865	267	107	84		
7	5,974	3,655	1,797	345	158	19		
8	5,730	3,475	1,618	504	81	52		
				Mathematics				
3	8,854	4,328	2,182	2,323	21			
4	8,214	3,850	1,971	2,339	54			
5	8,148	4,437	1,899	1,797	15			
6	8,097	5,063	1,367	1,647	20			
7	8,398	4,424	1,893	2,036	45			
8	8,668	5,074	1,596	1,969	29			

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

Grade	**Block**	**Number of IABs Taken**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
3	Brief Writes		3	10	12	7	8	
	Editing and Revising	813	1,061	624	238	173	8	
	Listening and Interpretation	1,149	866	481	152	173	8	
	Performance Task	2						
	Reading Informational Text	1,622	231	384	230	173	8	
	Reading Literary Text	190	340	419	178	170	8	
	Research	825	359	392	194	169	8	
4	Brief Writes				1	3		
	Editing and Revising	637	970	562	216	69		
	Listening and Interpretation	1,611	859	553	227	69		
	Performance Task	82			9	1		
	Reading Informational Text	1,581	268	511	209	69		
	Reading Literary Text	107	184	457	211	68		
	Research	1,018	259	407	179	66		
5	Brief Writes					1	7	
	Editing and Revising	451	1,292	534	22	81	7	
	Listening and Interpretation	861	1,327	488	22	80	7	
	Performance Task				1			
	Reading Informational Text	1,352	612	306	23	81	7	
	Reading Literary Text	79	85	182	9	81	7	
	Research	883	586	362	23	81	7	
6	Brief Writes		1					
	Editing and Revising	474	1,684	253	103	84		
	Listening and Interpretation	798	1,554	241	106	84		
	Performance Task							
	Reading Informational Text	1,938	347	89	88	84		
	Reading Literary Text	143	100	59	40	84		
	Research	679	44	159	91	84		
7	Brief Writes		1					
	Editing and Revising	475	1,444	319	145	19		
	Listening and Interpretation	650	1,328	292	140	19		
	Performance Task		4	2				
	Reading Informational Text	1,777	533	94	152	19		
	Reading Literary Text	110	207	44	40	19		
	Research	643	77	284	155	19		
8	Brief Writes							
	Editing and Revising	554	1,256	503	80	52		
	Listening and Interpretation	687	1,314	324	79	52		
	Performance Task	2		1				
	Reading Informational Text	1,585	394	192	68	52		
	Reading Literary Text	42	199	16	74	52		
	Research	605	73	476	23	52		

Table A–4: Mathematics Number of Students Who Took IABs by Block Labels

Grade	Block	Number of IABs Taken			
		1	2	3	4
3	Measurement and Data	345	684	2,246	21
	Number and Operations – Fractions	1,067	1,742	2,315	21
	Operational and Algebraic Thinking	2,910	1889	2,323	21
	Performance Task	6	49	85	21
4	Number and Operations in Base Ten	1,580	1,284	2,339	54
	Number and Operations – Fractions	1,116	1,004	2,306	54
	Operational and Algebraic Thinking	1,149	1,564	2,308	54
	Performance Task	5	90	64	54
5	Measurement and Data	443	468	1,795	15
	Number and Operations in Base Ten	2,101	1,702	1,792	15
	Number and Operations – Fractions	1,855	1,545	1,796	15
	Performance Task	38	83	8	15
6	Expressions and Equations	1,473	939	1,643	20
	Geometry	493	109	1,643	20
	Performance Task	102	370	9	20
	Ratios and Proportional Relationships	2,995	1,316	1,646	20
7	Expressions and Equations	1,299	962	2,034	45
	The Number System	2,217	1,383	2,035	45
	Performance Task	8	90	107	45
	Ratios and Proportional Relationships	900	1,351	1,932	45
8	Expressions and Equations	1,310	1,284	1,929	29
	Functions	1,952	1,179	1,960	29
	Geometry	1,624	510	1,955	29
	Performance Task	188	219	63	29

Appendix B: Percentage of Proficient Students in 2014-2015 and 2015-2016 for All Students and by Subgroups

Table B-1. ELA/L Student Performance Across Years (Grades 3–5)

Group	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 3									
All Students	37,987	2436	88	54	38,942	2438	89	54	0
Female	18,577	2447	86	58	19,139	2447	88	58	0
Male	19,410	2426	89	49	19,803	2430	90	50	1
American Indian or Alaska Native	109	2410	80	40	90	2422	78	48	8
Asian	1,917	2479	84	73	2,151	2480	84	74	1
African American	4,922	2386	79	28	4,874	2392	81	31	3
Hispanic or Latino	8,995	2390	80	31	9,854	2395	82	33	2
Native Hawaiian/Pacific Islander	32	2435	84	59	47	2420	92	38	-21
White	20,815	2464	79	68	20,601	2465	82	67	-1
Multiple Ethnicities	1,197	2442	87	55	1,325	2450	87	57	2
LEP	2,852	2354	68	13	3,554	2362	70	16	3
IDEA Eligible	4,363	2349	78	16	4,332	2357	78	17	1
Grade 4									
All Students	38,597	2479	93	55	38,450	2480	96	56	1
Female	19,065	2491	90	60	18,805	2490	94	59	-1
Male	19,532	2467	93	50	19,645	2471	97	52	2
American Indian or Alaska Native	113	2454	86	42	102	2446	98	42	0
Asian	1,969	2525	84	75	1,996	2526	91	74	-1
African American	4,778	2424	84	29	4,955	2427	88	31	2
Hispanic or Latino	8,770	2429	87	32	9,383	2430	89	33	1
Native Hawaiian/Pacific Islander	40	2480	102	58	29	2486	89	55	-3
White	21,936	2506	83	68	20,825	2511	85	70	2
Multiple Ethnicities	991	2489	92	57	1,160	2493	95	59	2
LEP	2,692	2389	76	14	2,962	2384	78	14	0
IDEA Eligible	4,695	2384	80	15	4,934	2390	84	17	2
Grade 5									
All Students	38,817	2516	92	59	39,010	2517	97	59	0
Female	18,884	2529	90	64	19,273	2531	94	64	0
Male	19,933	2503	93	53	19,737	2504	98	53	0
American Indian or Alaska Native	96	2496	80	46	112	2501	95	54	8
Asian	1,996	2559	86	76	2,003	2563	90	77	1
African American	4,876	2460	85	33	4,840	2461	90	33	0
Hispanic or Latino	8,382	2465	87	35	9,201	2467	92	37	2
Native Hawaiian/Pacific Islander	29	2528	95	62	43	2525	109	63	1
White	22,476	2542	82	71	21,826	2547	86	72	1
Multiple Ethnicities	962	2520	90	60	985	2528	96	62	2
LEP	2,351	2418	70	12	2,694	2412	75	13	1
IDEA Eligible	4,955	2418	81	16	5,070	2420	84	17	1

Table B-2. ELA/L Student Performance Across Years (Grades 6-8)

Group	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 6									
All Students	39,710	2538	92	56	39,071	2536	98	55	-1
Female	19,307	2552	88	62	18,963	2548	95	60	-2
Male	20,403	2524	93	49	20,108	2525	100	50	1
American Indian or Alaska Native	119	2515	84	47	95	2527	94	47	0
Asian	1,959	2590	82	78	1,990	2581	90	73	-5
African American	4,833	2485	83	30	4,881	2482	91	31	1
Hispanic or Latino	8,454	2487	88	32	8,794	2481	94	31	-1
Native Hawaiian/Pacific Islander	41	2556	100	56	32	2541	105	50	-6
White	23,295	2563	82	67	22,299	2565	87	68	1
Multiple Ethnicities	1,009	2545	92	59	980	2542	95	56	-3
LEP	2,047	2428	74	8	2,112	2411	75	6	-2
IDEA Eligible	5,042	2441	81	14	5,193	2438	87	15	1
Grade 7									
All Students	38,782	2560	95	57	40,085	2559	100	55	-2
Female	18,838	2576	91	64	19,410	2573	96	61	-3
Male	19,944	2545	97	50	20,675	2546	101	50	0
American Indian or Alaska Native	87	2531	83	39	113	2537	95	43	4
Asian	1,876	2613	87	79	1,994	2613	91	77	-2
African American	5,001	2507	88	32	4,917	2502	89	29	-3
Hispanic or Latino	8,082	2507	92	34	8,836	2505	95	32	-2
Native Hawaiian/Pacific Islander	24	2564	96	58	43	2555	117	56	-2
White	22,837	2586	85	69	23,119	2587	89	67	-2
Multiple Ethnicities	875	2567	91	60	1,063	2566	101	59	-1
LEP	1,827	2439	70	7	2,074	2430	71	5	-2
IDEA Eligible	4,948	2457	80	13	5,232	2460	86	15	2
Grade 8									
All Students	39,610	2572	96	54	39,351	2574	100	55	1
Female	19,223	2589	92	62	19,157	2589	96	62	0
Male	20,387	2556	97	47	20,194	2559	102	49	2
American Indian or Alaska Native	106	2541	92	43	94	2556	93	44	1
Asian	1,752	2625	88	76	1,925	2627	93	76	0
African American	5,067	2519	86	29	5,068	2520	92	32	3
Hispanic or Latino	8,059	2520	91	31	8,546	2519	95	33	2
Native Hawaiian/Pacific Islander	36	2564	96	50	26	2585	106	58	8
White	23,740	2597	87	65	22,770	2601	90	67	2
Multiple Ethnicities	850	2581	96	57	922	2582	100	59	2
LEP	1,723	2450	68	5	1,791	2437	68	4	-1
IDEA Eligible	4,941	2473	81	13	5,171	2473	85	15	2

Table B-3. Mathematics Student Performance Across Years (Grades 3–5)

Group	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 3									
All Students	38,249	2427	80	48	38,870	2438	81	53	5
Female	18,701	2427	77	47	19,109	2438	78	52	5
Male	19,548	2428	84	49	19,761	2439	84	53	4
American Indian or Alaska Native	111	2406	85	36	90	2431	77	51	15
Asian	1,961	2477	80	71	2,147	2491	76	78	7
African American	4,943	2379	72	21	4,860	2391	75	27	6
Hispanic or Latino	9,176	2385	73	24	9,833	2398	75	31	7
Native Hawaiian/Pacific Islander	32	2416	70	34	46	2421	77	46	12
White	20,829	2453	71	62	20,569	2463	72	67	5
Multiple Ethnicities	1,197	2433	79	49	1,325	2447	77	56	7
LEP	3,117	2359	68	11	3,546	2377	70	20	9
IDEA Eligible	4,384	2350	80	15	4,324	2360	82	18	3
Grade 4									
All Students	38,829	2470	80	44	38,387	2478	82	48	4
Female	19,180	2469	76	43	18,773	2476	78	47	4
Male	19,649	2471	84	45	19,614	2480	86	49	4
American Indian or Alaska Native	115	2452	74	34	102	2450	87	36	2
Asian	2,002	2523	79	70	1,992	2533	82	73	3
African American	4,783	2419	70	17	4,938	2427	73	21	4
Hispanic or Latino	8,929	2426	72	21	9,372	2434	74	24	3
Native Hawaiian/Pacific Islander	41	2468	96	46	29	2488	77	55	9
White	21,971	2494	71	57	20,794	2504	72	62	5
Multiple Ethnicities	988	2480	83	46	1,160	2488	81	51	5
LEP	2,942	2400	70	11	2,954	2405	69	12	1
IDEA Eligible	4,695	2392	76	11	4,916	2401	75	13	2
Grade 5									
All Students	39,044	2493	87	37	38,941	2501	89	41	4
Female	18,980	2492	83	35	19,242	2500	86	40	5
Male	20,064	2495	91	38	19,699	2502	93	42	4
American Indian or Alaska Native	96	2468	69	20	112	2488	84	32	12
Asian	2,019	2547	87	60	1,999	2562	87	68	8
African American	4,889	2434	75	11	4,830	2441	77	14	3
Hispanic or Latino	8,550	2444	78	15	9,173	2452	80	18	3
Native Hawaiian/Pacific Islander	30	2499	85	33	43	2511	103	37	4
White	22,499	2520	77	49	21,798	2530	79	54	5
Multiple Ethnicities	961	2498	86	35	986	2512	91	43	8
LEP	2,586	2410	70	5	2,688	2415	69	6	1
IDEA Eligible	4,958	2409	77	7	5,055	2416	78	9	2

Table B-4. Mathematics Student Performance Across Years (Grades 6-8)

Group	2014–2015				2015–2016				Change in %Proficient
	N	Mean	SD	%Prof	N	Mean	SD	%Prof	
Grade 6									
All Students	39,870	2513	100	37	38,965	2521	104	41	4
Female	19,372	2516	94	37	18,921	2523	99	41	4
Male	20,498	2511	105	37	20,044	2519	108	41	4
American Indian or Alaska Native	121	2483	92	21	95	2499	94	31	10
Asian	1,979	2584	96	65	1,988	2588	99	66	1
African American	4,841	2449	88	12	4,860	2452	95	14	2
Hispanic or Latino	8,577	2456	95	15	8,769	2461	97	17	2
Native Hawaiian/Pacific Islander	40	2537	112	53	32	2530	117	41	-12
White	23,299	2542	87	48	22,243	2553	89	53	5
Multiple Ethnicities	1,013	2520	100	39	978	2525	101	40	1
LEP	2,230	2402	88	4	2,107	2402	86	4	0
IDEA Eligible	5,042	2408	95	7	5,158	2412	96	7	0
Grade 7									
All Students	39,001	2530	106	39	39,961	2538	108	42	3
Female	18,952	2532	101	38	19,352	2540	102	42	4
Male	20,049	2528	111	39	20,609	2536	112	42	3
American Indian or Alaska Native	88	2491	92	18	113	2509	89	29	11
Asian	1,901	2605	101	68	1,988	2617	103	71	3
African American	5,026	2466	94	14	4,895	2467	95	14	0
Hispanic or Latino	8,270	2468	98	16	8,798	2477	101	19	3
Native Hawaiian/Pacific Islander	25	2525	101	32	43	2546	119	44	12
White	22,816	2560	93	50	23,063	2570	93	54	4
Multiple Ethnicities	875	2537	103	40	1,061	2544	108	44	4
LEP	2,053	2412	87	4	2,057	2415	89	5	1
IDEA Eligible	4,957	2421	93	7	5,189	2427	99	9	2
Grade 8									
All Students	39,764	2541	114	37	39,181	2551	116	40	3
Female	19,237	2546	108	38	19,069	2557	110	42	4
Male	20,429	2536	120	36	20,112	2546	121	39	3
American Indian or Alaska Native	105	2505	102	23	94	2509	107	20	-3
Asian	1,788	2621	113	64	1,922	2636	113	69	5
African American	5,058	2468	94	12	5,043	2479	100	15	3
Hispanic or Latino	8,166	2476	102	15	8,504	2485	103	17	2
Native Hawaiian/Pacific Islander	37	2521	112	32	26	2551	127	31	-1
White	23,669	2573	104	48	22,679	2585	104	52	4
Multiple Ethnicities	843	2544	112	35	913	2559	115	43	8
LEP	1,917	2417	90	4	1,779	2419	85	3	-1
IDEA Eligible	4,848	2429	94	6	5,131	2438	95	7	1

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	38,942	78	88	69	65	88	70	82	58	54	82
Female	19,139	78	88	69	65	88	70	80	58	54	83
Male	19,803	78	88	69	65	87	70	82	58	54	81
American Indian or Alaska Native	90	76	89	68	67	84	67	83	57	58	73
Asian	2,151	80	86	69	65	89	72	78	56	54	86
African American	4,874	79	89	69	65	84	70	85	58	54	74
Hispanic or Latino	9,854	78	89	69	65	85	70	84	59	53	75
Native Hawaiian/Pacific Islander	47	78	89	70	61*	89	70	78	60	53*	83
White	20,601	78	86	69	65	88	70	76	58	54	83
Multiple	1,325	78	87	69	64	88	70	78	60	53	83
LEP	3,554	80	90	69	65	80	73	86	58	52	61
IDEA	4,332	83	92	69	65	84	77	89	58	52	73
Grade 4											
All Students	38,450	77	89	61	62	87	69	83	48	51	82
Female	18,805	77	89	61	62	88	69	82	48	51	83
Male	19,645	77	90	61	62	87	69	84	48	51	80
American Indian/Alaska Native	102	80	91	62	64	85	73	88	49	51	81
Asian	1,996	79	87	61	61	91	73	79	47	51	87
African American	4,955	78	91	61	62	84	71	87	49	51	73
Hispanic or Latino	9,383	78	91	61	62	83	70	87	49	51	72
Native Hawaiian/Pacific Islander	29	77	83*	62*	66*	95*	68	80*	48*	55*	83*
White	20,825	76	86	60	62	88	68	77	48	51	83
Multiple	1,160	78	88	60	62	89	70	81	48	51	84
LEP	2,962	83	93	61	62	79	77	91	49	50	60
IDEA	4,934	83	93	61	62	83	77	91	48	50	68
Grade 5											
All Students	39,011	78	89	63	72	86	70	83	51	62	79
Female	19,274	78	88	63	72	86	70	81	51	62	81
Male	19,737	78	90	63	72	85	70	84	51	62	78
American Indian/Alaska Native	112	77	90	62	74	81	68	84	49	65	70
Asian	2,004	80	87	63	71	88	73	79	49	61	85
African American	4,840	79	91	63	71	82	71	86	52	61	70
Hispanic or Latino	9,201	79	90	63	72	82	71	86	52	63	71
Native Hawaiian/Pacific Islander	43	81	88	67*	69	92	73	82	51*	62	85
White	21,826	78	86	63	72	86	69	76	51	62	81
Multiple	985	78	87	64	71	88	70	80	51	62	82
LEP	2,694	83	92	63	71	76	77	90	52	58	47
IDEA	5,070	83	92	63	72	81	77	90	52	59	66

*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	39,071	76	87	65	71	83	67	80	54	62	74
Female	18,963	76	87	65	71	83	67	78	54	62	75
Male	20,108	76	88	65	71	82	68	81	54	62	73
American Indian/Alaska Native	95	76	86	66	72	84	66	76	56	61	75
Asian	1,990	77	85	65	71	85	68	75	53	62	80
African American	4,881	77	88	66	71	78	68	83	55	62	61
Hispanic or Latino	8,794	77	89	65	71	79	69	84	55	61	65
Native Hawaiian/Pacific Islander	32	77	80*	64*	68*	89	69	73*	54*	54*	86*
White	22,299	75	84	65	72	83	66	73	54	63	76
Multiple	980	75	85	66	72	83	66	76	55	63	74
LEP	2,112	85	92	65	70	71*	80	90	53	52	43*
IDEA	5,193	82	91	65	70	77	75	88	54	57	63
Grade 7											
All Students	40,085	77	87	65	75	84	69	80	54	67	75
Female	19,410	77	86	65	75	84	68	77	54	67	76
Male	20,675	78	88	66	75	84	69	81	54	67	74
American Indian/Alaska Native	113	75	85	64	73	85	67	75	55	65	78
Asian	1,994	79	85	66	75	87	71	73	53	66	82
African American	4,917	78	88	65	75	80	70	83	54	65	65
Hispanic or Latino	8,836	78	89	65	74	81	70	84	54	65	68
Native Hawaiian/Pacific Islander	43	79	92	69*	70	83	71	86	56*	61	76
White	23,119	77	84	65	75	84	68	73	54	67	76
Multiple	1,063	78	89	65	75	85	70	81	54	67	76
LEP	2,074	86	92	65	71	75	81	90	52	52	60
IDEA	5,232	82	91	65	73	79	76	88	53	62	63
Grade 8											
All Students	39,351	78	87	69	77	83	70	80	58	70	74
Female	19,157	78	86	69	77	84	70	78	58	70	75
Male	20,194	79	88	70	77	82	70	81	58	69	72
American Indian/Alaska Native	94	77	89	69	75	84	68	79	61	65	72
Asian	1,925	80	85	70	76	87	72	74	58	69	81
African American	5,068	79	88	70	77	81	71	82	59	68	66
Hispanic or Latino	8,546	80	89	69	77	80	72	84	59	69	64
Native Hawaiian/Pacific Islander	26	78	79*	71*	78*	83*	70	73*	60*	72*	73*
White	22,770	78	84	69	77	83	69	74	58	70	74
Multiple	922	78	83	69	77	84	70	77	57	69	76
LEP	1,791	87	92	68	72	78*	82	91	55	54	68*
IDEA	5,171	83	90	69	76	80	76	87	58	65	60

*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	38,870	82	90	73	79	89	75	84	64	72	84
Female	19,109	82	89	73	79	88	75	83	64	72	83
Male	19,761	83	90	73	79	89	76	85	64	71	84
American Indian/Alaska Native	90	81	90	69	83	81	74	85	61	74	73
Asian	2,147	85	86	74	79	92	79	78	62	72	89
African American	4,860	83	91	74	78	84	76	87	64	69	76
Hispanic or Latino	9,833	83	91	73	78	85	76	86	64	70	77
Native Hawaiian/Pacific Islander	46	82	91	73	82	86*	74	82	64	73	80*
White	20,569	82	87	73	79	89	75	78	63	72	84
Multiple	1,325	82	88	74	80	89	75	82	64	73	84
LEP	3,546	84	92	74	78	84	78	88	64	69	72
IDEA	4,324	87	94	73	77	86	82	92	62	68	78
Grade 4											
All Students	38,387	84	89	81	79	89	77	83	73	71	84
Female	18,733	83	88	80	79	88	76	82	73	71	82
Male	19,614	84	90	81	79	89	78	84	73	71	85
American Indian/Alaska Native	102	83	91	78	74	88	76	86	71	67	79
Asian	1,992	86	85	80	78	93	80	76	71	71	90
African American	4,938	85	91	80	78	84	78	86	73	68	76
Hispanic or Latino	9,372	84	90	80	79	86	78	85	73	70	77
Native Hawaiian/Pacific Islander	29	79	99*	70	75	92*	72	78*	62	67	89*
White	20,794	83	87	81	79	89	76	78	73	72	84
Multiple	1,160	84	87	80	79	90	77	79	74	72	85
LEP	2,954	86	92	80	77	84	81	89	72	66	74
IDEA	4,916	88	93	80	78	86	82	90	72	67	78
Grade 5											
All Students	38,941	83	91	78	71	89	76	86	69	61	84
Female	19,242	82	90	78	71	88	76	85	70	61	83
Male	19,699	83	91	78	71	89	77	87	69	61	84
American Indian/Alaska Native	112	83	89	79	68	92	77	84	71	58	86
Asian	1,999	84	88	77	71	93	78	80	68	61	90
African American	4,830	86	92	77	71	86	80	90	68	59	75
Hispanic or Latino	9,173	85	92	77	71	84	79	89	68	60	76
Native Hawaiian/Pacific Islander	43	85	91	72	74*	93	80	87	67	54*	93
White	21,798	81	88	78	72	88	74	81	70	62	84
Multiple	986	83	90	78	72	92	77	84	71	61	88
LEP	2,689	88	94	74	70	88	84	92	63	58	75
IDEA	5,055	90	95	77	70	85	86	93	66	59	77

*The classification index is based on n<10.

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

Group	N	%Accuracy					%Consistency				
		All	**L1**	**L2**	**L3**	**L4**	**All**	**L1**	**L2**	**L3**	**L4**
Grade 6											
All Students	38,965	82	91	77	72	89	75	86	69	61	83
Female	18,921	82	91	77	72	88	75	85	69	61	82
Male	20,044	83	92	77	72	89	76	87	69	61	84
American Indian/Alaska Native	95	82	91	76	72	90	75	84	68	62	80
Asian	1,988	84	88	77	71	93	78	82	68	61	90
African American	4,860	86	93	76	71	83	80	90	68	59	73
Hispanic or Latino	8,769	85	93	77	71	86	79	89	69	59	75
Native Hawaiian/Pacific Islander	32	82	87	71*	75*	87*	76	85	60*	58*	85*
White	22,243	81	89	77	72	89	73	80	69	62	83
Multiple	978	82	90	78	71	90	75	84	70	61	85
LEP	2,107	91	95	75	72	89	88	94	64	55	81
IDEA	5,158	90	95	76	70	86	86	94	67	58	77
Grade 7											
All Students	39,961	83	91	77	75	90	76	86	69	66	84
Female	19,352	82	90	77	75	89	75	85	69	66	83
Male	20,609	84	92	77	75	91	77	87	69	66	85
American Indian/Alaska Native	113	82	90	77	75	86*	75	86	66	67	78*
Asian	1,988	86	88	77	76	94	80	80	68	67	91
African American	4,895	86	92	77	75	84	80	89	68	63	73
Hispanic or Latino	8,798	85	93	77	74	87	79	89	69	64	77
Native Hawaiian/Pacific Islander	43	86	93	78	75*	93	79	89	72	62*	88
White	23,063	82	88	77	75	90	74	80	69	66	84
Multiple	1,061	83	91	77	76	90	76	86	69	67	85
LEP	2,057	91	95	76	72	90	88	94	64	57	83
IDEA	5,189	90	95	76	74	87	86	94	66	62	79
Grade 8											
All Students	39,181	82	90	72	72	90	75	85	62	62	85
Female	19,069	81	89	72	72	89	74	83	62	62	84
Male	20,112	83	91	72	72	90	76	86	61	62	86
American Indian/Alaska Native	94	83	92	71	68*	87	76	87	64	51*	80
Asian	1,922	84	86	71	72	94	78	78	61	62	91
African American	5,043	84	92	71	72	85	78	89	60	59	77
Hispanic or Latino	8,504	84	92	71	72	87	78	89	60	60	79
Native Hawaiian/Pacific Islander	26	83	88*	71*	80*	93*	76	81*	65*	44*	94*
White	22,679	80	87	72	72	89	72	80	63	62	85
Multiple	913	82	89	72	72	90	75	84	61	63	85
LEP	1,779	91	95	69	73	89	88	94	53	53	82
IDEA	5,131	89	94	70	71	87	85	93	57	57	78

*The classification index is based on n<10.